

SIREN DISTRIBUTION REPORTS

EDITOR: ANDRÁS KORNAI

TECHNICAL EDITOR: MÁRTON MAKRAI

Draft version 1.0, January 17 2016. Please do not circulate, quote, or cite without express permission from the editor, who can be reached at andras@kornai.com.

Introduction (András Kornai)

This is a report on the distribution and digital status of some selected languages discussing each language in a separate chapter. This introductory chapter summarizes data from resources that are available for nearly all of the selected languages, and introduces a uniform structure that is followed by all reports to the extent feasible.

Universal resources

Table 1 summarizes data from resources that are available for nearly all of the reported languages.

The number of native (L1) speakers has been extracted from the [Ethnologue](#) database and the [Endangered Languages Project](#), a catalog of 3346 languages. When the data is available at both resources, the mean of the two values is reported.

An [Crúbadán](#) is collection of (crawled) textual data for 2000 languages. The table includes the following numbers: `docs` is the number of documents (as reported on the site), `words` is the number of words (based on the downloadable language datasets), `floss splchk` shows the existence of a free, open source spell checker, `watchtower` shows the existence of a Watchtower edition, and `udhr` shows the existence of an edition of the Universal Declaration of Human Rights.

It is also shown whether the language is listed in [Omniglot](#), an Encyclopedia of writing systems and languages.

[language-archives.org](#) is an archive of language resources, maintained by the Open Language Archives Community with resources categorized as Primary texts (`primary texts`), lexical resources (`lex res`), language descriptions (`lang descr`), resources about the language (`res about` and `oth res about`), or resources in the language (`res in` and `oth res in`). For all categories two features are extracted, one for the number of all resources listed (`all`), and one for number of the resources available online (`online`).

The table includes a end-user software with binary features (whether the language is supported). The list of Office 13 language packs have been downloaded from <https://support.office.com>.

The [World Atlas of Language Structures \(WALS\)](#) is a database of structural properties of languages gathered from descriptive materials.

Wikipedia has an edition in all the reported languages, but the Bengali one is still in so called incubator state. The column `real articles` shows the number of articles that are supposedly not machine-generated as their length exceeds a threshold, and `adjusted size` is an approximation of the information content in the edition (character length normalized by unigram character entropy).

The next column shows the number of (linguistic) features reported in [Uriel](#), a structured compendium of information on language typology and language universals.

Most of the features in the last two panels are binary, showing whether some resource is available for the language: a translation of the Bible at <https://bibles.org> which is a Collection of Bible

name	code	An Críobháin												language-archives.org												Wikipedia												end-user software																
		L1				Ethnologue status				An Críobháin				language-archives.org				Wikipedia				end-user software				WALS																												
		docs		words		floss spchlk		watchtower		udhr		in Omniglot		primary texts online		primary texts all		lang descr online		lang descr all		la lex res online		lex res all		oth res in online		oth res in all		oth res about online		oth res about all		Mac input (language pack)																				
Akan	aka	1	8314600	3	176	267975	1	1	1	3	3	12	16	1	2	28	40	0	1	0	0	1	0	1	0	0	259	1842	20697	0	6144	13	0	0	23	212505.634726	244	0	1	4	0	0	0											
Amharic	amh	4	2181600	1	3726	9072217	1	1	1	1	3	9	13	6	31	161	32	50	0	0	0	1	1	1	1	0	13140	42644	346101	3	21487	29	1595	41	0	1846	1654493.2994	260	0	1	12	0	1	0										
Arabic	ara	0	20787180	0.0	9542	652506	1	1	1	0	7	10	20	28	5	11	110	115	4	41	1	1	1	1	1	1	0	44870	2619221	23717395	34	1221400	36889	29020	221	0	129987	1678904836.13	0	0	1	48	0	1	0									
Bengali	ben	0	189261200	1	899	1027298	0	0	0	1	5	5	12	16	4	10	24	24	22	29	0	0	1	1	1	1	1	0	4379	463822	2406277	17	122867	598	3640	473	0	25151	26923808.779	232	1	1	11	0	1	1								
Mandarin	cmn	178	847898270	0	1	202	0	0	1	5	5	10	14	8	9	262	299	255	266	0	0	1	0	1	0	0	0	503472	3324144	2277666	32	5060606	3208	38449	215	0	88351	884553402.699	273	1	0	43	0	0	0									
Hindi	hin	120	203033620	1	1552	2439262	1	1	1	1	7	7	9	25	4	10	41	52	496	511	0	1	1	0	1	1	1	0	107032	506376	3256543	4	235882	543	2633	114	0	28383	534533.431.824	265	0	1	9	0	1	1								
Hungarian	hun	0	12603560	1	4526	1350172	1	1	1	1	11	12	10	17	12	15	41	32	35	28	38	1	1	1	1	1	1	0	301920	1186351	1858971	34	104468	1281660	813308	230	0	193051	26928715983.73	274	1	1	18	0	0	0								
Indonesian	ind	140	23230480	1	1660	2285404	1	1	1	10	12	20	148	279	230	21	81	1	1	1	1	1	1	1	0	384876	1900941	1601660	29	813308	2083	61297	105	0	104468	118251792.32	278	1	1	11	0	1	1											
Russian	rus	110	166167580	0	31	658898	0	0	1	6	7	8	34	7	30	63	208	283	342	0	0	0	1	1	1	1	0	1338620	4595027	92887520	89	195649	9751	191495	137	0	890160	1069763744.34	273	1	1	22	1	0	1									
Somali	som	0.095	14763500	1	13490	1262897	1	0	1	1	1	4	7	2	4	16	23	0	1	0	1	1	0	0	0	4135	16627	171475	2	14721	59	126	94	0	1864	201066516894	264	1	1	6	0	1	0											
Spanish	spa	60	308031840	0	106	9781	1	1	0	1	2	18	219	12	16	42	150	2583	4226	3123	3339	0	0	1	0	0	0	0	1280301	5562250	9288490	71	4328004	15094	188	0	774766	10443846728.9	282	1	1	44	1	0	0									
Swahili	swh	0	7718895	1	3	1712136	1	1	1	2	5	10	21	1	6	4	43	39	65	0	0	1	0	1	1	1	0	1	34078	81061	1023719	9	28571	69	2017	24	6435	69509111.6735	270	1	1	7	1	0	1									
Tamil	tam	8	68771640	1	6363	914673	1	1	1	1	6	9	17	4	5	20	241	253	0	1	1	1	1	1	0	0	88322	25931	2190620	37	202114	208	8654	32	0	52673	493182075.8033	228	1	1	8	0	0	0										
Togalog	tgl	0	2431000	3	150	823	0	1	1	4	19	41	6	12	38	53	47	84	0	1	0	1	1	1	1	0	1	1	110168	604835	6557104	15	270334	864	31815	232	64634	78544867.971	277	1	1	12	0	1	1									
Thai	tha	40	2030830	1	1160	113962	1	1	1	5	7	23	24	17	28	42	46	127	0	0	1	1	0	1	1	0	1	1	280437	1442559	1850252	27	897283	3311	29583	230	0	141083	1401067290.43	273	1	1	5	0	0	0								
Turkish	tur	0	205801350	1	655	105320	1	1	1	10	10	12	14	4	6	76	83	3607	3618	0	0	0	1	1	1	1	0	0	128695	614484	2645895	7	28083	82	391	61	0	27826	210210658.858	272	1	1	13	0	0	1								
Uzbek	uzb	0	207001230	1	17	325381	0	0	1	1	1	2	2	2	5	2	3	0	0	1	1	1	1	1	1	1	1	1	1148284	3231707	24274551	23	508317	1356	21682	25	0	97310	101317	2	806	10	0	307	0	314	94255893.4103	257	0	1	1	0	0	0
Vietnamese	vie	0	6777830	1	1766	2759176	1	1	1	1	26	26	20	157	172	317	424	94	171	0	1	1	1	1	1	0	0	1055	5271	103137	2	307	0	314	94255893.4103	257	0	1	1	0	0	0												
Wolof	wol	0	397550	4	493	1161518	1	1	1	1	6	7	1	4	18	35	0	1	1	1	1	1	1	0	0	31440	54041	554002	1	14016	33	196	38	0	1058	1078546.258	278	1	1	6	0	1	0											
Yoruba	yor	0	1938080	2	401	822105	1	1	1	2	30	34	5	1	2	20	31	0	1	1	1	1	1	0	0	897850	4751896	4285452	82	22988919	6651	42410	165	0	787	891357.98776	276	1	1	7	0	0	0											
Chinese	chi	0	1185213640	1	19	1046625	1	1	1	1	20	24	3	40	27	26	31	0	0	1	1	1	1	1	0	0	897850	4751896	4285452	82	22988919	6651	42410	165	0	82	1076100.99546	269	1	1	6	0	1	0										
Zulu	zul	157	11869100	1	1714	1046625	1	1	1	1	3	4	6	10	2	3	1	3	0	0	1	1	1	1	1	0	0	776	4030	39055	0	8713	11	0	170	0	82	1076100.99546	269	1	1	6	0	1	0									

Table 1: Overview of digital resources for languages in this report

versions maintained by the American Bible Society, a dictionary in [Leipzig Corpora](#), a collection of Corpus-Based Monolingual Dictionaries, or a parameter file for the language in [TreeTagger](#), a Part of Speech tagger for many languages. [Find-A-Bible](#) all versions is the number of all (printed or online) versions of the Bible listed at [Find-A-Bible](#), a platform for collaboration of the major Bible agencies for making biblical resources available. The last panel show data from the [Endangered Languages Project](#), a catalog of 3346 languages. `langspec` denotes [language specific projects](#), and `lex` stands for multi-lingual (“generic”) [massive dictionary and lexicography projects](#).

Organization of chapters

To the extent feasible, all reports following a uniform structure:

X.1 Demography and ethnography

Here we discuss the **X.1.1 name variants** of the language, its **X.1.2 geographic spread**, the L1 and L2 **X.1.3 speaker populations** (including émigré communities), and the **X.1.4 dialect situation**. In addition to summarizing the data available through direct links to the [Ethnologue](#) and [Wikipedia](#), an effort has been made to provide additional data sources and to incorporate expert opinion.

Some chapters include a so called language cloud by Ethnologue, plots with the horizontal axis corresponding to the EGIDS language status, a manually specified number from 0 (international) to 10 extinct. The vertical axis shows the logarithm of the native speaker population. Logarithm is applied to make the manifold of languages arranged along a straight line. Most languages lie in the middle. Plotting population against EGIDS status shows such differences as Wolof spoken in Senegal and Gambia: the L1 population is an order of magnitude greater in Senegal, but the status is greater in Gambia where the use of Wolof extends to work and mass media (level 3) as opposed to Senegal, where it still has a vigorous use, but only in education (4).

X.2 The main typological and syntactic features

The language is described in standard terms of **X.2.1 linguistic typology**. Special emphasis is put on four syntactic constructions: the **X.2.2 predicative** “X is Y” (the water is dirty); the **X.2.3 possessive** “X’s Y” (*my/his leg* is broken); the **X.2.4 imperative** “do X!” (move the tent!); and the **X.2.5 interrogative** “wh X?” (where is he?).

X.3 Writing system, transcription

X.4 Previous research on the language

Basic introduction and description of the language as provided in school books texts and scholarly papers/monographs, with emphasis on online available resources.

X.5 Data and sources

In addition to the data in each chapter, in this introduction we provide [Table 1](#) that summarizes the digital vitality of reported languages with two families of figures: their EGIDS level in the country

where this level is the highest (an estimate of the overall [development/endangerment status](#) of the language), and the size of their Wikipedia dumps as of 1/1/2016.

X.5.1 Basic vocabulary

X.5.2 Dictionaries

Traditional (paper) dictionaries

Online dictionaries

Each chapter features a survey of the various open and paid lexical resources available to a given language. [Table 2](#) offers a grand-summary of those online resources that a) were freely accessible (at least for non-commercial purposes), b) were not mere wrappers around popular translation engines, and c) were not static contents (e.g., openly available PDF word listings). [Table 2](#) evaluates the sources by referring to two major feature bundles that are revealing about the *quality* and *scrapability* (i.e., harvestability) of the resources.

Quality Signs in the Quality multicolumn indicate whether:

- a page has IPA transcription for the entries (**IPA**)
- part-of-speech labels are available to entries (**POS**)
- transliteration is available to a given entry (**Trnslit**)
- entries are listed in Latin script (**LatSc**)
- audio content is available for words (**Audio**)
- sample sentences are available that illustrate the target words in context (**Context**).

Scrapability To assess the harvestability of various resources, different features were taken into account. Ideal resources feature a word listing, usually sorted by letter, that can be used as a router for a scraper (i.e., an automated algorithm programmed to download predefined contents, e.g., dictionary entries for a given word-pair). In the absence of such a feature, the scraping of a given resource must be executed by searching for terms in its database by relying on another language (e.g., English or German). Due to various causes, the scraping of a resource may not be executed. Some dictionary service providers, for instance, prevent the harvesting of their resources by blocking the access of “notorious” users (i.e., users who send numerous rapid requests to a server) by banning their IP addresses. At other instances, especially in the case of service providers based in economically less advanced countries, scraping might fail due to the poor resources on the server-side (e.g., the server might not be able to deal with rapid requests).

Signs in the Scrapability multicolumn of [Table 2](#) indicate how the various online dictionaries are subject to the factors introduced above:

- **List** — indicates whether the given source offers a word listing.
- **Query** — indicates if an URL-query is possible on the given site.

- **POST** — if a word listing is not available on the given site and an URL-query is not possible, it indicates whether the content sites are downloadable by sending POST requests to the site through an automated tool, such as [cURL](#).
- **Clone** — indicates whether a given site is a clone of another. (In certain cases a clone of another site might be useful when the source site is down or slow; clone sites that offer no performance, or other gains are not listed in the table.)
- **Blocking** — crucially, this column highlights if a blocking occurs due to rapid requests. (Signs indicate whether the site returned errors or failed to respond while executing 50 automated site downloads through [wget](#)).

X.5.3 Corpora

When available, we report the translations of sacred texts and the Universal Declaration of Human Rights as these are appropriate for creating a parallel corpus of moderate size but with paragraph-level alignment.

X.5.4 News portals

X.6 Computational tools

This section introduces the main computational tools developed for the language:

X.6.1 Language identification

X.6.2 Tokenizer

X.6.3 Stemmer

X.6.4 Spell checker

X.6.5 Phrase level and higher tools

X.6.6 End-user support

Language	Dictionary	Developer	Quality						Scrapeability				
			IPA	POS	Trnslit	LatSc	Audio	Context	List	Query	POST	Clone	Blocking
Akan	Twi dictionary	N/A		x		x			x		x		
Akan	Glosbe Dictionary	N/A		x		x				x	N/T	x	
Akan	GhanaWeb	GhanaWeb		x		x				x	N/T		x
Amharic	Amharicdictionary.com	SelamSoft	x	x						x	N/T		
Arabic	Ectaco	ECTACO Inc.		x						x	N/T		x
Arabic	bab.la – Arabic–English dictionary	Andreas Schroeter and Patrick Uecker	x	x	x	x			x	x	N/T		
Arabic	Google Translate	Google Inc.	x				x			x	N/T		N/T
Bengali	Bengali to English Dictionary	BDWord	x						x	x	N/T		
Bengali	English & Bengali Online Dictionary	N/A	x						x	x	N/T		
Bengali	English to Bengali dictionary	Ankur	x						x		N/T		x
Bengali	English Bengali Dictionary online	Glosbe	x			x			x		N/T	x	
Bengali	English Bengali Dictionary	Shyam Krishnan							x		N/T		
Bengali	Shabdkosh	Maneesh Soni	x	x	x	x	x		x	x	N/T		
Bengali	Samsad Bengali–English Dictionary	Shishu Sahitya Samsad	x	x	x				x		N/T		
Bengali	Online Bangla Dictionary	Abdullah Ibne Alam	x						x		N/T		N/T
Bengali	ALDictonary	Adept Leal Software	x							x		N/T	
Hindi	Collins English–Hindi	HarperCollins Publishers	x						x	x	N/T		
Hindi	Hindicube.com	Comsys Technologies Pte. Ltd.							x	x	N/T		
Hindi	HamariWeb.com	Abrar Ahmed		x	x					x	N/T		
Hindi	Universal Word – Hindi Dictionary	Pushpak Bhattacharyya	x	x	x						N/T		
Hindi	Shabdkosh	Maneesh Soni	x		x	x	x		x	x	N/T		
Hindi	HinKhoj Hindi Dictionary	HinKhoj InfoLabs Llp.	x	x	x	x	x		x	x	N/T		
Hindi	Hindi Dictionary	Shyam Krishnan							x		N/T		
Hindi	A Practical Hindi–English Dictionary	Mahandra Caturvedi	x	x	x	x			x	x	N/T		
Hindi	Hindlish.com	Wordtech Co. Ltd.	x	x			x	x	x	x	x	N/T	
Hindi	bab.la – English–Hindi dictionary	Andreas Schroeter and Patrick Uecker	x	x	x				x	x	N/T		
Hindi	English to Hindi dictionary	Zdenek Broz		x	x					x	N/T		
Hindi	ALDictionary	Adept Leal Software	x							x	N/T		
Hungarian	Szótár.net	Akadémiai Kiadó	x		x					x	N/T		
Hungarian	SZTAKI Szótár	MTA SZTAKI	x		x	x			x		N/T		x
Indonesian	Kasmus.net	STANDS4 LLC.	x		x				x	x	N/T		x
Indonesian	bab.la – English–Indonesian dictionary	Andreas Schroeter and Patrick Uecker	x		x	x	x		x	x	N/T		
Indonesian	Kateglo.com	kateglo	x		x					x	N/T		x
Indonesian	Sederet.com	N/A			x					x			
Mandarin	Cambridge English–Mandarin	Cambridge University Press	x				x			x	N/T		
Mandarin	mdbg.net	MDBG		x	x					x	N/T		x
Mandarin	yellowbridge.com	J. Lau	x	x	x						N/T		x
Mandarin	Google Translate	Google Inc.	x	x	x	x	x			x	N/T		N/T
Persian	MyMemory	T-labs								x	N/T		x
Persian	Dictionary–Farsi	Perdic.com								x	N/T		
Persian	Aryanpour	Aryanpour.com								x			
Persian	English Farsi Advanced Dictionary	Farsidics.com					x			x	N/T		N/T
Persian	Ectaco	ECTACO Inc.	x							x	N/T		x
Russian	ABYYY Lingvo	ABYYY Production LLC.					x			x	N/T		
Russian	Rustran	ECTACO Inc.	x							x	N/T		x
Russian	PONS	PONS GmbH	x			x				x	N/T		x
Russian	LEO	LEO GmbH	x		x					x	N/T		x
Russian	bab.la – English–Russian dictionary	Andreas Schroeter and Patrick Uecker	x		x	x	x	x	x	x	N/T		
Russian	LEGO	The LINGUIST list et al.							x		N/T		
Russian	Google Translate	Google Inc.								x	N/T		N/T
Somali	Somali–English–Italian trilingual dictionary	Redsea-online.com cultural foundation		x		x			x	x	N/T		x
Somali	English to Somali dictionary	Zdenek Broz	x	x						x	N/T		
Somali	Freelang	R. B. Figueiredo			x						x		
Somali	WikiQaamus	Wikimedia Inc.	x		x		x		x	N/T	N/T	N/T	N/T
Spanish	bab.la – Spanish–English dictionary	Andreas Schroeter and Patrick Uecker	x		x	x	x		x	x	N/T		
Spanish	Collins English–Spanish	HarperCollins Publishers	x		x	x	x		x	x	N/T		
Spanish	Oxford English–Spanish	Oxford University Press	x		x	x	x	x	x	x	N/T		x
Spanish	Cambridge English–Spanish	Cambridge University Press	x		x			x		x	N/T		
Spanish	Larousse Enlish–Spanish	Isabelle Jeuge-Maynart et al.	x		x	x	x			x	N/T		
Spanish	Google Translate	Google Inc.	x		x	x	x			x	N/T		N/T
Swahili	Tuki English–Swahili Dictionary	University of Dar Es Salaam	x		x				x		N/T		
Swahili	Swahili–English Dictionary	TshwaneDJe	x		x	x		x			x		
Swahili	Swahili–English Dictionary	Nino Vessella			x					x	N/T		
Swahili	English Swahili Dictionary online	Glosbe	x		x					x	N/T	x	

(conted)

Language	Dictionary	Developer	Quality						Scrapability				
			IPA	POS	Trnslit	LatSc	Audio	Context	List	Query	POST	Clone	Blocking
(cont'd)													
Swahili	Swahili–English xfried Dictionary	Morris Fried			x				x		N/T		
Swahili	English to Swahili dictionary	Zdenek Broz			x	x				x	N/T		
Swahili	bab.la – Swahili–English dictionary	Andreas Schroeter and Patrick Uecker			x				x	x	N/T		
Swahili	freedict.com	Parvis			x						x		
Tagalog	Tagalog dictionary	Pinoy Dictionary	x		x		x	x	x	x	N/T		
Tagalog	Tagalog dictionary	N/A	x		x					x	N/T		
Tagalog	LingvoSoft Online	ECTACO Inc.	x		x					x	N/T		x
Tagalog	TagalogTranslate.com	N/A			x				x		N/T		
Tagalog	SEAlang Tagalog dictionary	University of Hawaii	x		x		x						N/T
Tamil	Tamil-cube	Comsys Singapore											
Tamil	English–Tamil–German Dictionary	Tamildict.com	x						x	x	N/T		x
Tamil	ShabdKosh	Maneesh Soni	x	x	x	x	x		x	x	N/T		
Tamil	MyMemory	T-labs								x	N/T		x
Tamil	Google Translate	Google Inc.	x	x	x	x	x			x	N/T		N/T
Thai	Cambridge English–Thai Dictionary	Cambridge University Press	x					x		x	N/T		
Thai	Thai–English Dictionary	National Electronics and Computer Technology Center	x								x		
Thai	English–Thai Bilingual Dictionary	N/A	x	x						x	N/T		x
Thai	LongdoDict	Metamedia Technology	x							x	N/T	x	
Thai	Thai–English Dictionary	Glenn Slayden	x	x	x	x	x	x	x	x	N/T		
Thai	Thai–English dictionary	N/A	x	x	x					x	N/T	x	
Thai	MyMemory	T-labs								x	N/T		x
Thai	Forvo Pronunciation Dictionary	Félix Vela					x		x	x	N/T		
Thai	Thai Wiktionary	Wikimedia Inc.	x	x						N/T	N/T	N/T	N/T
Turkish	bab.la – Turkish–English dictionary	Andreas Schroeter and Patrick Uecker	x		x	x	x		x	x	N/T		
Turkish	Cambridge English–Turkish	Cambridge University Press	x					x		x	N/T		
Turkish	Babylon	Babylon Software Ltd.				x							
Turkish	turkishdictionary.net	N/A			x					x	N/T		
Turkish	Tureng	Tureng Translation Ltd.	x		x					x	N/T		x
Turkish	Ectaco	ECTACO Inc.	x							x	N/T		x
Uzbek	Online Pocket Uzbek Dictionary	N/A	x		x					x	N/T		x
Uzbek	Uzbek Dictionary	N/A	x		x						x		
Uzbek	Uzbek Dictionary	Indiana University	x		x								N/T
Uzbek	Russian–Uzbek online dictionary	AwardSofts	x		x					x	N/T		
Vietnamese	VDict	VDict.com		x	x		x		x	x	N/T		
Vietnemese	Cambridge English–Vietnamese	Cambridge University Press	x			x	x		x	x	N/T		
Vietnemese	Google Translate	Google Inc.	x				x			x	N/T		N/T
Wolof	Afroweb	N/A	x		x				x		N/T		N/T
Wolof	xLingua	Computer Zentrum GmbH	x		x					x	N/T		
Wolof	Freelang	R. B. Figueiredo			x						x		
Wolof	home2.swipnet.se	N/A							x		N/T		N/T
Yoruba	yorubadictionary.com	Pamela Smith and Adebusola Onayemi	x		x				x		N/T		
Yoruba	Freelang	R. B. Figueiredo			x						x		
Zulu	Isizulu	Carsten Gaebler	x		x					x			x
Zulu	Isizulu2	Bilingo			x				x		N/T		x
Zulu	MyMemory	T-labs								x	N/T		x
Zulu	Zulu Wiktionary	Wikimedia Inc.				x			N/T	N/T	N/T		N/T

Table 2: Summary table of the online dictionaries featured in the present volume (in the order they are reported). For more details about the Quality and Scrapability factors indicated in the table, please see the [Online dictionaries](#) section. (Note: scrapability indicators for openly available resources are not evaluated.)

language	name	trial	tokens
Arabic	Arabic Web 2012 (arTenTen12, Stanford tagger)	trial	8,322,097,229
Arabic	OPUS2 Arabic	paying	406,527,277
Arabic	Arabic Web	paying	174,239,600
Arabic	Arabic Web 2012 sample 115M (arTenTen12, Mada tagger)	paying	131,159,731
Arabic	KSUCCA (Classical Arabic)	paying	59,693,146
Arabic	Arabic Learner Corpus (ALC)	paying	386,583
Arabic	Quran annotated corpus [vowelled Latin]	paying	128,243
Arabic	Quran annotated corpus [vowelled Arabic]	paying	128,243
Arabic	Quran annotated corpus [unvowelled Latin]	paying	128,243
Arabic	Quran annotated corpus [unvowelled Arabic]	paying	128,243
Bengali	Bengali Web (BengaliWaC)	trial	13,719,158
Chinese Simplified	Chinese Web 2011 (zhTenTen11)	trial	2,106,661,021
Chinese Simplified	OPUS2 Chinese Simplified	paying	299,338,099
Chinese Simplified	Chinese Web (Internet-ZH)	paying	277,931,664
Chinese Simplified	Chinese GigaWord 2 Corpus: Mainland, simplified	paying	250,124,230
Chinese Simplified	Chinese Web 2011 (zhTenTen11, sample 10M)	paying	11,028,308
Hindi	Hindi Web (HindiWaC)	trial	65,772,188
Hindi	OPUS2 Hindi	paying	1,642,973
Hungarian	Hungarian Web 2012 (huTenTen12)	paying	3,184,161,466
Hungarian	Araneum Hungaricum Maius [2014]	trial	1,200,001,609
Hungarian	EUR-Lex Hungarian 2/2016	paying	499,799,589
Hungarian	OPUS2 Hungarian	paying	218,409,426
Hungarian	DGT, Hungarian	paying	55,276,730
Hungarian	EUROPARL7, Hungarian	trial	14,655,015
Hungarian	CHILDES Hungarian Corpus	paying	311,543
Indonesian	Indonesian Web (IndonesianWaC)	trial	109,281,359
Persian	TalkBank Persian	paying	549,165,952
Persian	OPUS2 Persian	trial	5,367,401
Persian	CHILDES Farsi Corpus	paying	150,505
Russian	Russian Web 2011 (ruTenTen11)	trial	18,280,486,876
Russian	Araneum Russicum Maius [2013]	paying	1,216,800,424
Russian	Araneum Russicum Externum Maius (non-Russia Russian, 15.03) 1,20 G	paying	1,200,053,619
Russian	Araneum Russicum Maius (Russian, 15.02) 1,20 G	paying	1,200,001,911
Russian	Araneum Russicum Russicum Maius (Russia-only Russian, 15.03) 1,20 G	trial	1,200,000,258
Russian	OPUS2 Russian	paying	381,468,257
Russian	Russian web corpus	paying	187,965,822
Russian	CHILDES Russian Corpus	paying	59,759
Spanish	Spanish Web 2011 (esTenTen11, Eu + Am, Freeling v4)	trial	10,994,616,207
Spanish	American Spanish Web 2011 (esamTenTen11, Freeling v4)	trial	8,640,399,540
Spanish	European Spanish Web 2011 (eseuTenTen11, Freeling v4)	trial	2,354,216,667
Spanish	Araneum Hispanicum Maius [2013]	paying	1,200,000,609
Spanish	OPUS2 Spanish	paying	870,615,999
Spanish	EUR-Lex Spanish 2/2016	paying	836,039,928
Spanish	Spanish web corpus	paying	116,900,060
Spanish	esTenTen [2011, Eu + Am, Freeling v4, sample]	paying	73,597,801
Spanish	DGT, Spanish	paying	68,721,827
Spanish	EUROPARL7, Spanish	trial	60,862,330
Spanish	CHILDES Spanish Corpus	paying	1,358,475
Tamil	Tamil Web (TamilWaC)	trial	32,861,569
Tamil	CHILDES Tamil Corpus	paying	21,865
Thai	Thai Web (ThaiWaC)	trial	108,013,897
Thai	CHILDES Thai Corpus	paying	299,962
Turkish	Turkish Web 2012 (trTenTen12)	trial	4,124,558,200
Turkish	OPUS2 Turkish	paying	207,223,730
Turkish	TurkishWaC	paying	40,539,507
Turkish	CHILDES Turkish Corpus	paying	233,097
Uzbek	Turkic web – Uzbek	trial	24,570,516
Vietnamese	Vietnamese Web Corpus (VietnameseWaC)	trial	129,781,089
Yoruba	Yoruba WaC [2015]	trial	3,500,353

Contents

1 Akan (Nikolett Mus)	1
2 Amharic (Levente Madarász)	17
3 Arabic (Nikolett Mus)	29
4 Bengali (Levente Madarász)	47
5 Hindi (Levente Madarász)	67
6 Hungarian (László Kálmán and András Kornai)	83
7 Indonesian (Dávid András Imre)	97
8 Mandarin (Dávid András Imre)	111
9 Persian (Nikolett Mus)	127
10 Russian (Nikolett Mus)	139
11 Somali (Levente Madarász)	153
12 Spanish (Zsuzsanna Bárkányi)	169
13 Swahili (Levente Madarász)	185
14 Tagalog (Noémi Vadász)	203
15 Tamil (Nikolett Mus)	221
16 Thai (Ekaterina Georgieva)	233
17 Turkish (Noémi Vadász)	253
18 Uzbek (Nikolett Mus)	271
19 Vietnamese (Imre Dávid András)	285
20 Wolof (Noémi Vadász)	297
21 Yoruba (Noémi Vadász)	317
22 Zulu (Ekaterina Georgieva)	337

Chapter 1

Akan (Nikolett Mus)

Contents

1.1 Demography and ethnography	1
1.2 Main typological and syntactic features	5
1.3 Writing system, transcription	8
1.4 Previous research on the language	9
1.5 Data and sources	10
1.6 Computational tools	12
Bibliography	14

Introduction

This chapter reviews some aspects of the Akan language, a member of the Niger-Congo language family, mainly spoken in Ghana. Section 1.1 offers an ethno-linguistic outline of Akan and its speaking community, which is followed by an overview of the linguistic typology of the language (see 1.2), and information about its writing system and linguistic transcriptions (see 1.3). After reporting on the previous research concerning the grammatical structure of Akan (see 1.4), the available offline and online linguistic data sources will be introduced (see 1.5). The final Section addresses the availability of various computational tools for the processing of the language (see 1.6).

1.1 Demography and ethnography

1.1.1 Name variants

Akan is a Central Tano language, which belongs to the Kwa branch of the Niger-Congo language family. The name *Akan* is used to denote a macrolanguage, i.e. a set of individual languages that are related to each other and their speakers share a common ethnolinguistic identity even though they do not understand each other (for a detailed definition and further examples of the macrolanguages consult the site called the [Scope of denotation for language identifiers](#) of SIL). The Akan macrolanguage thus subsumes the Twi, and the Fante languages. Twi can be further divided into two dialectal groups which are the Asante (or Asante Twi) and the Akwapem (or Akwapem Twi) groups. So, the three main dialectal variations of Akan are the Asante, the Akwapem, and the Fante languages. These languages are the basis on which the literary standard of Akan has been created. The corresponding

entries of Wikipedia, i.e. the [Akan language](#), the [Twi language](#), and the [Fante language](#) provide more information about Akan.

The ISO 639-3 code of the Akan macrolanguage is **aka**. The ISO code used for Twi is **twi**, and the code **fat** belongs to Fante (cf. the [Documentation for ISO 639 identifier: aka](#)).

1.1.2 Geographic spread

The Akan language is spoken in Ghana, in a subregion of West Africa along the Gulf of Guinea and the Atlantic Ocean. Ethnic groups of Akan can be found in the southeastern part of the country. The Akan people accounts for the 58% of the population in Ghana (cf.). The map (1.1) illustrates the linguistic diversity of Ghana (cf. the [Languages of Ghana](#) Wikipedia entry).

Additionally, the language is also spoken in Cote d'Ivoire (or Ivory Coast). The country is located in West Africa. The Akan speakers represent the 30-40% of the total amount of the population, which means c. 8 million Akan speaker in Cote d'Ivoire (cf. the [Culture of Ivory Coast](#) Wikipedia entry). The map (1.2) illustrates the Akan speaking territories of Cote d'Ivoire.

According to the [Akan entry of the Joshua Project](#) there are Akan speakers outside of Africa too. For instance, there are 26,000 speakers who live in the United Kingdom. Further c. 22,000 speakers can be found in the northern countries of Europe, such as in Norway, Finland, Denmark and the Netherlands. Additionally, c. 77,000 Akan speaker lives in the United States and Canada.

1.1.3 Speaker populations

According to the [Akan language](#) entry of Wikipedia, there are c. 11 million people who speak Akan as their mother tongue. Besides, c. 1 million people use the language as their secondary language (L2) in the area.

In contrast, the [Akan Ethnologue entry](#) estimates 8,186,600 L1 and 1,000,000 L2 Akan speakers.

Finally, [Abakah \(2004\)](#) reports on the speaker population and dialectal situation of Akan (see the PDF). Based on [Abakah \(2004\)](#), the estimated number of L1 speakers is approximately 20 million, which is c. the 44% of Ghana's population.

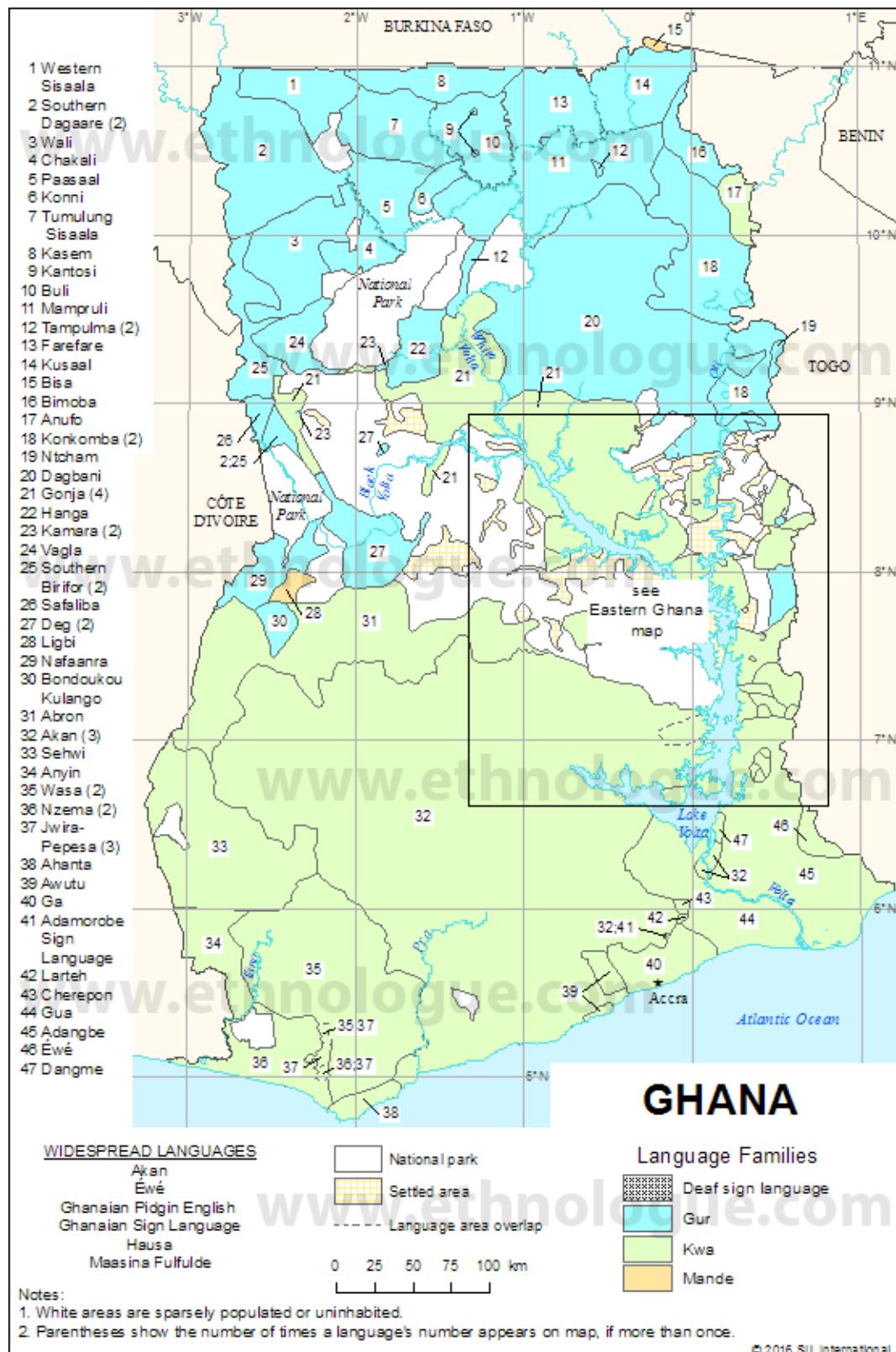
According to [Ethnologue](#), the EGIDS (Expanded Graded Intergenerational Disruption Scale) level for Akan is **3** in Ghana. It means that the language is used in work and mass media without being an official language of the country, i.e. Ghana. [Figure 1.3](#) shows the position of the Akan language within the cloud of the languages of the world, see the explanation in [section](#).

Although a highly significant amount of the population speaks Akan in Cote d'Ivoire, no data is available concerning the EGIDS level of the language in this country. However, the official language of Cote d'Ivoire is French and it is also the *lingua franca* in the country (given that there are 83 different languages spoken in the country). It is then implicitly assumed that the EGIDS level cannot be 0, 1 or 2, as these values are given to the official or widely used languages.

1.1.4 Dialect situation

While both **Fante** and **Twi** belong to the Akan macrolanguage, they also have further (sub)dialects. The Agona, Anomabo Fanti, Abura Fanti and Gomua dialects constitute the Fante language (see the [Fante dialect](#) of Wikipedia).

As mentioned, the Twi language can be divided into two dialects, which are Akuapem and Asante. The Akuapem was the first dialect that was used for Bible translation, therefore became the prestige

Figure 1.1: The linguistic diversity of Ghana (source: [Ethnologue](http://www.ethnologue.com))

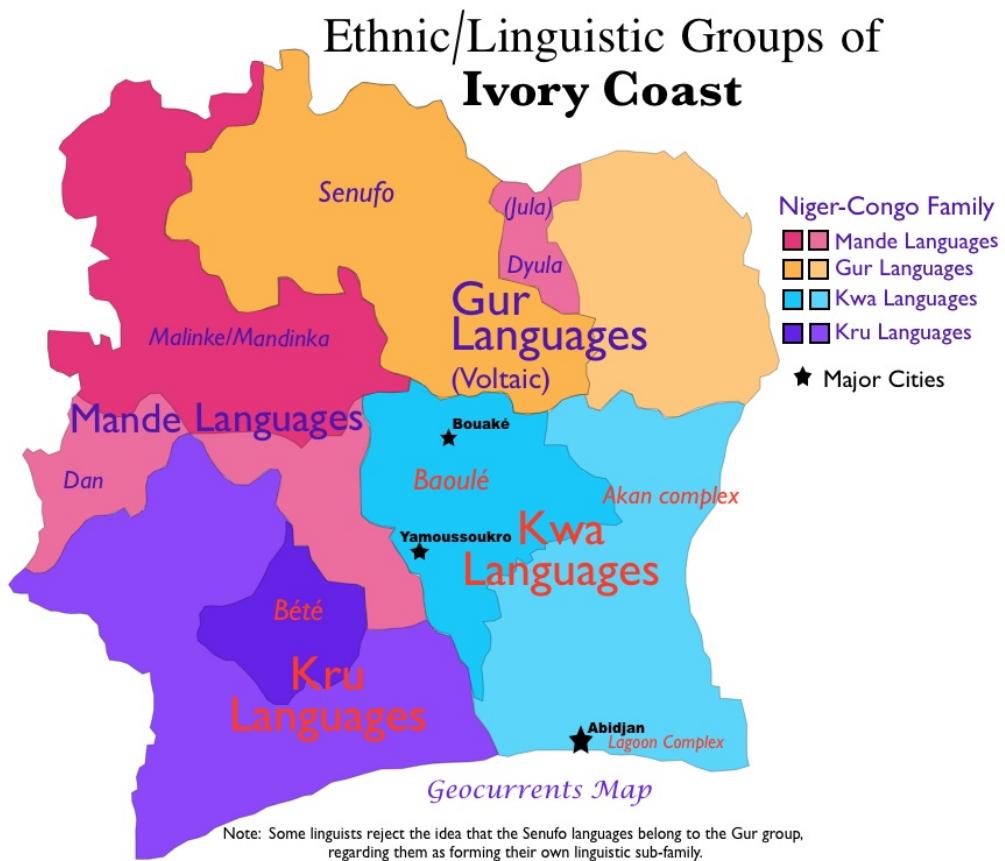


Figure 1.2: The linguistic diversity of Côte d'Ivoire (source: [Geocurrents](#))

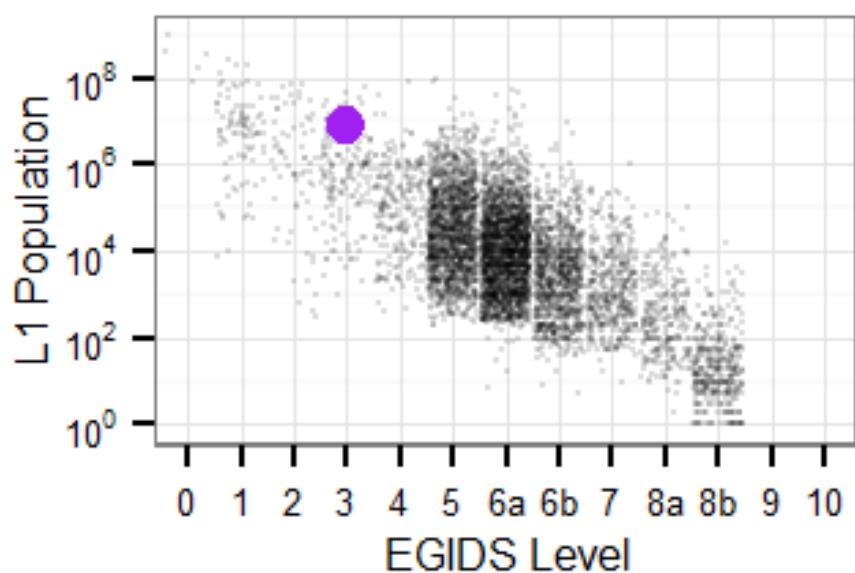


Figure 1.3: The EGIDS level for Akan (source: [Ethnologue](#))

dialect (see the Wikipedia entry called [Asante dialect](#)). Additionally, the Ahafo, the Akyem, the Asen, the Dankyira and the Kwawu dialects belong to the Twi language.

1.2 Main typological and syntactic features

1.2.1 Linguistic typology

This Subsection is based on the parameters identified and collected by the team of WALS. For further description see [the Akan entries of WALS](#).

Phonological level According to [Brown and Ogilvie \(2010\)](#), the phoneme inventory of Akan contains 15 consonants (p, t, k, b, d, g, m, n, f, s, h, w, l, r, and y) and 9 oral vowels (/e ɔa o ɔi u iʊ/) and 5 nasal ones (/i ì ə ù ù/). In the Asante and the Akupaem dialects there is an additional oral vowel: /œ/. The vowels show tongue root vowel harmony. In the Fante dialects, the vowels pattern according to rounding harmony as well (cf. [TypeCraft Tools for Akan](#) and [Phonology of Akan](#) on Wikipedia).

As already mentioned Akan belongs to the Niger–Congo language family which is the largest language group of Africa. The vast majority of languages of this family are tonal such as Akan (see the [Niger Congo Languages on the Languages of Africa](#)). Every syllable carries contrastive tone with two levels (high and low) ([Brown and Ogilvie, 2010](#)).

Morphological level In Akan the grammatical relations are mainly expressed by prefixes. It is only the possessive paradigm in which the nouns take suffixes. Besides, Akan employs the strategy of (productive full and partial) reduplication as a grammatical device.

Morphosyntactic level The Akan nouns are specified for number, i.e. singular and plural number that is expressed by prefixes. As mentioned, in the possessive paradigm they can take person markers. In the singular possessive paradigm the suffix *-ni* is attached to the noun, and in the plural paradigm the noun takes the *-fo* suffix. The adjectives functioning as modifiers agree with the head noun in number, so they take number markers. The verbs are inflected for aspects and there is a future marker.

Syntactic level The basic word order of the language is SVO, illustrated in the following example.

- (1) *Kofi hu-u Am(m)a.*

Kofi see-REM.PST Amma

‘Kofi saw Amma.’ ([Ofori, 2006](#), p. 12)

The Akan language is a head initial language, which means that the dependents usually follow their heads in the phrases. Consequently, the elements occur in the following orders: NAdj, NDem, NNum, AuxV. The only exception is the possessive phrase, in which the dependent appears before the head, i.e. GenN. For a detailed description about the structure of Akan from a typological point of view see the corresponding Chapters of [wals](#)). Additionally, the Typecraft provides a [Typological Features Template for Akan](#).

1.2.2 Predication

The verbs functioning as predicates in clauses can get inflectional morphemes, e.g. tense, aspect, etc. expressed either by prefixes or by suffixes attached to the verb.

- (2) *Kofi a-ba.*

Kofi REC.PT-COME

'Kofi has come.' (Ofori, 2006, p. 10)

- (3) *Kofi ba-i.*

Kofi come-REM.PST

'Kofi came.' (Ofori, 2006, p. 10)

The nominals, i.e. the nouns and adjectives, functioning as predicates always require an overt copula in the construction:

- (4) *ɔ-yεɔbarima.*

3SG-BE

'He is a man.' (Ellis and Boadi, 1969, p. 18)

- (5) *Kofi a-yεfi.*

Kofi PERF-BE

'Kofi was dirty.' (Osam, 1993, p. 159)

1.2.3 Possession

In the adnominal possessive structures, the possessor precedes the possessed item. The possessor may be either a pronominal or a lexical one.

- (6) *me ho*

POSS1SG body

'my body'

- (7) *Amma sika*

Amma money

'Amma's money' (Brown and Ogilvie, 2010, p. 19)

In the predicational possessive structures a so-called *light* verb is used. This light verb differs from the regular copula used in the nominal predicates. In the possessive clauses, the possessed NP is the grammatical subject of the existential predicate. By contrast, the possessor NP is construed as the topic of the sentence that indicates the setting of the clause.

- (8) *Me wo wodan bi.*
 1SG be.at house one
 ‘I have a house.’ (Christaller, 1875, p. 66)

1.2.4 Imperative

Imperative is one of the two moods that can be expressed in Akan (the other one is the indicative mood). Within the imperative modal system an imperative proper and an optative can be differentiated. The imperative proper is not marked morphologically.

- (9) *Kɔ!*
 go.IMP
 ‘Go!’ (Osam, 2003, p. 13)

In contrast, the optative involves the presence of a subject marker and a high tone homorganic nasal prefix on the verb.

- (10) *Mo-n-da!*
 2PL.SUBJ-OPT-SLEEP
 ‘You (should) sleep.!’ (Osam, 2003, p. 14)

The prohibitives are expressed by the combination of the normal imperative construction and the normal negative structure, i.e. negative affixes.

- (11) *N-kɔ!*
 NEG-GO
 ‘Don’t go!’ (Osam, 2003, p. 13)

- (12) *Mo n-n-da!*
 2PL.SUBJ NEG-OPT-SLEEP
 ‘You shouldn’t sleep.!’ (Osam, 2003, p. 14)

In Fante, the negative imperative is preceded by the negative form of the verb *ma* ‘let’ (cf. Osam, 2003).

- (13) *M-ma n-kɔ!*
 NEG-LET NEG-GO
 ‘Don’t go!’ (Osam, 2003, p. 13)

1.2.5 Interrogative

In polar interrogatives, two question particles are used to mark the interrogativity: *ana* and *aso*. The question particle *ana* occurs at the sentence final position, while *aso* appears initially.

- (14) *Yaw a-da ana.*
Yaw PERF-SLEEP QM

‘Has Yaw slept?’ (Marfo, 2005, p. 32)

- (15) *Aso Yaw a-da?*
QM Yaw PERF-SLEEP

‘Has Yaw slept?’ (Marfo, 2005, p. 159)

Additionally, it can only be the intonation of the polar interrogatives that shows systematic differences from those of the respective declaratives. In polar interrogatives, indeed, the pitch level of the final tone of the sentence-final word rises and falls steadily (cf. Marfo, 2005).

In wh-interrogatives, the interrogative phrase either remains *in-situ* or it appears in the sentence initial position. The canonical position for wh-words in Akan is their *in-situ* position, i.e. they remain in the same position within the clause in which a non-question word fulfilling the same grammatical function is located.

- (16) *Baa re-sere hwan?*
Baah PROG-LAUGH who

‘Baagh is laughing at who?’ (Marfo, 2005, p. 119)

If they are fronted, they are left-dislocated outside from the clause and a clitic morpheme, i.e. a focus marker (*na*) appears at the right edge of the dislocated wh-word.

- (17) *Hwan na Baa re-sere no?*
who FOC Baah PROG-LAUGH 3SG

‘Whom is Baagh laughing at?’ (Marfo, 2005, p. 119)

1.3 Writing system, transcription

Although the earliest recordings of the Akan language were taken by missionaries as early as the 17th and 18th centuries, a unified orthography has only been created during the 1980s by the *Akan Orthography Committee*. This unified orthography is based on the Latin script and is primarily used in schools at the lower level (Primary 1-3). The literacy rate is 30-60 % among speakers who speak the language as their mother tongue. This rate is decreased to 5-10 % among L2 speakers (cf. [Ethnologue report for language code: aka](#)).

Additionally, there are three standardized orthographies for Asante, Akupaem and Fante (cf. [Om-niglot](#)). These standards are mutually intelligible, however, the speakers of the other standards usually do not accept the other forms.

Vowels									
a	e	e	i	o	o	u	ɛ	ɔ	
[a]	[i]	[ɛ]	[ɪ]	[ʊ]	[ɔ]	[u]	[ɛ]	[ɔ]	

Consonants										
b	d	dw	dwi	f	g	gw	gyi	h	hw	hwi
[b]	[d]	[dʒ]	[dʒwɪ]	[f]	[g]	[gʷ]	[dʒi~ɣi]	[h]	[hʷ]	[ɣʷɪ]
hyi	k	kw	kyi	l	m	n	ng	ngi	nw	nwi
[çɪ]	[kʰ]	[kʷ]	[tɕʰi~ççʰɪ]	[l]	[m]	[n, n̥, n̥]	[ɳ:]	[ɲi]	[ɳɳʷ]	[ɲɳ̩ɪ]
nyi/nnyi	p	r	s	t	ti	twi	w	wi		
[n̥i]	[pʰ]	[r, r̥, t̥]	[s]	[t̥]	[tɕi]	[tɕʷi]	[w]	[ɥi]		

Figure 1.4: The Akan alphabet and pronunciation

In (scientific) publications, the tones, that can be high, mid, and low, are usually marked by acutes and graves.

Figure 1.4 illustrates the Akan alphabet and the pronunciation of the phonemes. The Akan character chart is provided by the [Akan Omniglot](#) side.

1.4 Previous research on the language

The earliest descriptive grammars of Akan are provided by Christaller (1875); Balmer and Grant (1929); Welmers (1946). The sound system and tonal structure of the language is discussed by Dolphyne (1988) in detail. The structure of Akan is described by Osam (2004). Additionally, papers on different aspects of the languages of Ghana are gathered and edited by Trutenuau (1976). For a more detailed bibliography, see the OLAC entries for [Fanti](#), [Akan](#) and [Twi](#).

Furthermore, there are (ongoing) projects that attempt to describe and document the language. For instance, the [300 Languages](#) subproject of The Rosetta Project aims at archiving Akan, as well as, constructing a universal corpus of human language by collecting parallel texts and audio recordings. The project collects recordings and translations of *The Swadesh List*, *The Universal Declaration of Human Rights* and the *Genesis* (Chapters 1-3). The data are available via [The Internet Archive](#).

Besides, there is the project of the [Kasahorow Foundation](#) in Ghana, which aims at modernizing the African languages.

Research and control bodies The African Studies and Linguistics departments of the [Ohio University](#) offer possibilities and supports in Akan linguistic and anthropological research. Furthermore, linguistic research concerning several aspects of Akan (including language teaching, language and business, language and society, language policy, language and the deaf, language and culture, language and law, language and politics etc.) is carried out at the [Department of Linguistics at the University of Ghana](#).

1.5 Data and sources

This section introduces some of the available sources of the Akan language. The sources contain vocabularies, dictionaries (both paper editions and online ones), and texts. Additionally, an overview of the most prominent news portals of Akan is provided here.

1.5.1 Basic vocabulary

The [An Crubadan](#) project also provides Akan sources (such as character trigrams, word bigrams, word frequency lists, as well as, URLs of Akan pages) based on collected texts containing 367,192 words. Additionally, there is an online word list available on the homepage of the Ghana Magazine (see [Twi dictionary & translator](#)). Besides, the UCLA Phonetics Laboratory provides material, i.e. recordings, word lists, phrasebooks, etc. that are gathered for phonetic and phonological research purposes. The data are available online in the [Archive](#) of the UCLA Phonetics Lab. Furthermore, the following online word lists are available:

- the [Akan Vocabulary List](#) (a thematized word list available in a digitized format)
- the [Stanford-Twi sample](#) -parallel word list in English and Akan, and a preliminary (basic) phrasebook organized by different topics, such as family relations, types of food, etc.
- the [Akan-Twi Religion Vocabulary](#) provided by [memrise](#). (The webpage requires user account.)
- the [Akan Grammar Vocabulary](#) by [cram.com](#)
- a basic vocabulary by [ofmtv.com](#) which provides parallel word lists in English and Akan (Twi) organized by basic topics, for instance family relations, cultural terms and customs, animals, greetings, etc. and audio files for supporting language learners (i.e. pronunciation)
- support materials for learners by the [UCLA Language Materials Project](#)
- the [Learn foreign languages](#) portal provides possibility to learn the target language via an online learning partner exchange program

1.5.2 Dictionaries

Traditional (paper) dictionaries

Printed dictionaires of Akan have been edited recently within the frame of the Kasahorow project:

- the *Akan Learner's Dictionary: Akan-English & English-Akan* (edited in 2012, 174 pp.) which contains word entries and examples in Akuapem, Fanti and Twi dialects, as well as the POS tags and the English translations of the words;
- the *Modern Akan Dictionary: Akan-English Dictionary: Akan-English & English-Akan* (edited in 2012, 130 pp.) consisting of example sentences for each entry with English translations and POS tags;
- the *My First Akan Dictionary: Colour and Learn* (edited in 2016, 70 pp.) which is a picture book for introducing Akan for children;

Additionally, [A dictionary of the Asante and Fante language called Tshi \(Chwee, Twi\)](#) : with a grammatical introduction and appendices on the geography of the Gold Coast and other subjects, which is basically a paper-like dictionary (in the sense, that it cannot be searched via searching engines) edited by J. G. Christaller between 1827–1895 (716 pp.). The dictionary is uploaded to the [archive.org](#), therefore it is available on various digital formats.

Online dictionaries

The following English-Akan online dictionaires are available:

- The [Kasahorow online dictionary](#) is developed by the Kasahorow Foundation. This source is not appropriate for extracting Akan texts/data.
- The [Twi dictionary](#) contains 14,663 entries with English translations and POS tags.
- The [Glosbe Dictionary](#) contains the Akan entries, their POS categories and it also provides examples of usage. This source is useful for web scraping.
- The [GhanaWeb](#) translates basic clauses and provides additional basic grammatical informations: POS category of the words and the forms of the stems. Extracting the core content, i.e. the Akan data, from this site is not easily possible, i.e. this site is not scrapeable.

1.5.3 Corpora

Monolingual corpora

As of 29/09/2016, the Akan [wikipedia](#) contains 257 articles. For Akan monolingual texts see also the corresponding entries of [Indigenous tweets](#). There are further Akan translations available:

- the *Bible* is translated into [Twi](#) and [Asante](#)
- the full [twi translation](#) of *The book of Mormon* is also available
- there is an Asante translation of the [Quran](#)
- the *Universal Declaration of Human Rights* is translated into [Twi](#) and [Asante](#)

Bilingual corpora

[Typecraft](#) provides an Akan corpus consisting of 80,893 words of which 9,347 is assigned with POS tags. The annotators were linguistic students. The data and the annotation are freely available.

1.5.4 News portals

- Radio stations: Stations owned by the [Ghana Broadcasting Corporation](#).
 - [Radio 1 \(Ghana\)](#): a public radio station in Accra. The station broadcasts in English and other Ghanaian languages including Akan, Dagbani, Ewe, Ga, Hausa and Nzema. (Online broadcast available.)
 - Twin City Radio: a public radio station in Sekondi-Takoradi. The radio station broadcasts in English, Fante and Nzema languages. (Aerial broadcasting only.)

- Radio Central: a public radio station in Cape Coast which broadcasts in English and Fante. (Aerial broadcasting only.)
- Garden City Radio: a public radio station in Kumasi. Broadcasting in English and Twi. (Aerial broadcasting only.)
- Sunrise FM (Ghana): a public radio station in Koforidua, English and Twi. (Aerial broadcasting only.)
- Radio BAR: public radio station in Sunyani; English and Twi. (Aerial broadcasting only.)
- [Peace FM](#): a private radio station based in Accra. (Online broadcast available.)
- Television broadcasters/stations:
 - First Digital TV includes more TV channels, such as the Amansan Television.
 - TV Africa is a privately owned free to air television station (Languages: Ga, Twi, Hausa, English (UK)).
 - [OFM TV](#) broadcasts online.
- Online portals:
 - the Citifmonline portal provides [Regional News](#) in English.
 - according to [Alexa](#), the [GhanaWeb](#) is the 5th most commonly visited website in Ghana. (However, the news and articles are in English.)

1.6 Computational tools

This section summarizes the available computational tools developed for the Akan language. There is a comprehensive description by [Osborn \(2010\)](#) of Akan and other African languages, which provides detailed information about the language from computational point of view (see [PDF](#)).

1.6.1 Language identification

The [Compact Language Detector 2](#) provides full support for Akan.

1.6.2 Tokenizer

The Northwestern University Information Technology released the [MorphAdorner V2.0](#), which provides methods for tokenizing Akan texts. Furthermore, the [tools](#) developed by MarkLogic provide basic language support (language-specific tokenization, stemming) for Akan.

1.6.3 Stemmer

As of 29/09/2016, there is no stemmer for Akan available. Nevertheless, the [MorphAdorner V2.0](#) can be used for lemmatization.

1.6.4 Spell checker

An open source [HunSpell](#) spell checker for Akan is available. In addition, the Mozilla dictionary includes Akan [spell checker](#). Besides, [Aspell 0.60](#) supports Akan.

1.6.5 Phrase level and higher tools

Akan morphological analyzer The [TypeCraft](#) online application consists of a database and a linguistic editor for interlinear glossing of the Akan language.

Akan part-of-speech tagger There is a [POS tagger](#) developed within the frame of the An Gramadoir project, which aims at developing a grammar checker.

Akan chunker As of 10/01/2017, there is no chunker for Akan available.

Akan named entity recognizer As of 10/01/2017, there is no named entity recognizer for Akan available.

Akan sentence parser The [XLFG5](#) is a Lexical-Functional Grammar Parser device available online. [Jones \(2014, p. 92\)](#) reports on testing the device on Akan.

Akan speech recognizer [Cho et al.](#) describes the results of building a complete text-to-speech system of the Akan language, using Microsoft language technology.

Akan machine translator As of 10/01/2017, there is no named entity recognizer for Akan available.

Akan question answering machine As of 10/01/2017, no question answering system has been developed for the Akan language.

1.6.6 End-user support

The OS supports for the Akan language are the following:

- The Akan language is not supported by Mac OS X
- Microsoft Windows language pack is not available in Akan.
- There is no language pack of Linux for Akan.

There is a language pack of [Firefox for Akan](#).

There is a keyboard developed by the Kasahorow Foundation for Android phones that allows typing in Akan. (The blog called the [mightyafriican](#) discusses the set up and use of this keyboard in detail.) Furthermore, a [virtual keyboard for Akan provided by gate2home](#) is also available. In addition, there is a [Nkyea Keyboard](#) for iPhones and iPads available from iOS8 that also includes a nifty built-in English-to-Twi word translator.

Furthermore, there is no separate Unicode range for Akan as its writing system relies on no special characters. Therefore, it is situated in the standard Latin Unicode range.

Bibliography

Emmanuel Nicholas Abakah. Elision in fante. *Africa & Asia*, 4(1):181–213, 2004.

W. Balmer and F. Grant. *A Grammar of the Fante-Akan Language*. The Atlantic Press, London, 1929.

Keith Brown and Sarah Ogilvie. *Concise encyclopedia of languages of the world*. Elsevier, 2010.

Hyong Sil Cho, Gifty Akuamoah, Daan Baldewijns, Sara Candeias, Cristiano Chesi, Kofi Agyekum, and Miguel Sales Dias. A phone set of asante-twi defined in ipa and x-sampa.

J. G. Christaller. *A Grammar of the Asante and Fante Language called Tshi (Chwee, Twi). Gold Coast: Basel German Evangelical Mission*. Gregg Press, New Jersey, 1875.

Florence Abena Dolphyne. *The Akan (Twi-Fante) language: Its sound systems and tonal structure*. Ghana Universities Press, 1988.

Jeffrey Ellis and Lawrence Boadi. ‘to be’ in twi. In *The Verb ‘Be’ and its Synonyms*, pages 1–71. Springer, 1969.

Mari C Jones. *Endangered languages and new technologies*. Cambridge University Press, 2014.

Charles Ofosu Marfo. *Aspects of Akan grammar and the phonology-syntax interface*. PhD thesis, University of Hong Kong, Hong Kong, 6 2005.

Seth Antwi Ofori. *Topics in Akan grammar*. ProQuest, 2006.

E Kweku Osam. *An introduction to the structure of Akan: its verbal and multiverbal systems*. Department of Linguistics, University of Ghana, 2004.

Emmanuel Kweku Osam. Animacy distinctions in akan grammar. *Studies in the Linguistic Sciences*, 23(2):153–164, 1993.

Emmanuel Kweku Osam. An introduction to the verbal and multi-verbal system of akan. In *Proceedings of the workshop on Multi-Verb Constructions Trondheim Summer School*, 2003.

D. Osborn. *African languages in a digital age. Challenges and opportunities for indigenous language computing*. HSRC-IDRC, Cape Town and Ottawa, 2010.

H Max J Trutenau. *Languages of the Akan area: papers in Western Kwa linguistics and on the linguistic geography of the area of ancient Begho*, volume 14. Basler Afrika Bibliographien, 1976.

W. E. Welmers. *A Descriptive Grammar of Fanti*. Linguistic Society of America, Philadelphia, 1946.

Chapter 2

Amharic (Levente Madarász)

Contents

2.1 Demography and ethnography	17
2.2 Main typological and syntactic features	20
2.3 Writing system, transcriptions	21
2.4 Previous research on the language	22
2.5 Data and sources	22
2.6 Computational tools	23
Bibliography	25

Introduction

The present chapter aims at providing an overview of the Amharic language, mainly spoken in the Amhara region and Addis Ababa, Ethiopia. The first part will present the reader with the historical roots and the dialectal extension of the language, which is followed by a round-up on the writing system, the reference grammars and the different available corpora. The closing section will detail the various digital resources and the available computational tools for the language.

2.1 Demography and ethnography

2.1.1 Name variants

According to [Minahan \(2013\)](#), Amhara are a Semitic people descended from early Semites who migrated to the region from the Arabian Peninsula around 700 B.C. *Amarəñña* (/amarɪn:a/; አማራና), the people’s endonym stands for “Mountain People” and is adopted in English as an exonym denoting Amara people. Besides *Amharic*, the language of Amaras is also known as *Abyssinian*, *Amarigna*, *Amarinya*, *Amhara* and *Ethiopian*, identified with the *am* ISO 639-1, *amh* ISO 639-2, and ISO 639-3 codes ([Paul et al., 2015b](#)).

2.1.2 Geographic spread

Figure 2.1 provides a visual overview of the major Amharic speaking geographic areas. As for its condition, Amharic on the Expanded Graded Intergenerational Disruption Scale (EGIDS) is a first level, national language in Ethiopia, which means that it is widely used in education, work, mass media,

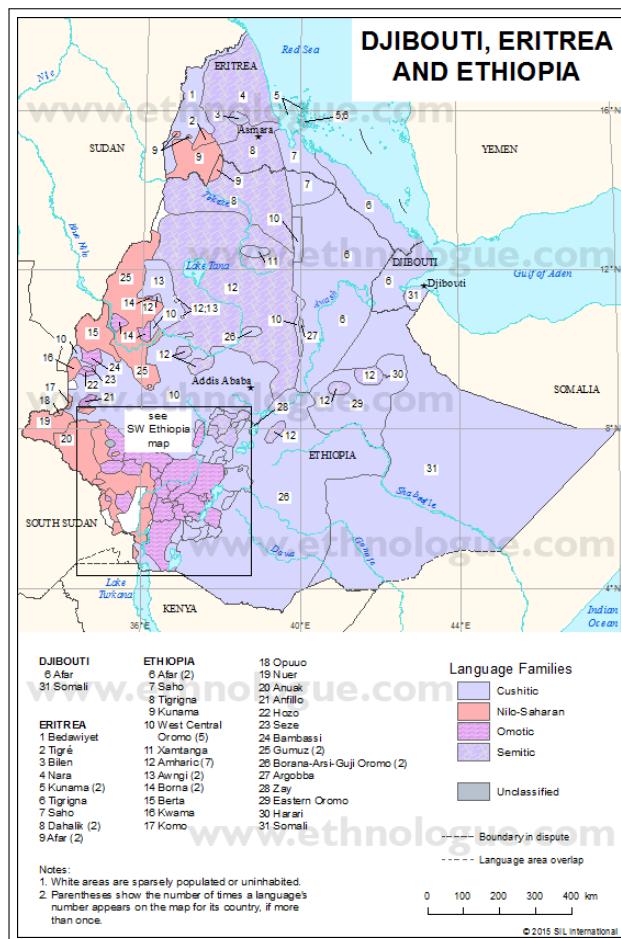


Figure 2.1: Areas of Amharic (Paul et al., 2015b).

and government at the national level (in the 1994 Ethiopian Constitution, Amharic is declared to be the working language of the Federal Democratic Republic of Ethiopia (Gazeta, 1994)). In addition, Amharic is an EGIDS level 5, dispersed language in Israel, where it is referred to as *Falasha* (pejorative) (Paul et al., 2015a). Amharic in Israel is primarily spoken in the Jerusalem district, in Netanya, Rehovot, Rishon LeZion and Petah Tikva belonging to the HaMerkaz (or Central) administrative district, Ashdod, Beersheba, Ashkelon and Kiryat Malachi in the HaDarom (Southern) administrative district, as well as Haifa and Hadera in the Hefa administrative district. Amharic is also used in other countries (see speaker populations below), but its condition in these areas is not documented.

2.1.3 Speaker population

According to Paul et al. (2015a), the Amharic speaking community consists of 21,811,600 speakers, 21,600,000 of whom resides in Ethiopia. In all areas, it is spoken by 14,800,000 monolingual speakers, but it is thriving as an L2 language as well (4,000,000 L2 speakers in Ethiopia). The Joshua Project (2014) cites somewhat different numbers (for a comparison, see Table 2.1): their data reports 29,333,100 Amharic speakers in total, but does not differentiate between L1 and L2 users.

2.1.4 Dialect situation

There are several regional variants of Amharic (although some of their current status is unknown as the latest comprehensive overview of these dates back to the late 1960s): the *Addis Abää* and *Shewa*

country	# L1 (JP)	# L1 (Eth)	EGIDS level	L1 ratio (JP)	L1 ratio (Eth)	# L2 (Eth)
Ethiopia	28,930,600	21,600,000	1	29.09	21.72	4,000,000
USA	130,000	N/A	N/A	0.04	N/A	N/A
Sudan	102,000	N/A	N/A	0.25	N/A	N/A
Israel	72,700	40,000	5	0.86	0.47	N/A
Eritrea	42,000	N/A	N/A	0.66	N/A	N/A
Yemen	11,000	N/A	N/A	0.04	N/A	N/A
Sweden	9,700	N/A	N/A	0.10	N/A	N/A
Italy	6,900	N/A	N/A	0.01	N/A	N/A
Egypt	6,000	N/A	N/A	0.01	N/A	N/A
Germany	5,700	N/A	N/A	0.01	N/A	N/A
Djibouti	4,600	N/A	N/A	0.57	N/A	N/A
Norway	3,400	N/A	N/A	0.07	N/A	N/A
Denmark	2,000	N/A	N/A	0.04	N/A	N/A
New Zealand	1,200	N/A	N/A	0.03	N/A	N/A
Netherlands	1,100	N/A	N/A	0.01	N/A	N/A
Belgium	1,000	N/A	N/A	0.01	N/A	N/A
Finland	1,000	N/A	N/A	0.02	N/A	N/A
Spain	800	N/A	N/A	0.00	N/A	N/A
Canada	700	N/A	N/A	0.00	N/A	N/A
Somalia	700	N/A	N/A	0.01	N/A	N/A
world	29,333,100	21,640,000				4,000,000

Table 2.1: Summary of Amharic speaker populations, based on data obtained from the Joshua Project (2014) (JP) and the Amharic Ethnologue entry (Paul et al., 2015a) (Eth).

(or *Shoa*) dialects (spoken in the central area of Ethiopia; the former being recognized as Standard Amharic), the *Gojjam* dialect (spoken in Däbrä Marqos), the *Mänz* dialect (spoken in Molale), the *Wällo* dialect (spoken in Sulula and Hayq), and the *Gondar* (or *Gonder*, or *Bägermeder*) dialect (spoken in the Amhara area (Parfitt and Semi, 2013)). The interested reader is pointed to the article entitled *Regional Variations in Amharic* (Habte, 1973), which provides an overview of the main features of the *Gojjam*, *Mänz* and *Wällo* dialects, while Parfitt and Semi (2013) details the regional variant called *Gondar*.

2.2 Main typological and syntactic features

Linguistically Amharic belongs to the Ethio-Semitic branch of the Semitic family (Hammarström et al., 2015). Since Amharic has been in contact with various Semitic, Cushitic, and Omotic languages, it resembles some of their features (e.g., contrast of plain and glottalized ejectives, verb idioms using the verb 'say', word order with the verb last, and word formation by the use of suffixes) (Teferra and Hudson, 2007).

2.2.1 Linguistic typology

Phonological level. Since Amharic employs 30 consonant and only 7 vowel phonemes, the consonant-to-vowel ratio is comparably high (Leslau, 1997; Dryer and Haspelmath, 2013). As regards voicing in stops and fricatives, Amharic maintains a strong voicing contrast in all positions (it is apparently a controversial question, while (Teferra and Hudson, 2007) argues for voicing in fricatives, the Amharic WALS entry (2013) only cites voicing in plosives, and the Wikipedia entry on Amharic (2015) notes that the voiced bilabial fricative is only present in borrowed English words). Amharic is not a tonal language (Dryer and Haspelmath, 2013). Amharic employs a (C)V(C)(C) syllable structure (where parentheses indicate optional elements) (Project, 2006a).

Morphological level. Amharic utilizes a root-and-pattern morphology, and it is a weakly suffixing language (Project, 2006a; Dryer and Haspelmath, 2013; Teferra and Hudson, 2007).

Morphosyntactic level. As summarized by the colleagues of the UCLA Language Materials Project (2006b), Amharic nouns are inflected for: gender (masculine/feminine), number (singular/plural), definiteness (definite/indefinite), case (nominative/accusative/genitive/vocative) and direct object status (following the order presented here).

Syntactic level. The typical word order in Amharic is SOV Project (2006b). Below the four basic sentence patterns are described.

2.2.2 Predication

The simple predicative structure follows SOV pattern Kebede (2002):

owtobəs-u	yihedal
bus-DET	3RD.SG.NEUTRAL-is-going
‘The bus is going.’	

2.2.3 Possession

Possessive pronouns are made using the possessive marker 'yeu' Kebede (2002):

yeu-ne	gwadeunya	nat
POSS-1ST.SG.	friend	3RD.SG.FEMALE-is
‘She is my friend.’		

2.2.4 Imperative

Simple imperative sentences follow the pattern below Kebede (2002):

sō-yō-n	t̥ora	
man-DET-ACC	2ND.SG.MALE-call	
‘Call the man!’		

2.2.5 Interrogative

A typical interrogative sentence may employ the word of existence 'alleu' (referred to as WOE), equivalent to the English "there exist" phrase Kebede (2002):

shama	alleu	
candle	WOE	
‘Is there candle?’		

Other forms of interrogation might involve question words, e.g. Kebede (2002):

man	nō	
who	3RD.SG.NEUTRAL-is	
‘Who it is?’		

2.3 Writing system, transcriptions

የኢትዮጵያ ዘመን በፊት አማካይ የግዢ ማኅበር የአማርኛ የግዢ ማኅበር የአማርኛ የግዢ ማኅበር :

Figure 2.2: Ge'ez (Ager, 2015).

Amhara people use Ethiopic (Ge'ez) script for writing (see Figure 2.2). Originally Ge'ez was an *abjad* alphabet (the vowels were omitted in writing). The current version of Ethiopic, however, is an *abugida* systems (where consonants and vowels in a syllable are treated as clusters). The original set of 182 characters is now over 500 symbols, marking such fine sub-phonemic differences as palatalization, pharyngealization and labialization (ScriptSource, 2015).

Ethiopic script is situated in Unicode ranges 1200–137F, 1380–139F, 2D80–2DDF and AB00–AB2F. There are several romanization guides available for Ge'ez, e.g., US Board on Geographic Names (BGN).

2.4 Previous research on the language

The history of Amharic grammars reach back to the late 17th century, but the first comprehensive writing on nuances of the Amharic phonology, morphology and syntax was published by [Wolf \(1995\)](#). Designed as an academic reference book, *Reference Grammar of Amharic* is also special in the sense that it offers illustrations to the linguistic phenomena drawn from everyday life, as well as insights in the specific features of literary language usage, and that of Wollo, Mänz, Gojjam and Gondar language variants.

Additional important sources are [Wolf \(1968\)](#); [Dawkins \(1969\)](#); [Appleyard \(1995\)](#) and [Teferra and Hudson \(2007\)](#). (Further digital resources are available on the website of the [Defense Language Institute Foreign Language Center](#).)

For an exhaustive literary list on the available reference grammars, the reader is pointed to the [Amharic entry of OLAC](#). For a typological overview, visit the [Amharic section of WALS](#).

2.5 Data and sources

2.5.1 Basic vocabulary

- The *An Crúbadán* project offers a [package of Amharic resources](#) (such as character trigrams, word bigrams and word frequency tables) compiled from 3726 documents, with a total of 5,944,494 words
- Based on the New Testament, a word frequency table is also available on the [academic site of Levente Madarász](#). (Login and password: DLD)

2.5.2 Dictionaries

The following online and offline dictionaries are available for Amharic: [Cain \(2015\)](#); [Turton et al. \(2008\)](#); [Akilu \(1973\)](#); [et al \(2010\)](#); [Gutt and Mohammed \(1995\)](#); [et al \(1997\)](#); [Neudorf and Neudorf \(2007, 2014, 2007, 2014\)](#); [Wolf \(1976\)](#).

2.5.3 Corpora

In this part, the reader is presented with a list on the available Amharic corpora.

Monolingual

By far, the best monolingual corpus of Amharic is provided by the Amharic Wikipedia. As of 2015-12-01, the [Amharic Wikipedia](#) consisted of 42,347 pages; these are available at the [Amharic section of Wikimedia Downloads](#). The interested reader may also consult with [Gambäck et al. \(2009\)](#); Gambäck and his colleagues constructed an Amharic corpus for machine learning consisting of 8715 Amharic news articles from between 2001 and 2004. For Amharic texts, see also the [Amharic edition of Voice of Africa](#).

Bilingual

Amharic [Microsoft](#) and [Ubuntu](#) localization files are good candidates for building parallel corpora.

The list below contains further candidates for bilingual corpus building:

- **The New Testament** in .XML format, made available on Christos Christodoulopoulos' academic profile
- **The Quran** in .XML format on OPUS, compiled by the Tanzil project
- **The Book of Mormon** in .PDF format, available on the web page of The Church of Jesus Christ of Latter-Day Saints
- **The Universal Declaration of Human Rights** in .TXT format on the web page of The Unicode Consortium

2.6 Computational tools

In this final section, the author reviews the available computation tools for Amharic.

2.6.1 Language Identification

For language identification, besides [Compact Language Detector 2 \(161\)](#), [TextCat](#) language guesser and the [saffsd langid.py LangID tool](#) are also compatible with Amharic.

[Polyglot3000](#) is also capable of identifying Amharic.

2.6.2 Tokenizer

MorphAdorner developed by Philip R. Burns supports Amharic ([2013](#)) and is [available online](#).

Rami Al-Rfou's [polygot](#) also support tokenization. (The above tools are also capable of language identification.)

2.6.3 Stemmer

Atelach Alemu Argaw and Lars Asker ([2007](#)) demonstrated an Amharic stemmer that is capable of 60% accuracy for old fashioned fiction and 75% accuracy for news text.

2.6.4 Spell checker

Open source spell checkers include [HunSpell](#) and [Mozilla dictionaries](#).

2.6.5 Phrase level and higher tools

In this section a collection of article references are provided, each dealing with higher-level computational tools for Amharic:

- **Amharic part-of-speech tagger:** [Adafre \(2005\)](#); [Gambäck \(2012\)](#); [Gebre \(2010\)](#); [Tachbelie et al. \(2011\)](#)
- **Amharic question answering system:** [Yimam and Libsie \(2010\)](#)
- **Amharic sentence parser:** [Ibrahim and Assabie \(2014\)](#)
- **Amharic speech recognizer:** [Abate \(2005\)](#)

2.6.6 End-user support

- **Mac OSX** (as of 2015-12-01): No OS-level support
- **Microsoft Windows** (as of 2015-12-01): Language pack available
- **Ubuntu** (as of 2015-12-01): GNOME translation updates available

Bibliography

- S. T. Abate. *Automatic speech recognition for Amharic*. PhD thesis, Universität Hamburg, 2005.
- S. F. Adafre. Part of speech tagging for Amharic using conditional random fields. In *Proceedings of the ACL workshop on computational approaches to semitic languages*, pages 47–54. Association for Computational Linguistics, 2005.
- S. Ager. Omniglot - writing systems and languages of the world, 2015. URL www.omniglot.com.
- A. Akilu. *English-Amharic Dictionary*. Oxford University Press, 1973.
- D. Appleyard. *Colloquial Amharic*. Routledge, 1995. ISBN 9780415100038.
- A. A. Argaw and L. Asker. An Amharic stemmer: Reducing words to their citation forms. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Semitic '07, pages 104–110, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=1654576.1654594>.
- P. R. Burns. Morphadotter v2.0, 2013. URL <https://devadotter.northwestern.edu/maserver/>.
- M. Cain. Amharicdictionary.com, 2015. URL <https://www.amharicdictionary.com/>.
- C. H. Dawkins. *The Fundamentals of Amharic*. Sudan Interior Mission, 1969.
- M. S. Dryer and M. Haspelmath, editors. *Language Amharic*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL http://wals.info/languoid/lect/wals_code_amh.
- Bryant et al. *Surichen ko Aranjacen ko Galachen (Suri - English - Amharic Latin dictionary)*. SIL Language and Culture Archives, 2010.
- Gutt et al. *Silt'e - Amharic - English dictionary*. Addis Ababa University Press, 1997.
- B. Gambäck. Tagging and verifying an Amharic news corpus. *Language Technology for Normalisation of Less-Resourced Languages*, page 79, 2012.
- B. Gambäck, F. Olsson, A. A. Argaw, and L. Asker. An Amharic corpus for machine learning. In *Proceedings of the 6th world congress of African linguistics*, 2009. URL https://www.sics.se/~gamback/publications/wocal09_abs.pdf.
- Federal Negarit Gazeta. The Constitution of the Federal Democratic Republic of Ethiopia. *Addis Ababa*, 1994.
- Binyam Gebrekidan Gebre. *Part of speech tagging for Amharic*. PhD thesis, University of Wolverhampton Wolverhampton, 2010.

- E. H. M. Gutt and H. Mohammed. *Yäsilt'iñä - Amariña - Ingliziña k'amus (inc'er kutab) : Yäsilt'iñä - Amariña - Ingliziña mäzgäba l'alat (ac'ar atam) = Silt'e - Ahmaric - English dictionary (shorter version)*. Summer Institute of Linguistics, 1995.
- M. M. Habte. Regional variations in Amharic. *Journal of Ethiopian Studies*, 11, 07 1973. doi: 10.2307/41988260.
- H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 2.6, 2015. URL <http://glottolog.org>.
- A. Ibrahim and Y. Assabie. Amharic sentence parsing using Base phrase chunking. In *Computational Linguistics and Intelligent Text Processing*, pages 297–306. Springer, 2014.
- T. Kebede. *Ethiopian Amharic Phrasebook*. Lonely Planet Phrasebook Guides. Lonely Planet, 2002. ISBN 9781740591331. URL <https://books.google.hu/books?id=fCQjQkaZBEwC>.
- W. Leslau. Amharic phonology. In A. S. Kaye, editor, *Phonologies of Asia and Africa (including the Caucasus)*, pages 399–430. Eisenbrauns, 1997.
- J. Minahan. *Miniature Empires: A Historical Dictionary of the Newly Independent States*. Taylor & Francis, 2013. ISBN 9781135940171. URL <https://books.google.hu/books?id=wSBeAgAAQBAJ>.
- A. Neudorf and S. Neudorf. *Bertha-English-Amharic dictionary*. Benishangul-Gumuz Language Development Project, 2007.
- A. Neudorf and S. Neudorf. *Bertha English Amharic Arabic Dictionary*. Benishangul-Gumuz Language Development Project, 2014.
- T. Parfitt and E.T. Semi. *The Beta Israel in Ethiopia and Israel: Studies on the Ethiopian Jews*. Taylor & Francis, 2013. ISBN 9781136816680. URL <https://books.google.hu/books?id=cg4fAgAAQBAJ>.
- L. M. Paul, G. F. Simons, and D. F. Charles. Amharic - ethnologue, 2015a. URL <http://www.ethnologue.com/language/amh>. [Online; accessed 29-November-2015].
- L. M. Paul, G. F. Simons, and D. F. Charles. Ethnologue: Languages of the World, 2015b. URL <http://www.ethnologue.com>.
- Joshua Project. Language - Amharic, 2014. URL <http://joshuaproject.net/languages/amh>. [Online; accessed 29-November-2015].
- UCLA Language Materials Project. Amharic, 2006a. URL <http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=7>.
- UCLA Language Materials Project. Somali, 2006b. URL <http://www.lmp.ucla.edu/Profile.aspx?LangID=202&menu=004>.
- ScriptSource. Scriptsource - writing systems, computers and people, 2015. URL <http://scriptsource.org>.
- M. Y. Tachbelie, S. T. Abate, and L. Besacier. Part-of-speech tagging for underresourced and morphologically rich languages—the case of Amharic. *HLTD (2011)*, pages 50–55, 2011.

- A. Teferra and G. Hudson. *Essentials of Amharic*. Afrikawissenschaftliche Lehrbücher. Rüdiger Köppe, 2007. ISBN 9783896455734. URL <https://books.google.hu/books?id=nwQaAQAAIAAJ>.
- D. Turton, M. Yigezu, and O. Olibui. Mursi-English-Amharic dictionary, 2008. URL <http://www.mursi.org/pdf/dictionary.pdf>.
- Wikipedia. Amharic — Wikipedia, the free encyclopedia, 2015. URL <https://en.wikipedia.org/w/index.php?title=Amharic&oldid=692206133>. [Online; accessed 30-November-2015].
- L. Wolf. *Amharic textbook*. Otto Harrassowitz Verlag, 1968. ISBN 9783447005548.
- L. Wolf. *Concise Amharic Dictionary*. University of California Press, 1976.
- L. Wolf. *Reference grammar of Amharic*. Harrassowitz, 1995. ISBN 9783447033725.
- S. M. Yimam and M. Libsie. Amharic question answering (aqa). In *10th Dutch-Belgian Information Retrieval Workshop*, page 98, 2010.

Chapter 3

Arabic (Nikolett Mus)

Contents

3.1 Demography and ethnography	29
3.2 Main typological and syntactic features	34
3.3 Writing system, transcription	38
3.4 Previous research on the language	38
3.5 Data and sources	39
3.6 Computational tools	41
Bibliography	44

Introduction

The chapter provides detailed information related to the following aspects of the Arabic language (Afro-Asiatic, Semitic): its demographic and ethnographic situation (see Section 3.1); its typological characterization (see Section 3.2); the writing system(s) of the language (see Section 3.3). Additionally, the previous research on the Arabic language will be surveyed in Section 3.4. Furthermore, Section 3.5 introduces the available primary data and sources on the Arabic language, such as basic vocabularies, dictionnaires, corpora, news portals, etc. Finally, Section 3.6 contains a collection of computational tools used for linguistic research of the Arabic language.

3.1 Demography and ethnography

3.1.1 Name variants

The **Arabic** language belongs to the Semitic branch of the Afro-Asiatic language family. Within the Semitic branch, Arabic is categorized as a Central Semitic language. Its closest relatives are Aramaic, Hebrew, Ugaritic and Phoenician.

The Arabic language has three main variants, which are the **Classical Arabic**, the **Modern Standard Arabic** (henceforth MSA) and the **colloquial Arabic**. Classical Arabic is the language of the Qur'an. MSA – which is based on Classical Arabic – is the only one official written form of the language, which is, nevertheless, not a widespread spoken variant of the Arabic language. It is only sporadically used by members of a particular social group. However, MSA is the language of, *inter alia*, education, mass media and official administration. Finally, the colloquial Arabic is the

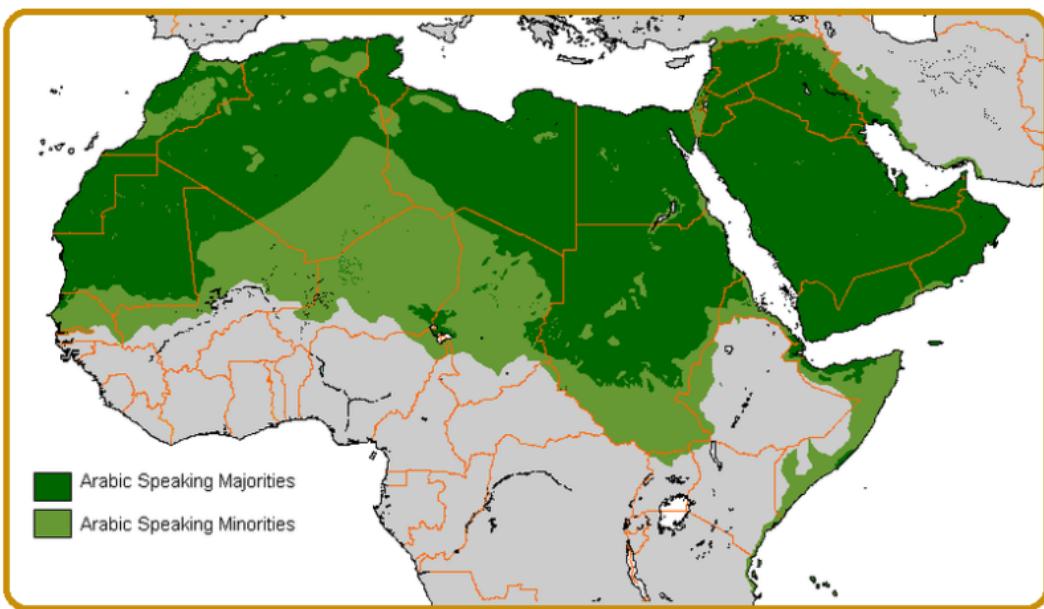


Figure 3.1: The Arabic speaking territories (source: [eRepublik](#))

spoken variant of Arabic. It has several spoken varieties or dialects itself. These so-called dialects (or varieties) are often mutually unintelligible and can rather be considered to be separate languages linguistically. However, on political and/or religious grounds they are treated as being one language.

The ISO 639-3 code of MSA is **arb**. In addition, there is an ISO 639-3 code for the spoken Arabic *macrolanguage*, which is **ara**. Furthermore, some individual spoken variants have separate ISO 639-3 code. These variants and their ISO 639-3 codes are summarized in Table 3.1 (cf. [SIL International](#)).

3.1.2 Geographic spread

The territory in which Arabic is spoken spreads over Western Asia and North Africa. Arabic is the sole official language in the following countries: Algeria, Bahrain, Egypt, Jordan, Kuwait, Lebanon, Libya, Mauritania, Oman, Qatar, Saudi Arabia, Syria, Tunisia, United Arab Emirates and Yemen.

In addition, Arabic is one of the official languages of Chad, Comoros, Djibouti, Eritrea, Iraq, Israel, Morocco, Palestine, Somalia and Sudan. The Map (3.1) illustrates the Arabic speaking territories.

3.1.3 Speaker populations

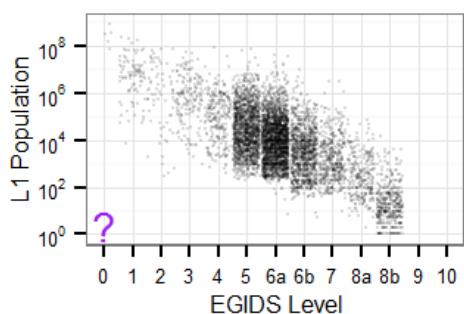
Arabic is the sixth most spoken language in the world. There are cc. **420 million** speakers who speak at least one variant of the Arabic macrolanguage.

The EGIDS (Expanded Graded Intergenerational Disruption Scale) level of MSA is **1** (national), i.e. the language is used in education, work, mass media, and government at the national level (see Figure 3.2).

This value may vary depending on the individual status of the local Arabic variants. For instance, the variety with the most speakers, i.e. the Egyptian Arabic is spoken by cc. 55 million people, exhibits **3** (wider communication) on the EGIDS scale. It means that Arabic in Egypt is used in work and mass media without official status to transcend language differences across a region (see Figure 3.3; for more details of the EGIDS levels of the Arabic varieties spoken in different countries see the corresponding entries of [Ethnologue](#)).

Table 3.1: The variants of the Arabic language

Variants of Arabic	ISO 639-3 code
Algerian Saharan Arabic	aao
Tajiki Arabic	abh
Baharna Arabic	abv
Mesopotamian Arabic	acm
Ta'izzi-Adeni Arabic	acq
Hijazi Arabic	acw
Omani Arabic	acx
Cypriot Arabic	acy
Dhofari Arabic	adf
Tunisian Arabic	aeb
Saidi Arabic	aec
Gulf Arabic	afb
South Levantine Arabic	ajp
North Levantine Arabic	apc
Sudanese Arabic	apd
Moderan Standard Arabic	arb
Algerian Arabic	arq
Najdi Arabic	ars
Moroccan Arabic	ary
Egyptian Arabic	arz
Uzbeki Arabic	auz
Eastern Egyptian Bedawi Arabic	avl
Hadrami Arabic	ayh
Libyan Arabic	ayl
Sanaani Arabic	ayn
North Mesopotamian Arabic	ayp
Babalia Creole Arabic	bbz
Sudanese Creole Arabic	pga
Chadian Arabic	shu
Shihhi Arabic	ssh

Figure 3.2: The EGIDS level for MSA (source:[Ethnologue](#))

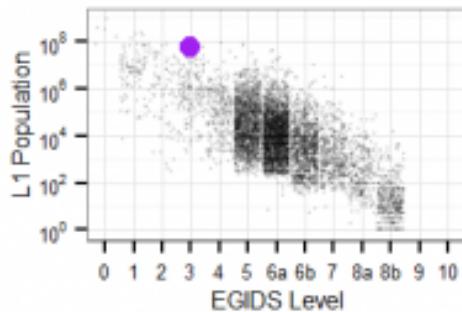


Figure 3.3: The EGIDS level for Egyptian Arabic (source:[Ethnologue](#))

Additionally, Arabic is one of the six official languages of the United Nations, as well as, the liturgical language of Muslims. The distribution, the number of the speakers in Arabic speaking countries the EGIDS level, etc. is illustrated in Table 3.2.

3.1.4 Dialect situation

Spoken Arabic is linguistically diverse. There are different variants of the spoken Arabic language, which can be identified with countries in which they are spoken. Additionally, the variants of spoken Arabic can be divided regionally. As seen, there are **Northern**, **Central**, **Western**, and **Southern** variants of the Arabic language.

The Northern variants:

- Levantine Arabic
 - North Levantine Arabic
 - * North Syrian Arabic
 - South Levantine Arabic
 - * Syrian Arabic
 - * Jordanian Arabic
 - * Palestinian Arabic
 - * Lebanese Arabic
 - +¹ Cypriot Maronite
 - Bedawi Arabic
- Mesopotamian Arabic
- qeltu variants
 - Baghdad Arabic
 - Khuzestani Arabic
- qeltu variants

– North Mesopotamian Arabic

The Central variants:

- Egyptian Arabic
- Sa’idi Arabic
- Sudanese Arabic

The Western variants:

- Maghrebi Arabic
 - Koines
 - * Moroccan Arabic
 - * Algerian Arabic
 - * Tunisian Arabic
 - * Libyan Arabic
 - Fully pre-Hilalian
 - * Jebli Arabic
 - * Jijel Arabic
 - * +Siculo-Arabic
 - 1. Maltese language
 - Bedouin

¹ The + symbol marks the extinct variants of Arabic

Table 3.2: The Arabic speakers

Country	Number of L1 speakers	EGIDS level	Ratio of L1 speakers	Number of L2 speakers
Egypt	52,500,000	3	63,64%	no data
Algeria	26,000,000	3	66,3%	3,000,000
Saudi Arabia	25,947,000	no data	90%	no data
Iraq	24,440,000-26,100,000	no data	75-80%	no data
Syria	19,000,000	no data	85%	no data
Morocco	18,800,000	3	57%	5,000,000
Sudan	17,000,000	3	44,8%	14,000,000
Yemen	15,100,000	6 ^a	61,85,%	2,400,000
Tunisia	10,880,000	no data	99%	no data
Lebanon	4,132,000	no data	70	no data
Libya	4,000,000	3	62,5%	no data
Mauritania	2,770,000	no data	68,1%	no data
Israel	1,688,600	no data	20,7%	no data
Chad	1,681,000	3	12,3%	no data
Palestine	1,600,000	3	33,7%	no data
Kuwait	1,403,962	no data	33%	no data
Oman	other Arab: 707,490	21%	19,83%	no data
Jordan	720,000	6 ^a	10,87%	no data
United Arab Emirates	710,000	no data	7,46%	no data
Bahrain	697,000	no data	46%	no data
Qatar	568,399	no data	5,4%	no data
Eritrea	other Arab: 66,903	36,27%	2%	no data
Djibouti	665,000	no data	4,8%	no data
Somalia	114,000	no data	0,1%	no data
Comoros	38,900	no data	no data	no data
Total	11,500	no data	no data	no data

- * Saharan Arabic
- * Hassaniya Arabic
- Andalusian Arabic
- The Southern variants (also called as Peninsular Arabic):
 - Gulf Arabic
 - Bahrani Arabic
 - Najdi Arabic
 - Hijazi/Hejazi Arabic
 - Yemeni Arabic
 - Hadhrami Arabic
 - Sanaani Arabic
 - Ta'izzi-Adeni Arabic
 - Dhofari Arabic
 - Omani Arabic
 - Shihhi Arabic

Furthermore, **Peripheral, Jewish** (e.g. Judeo-Arabic), **Creoles** (e.g. Nubi Creole Arabic), and **Pidgin** (e.g. Turku Arabic) variants are often considered (for more details see the corresponding [wikipedia entry](#)).

The Peripheral variants:

- Central Asian Arabic

The Table 3.3 shows the number of speakers of the dialects.

Map 3.4 illustrates the geographical spread of these main Arabic variants.

Furthermore, there is a **diglossic language situation**, whereas the written, i.e. MSA, and the spoken versions of Arabic are used simultaneously by the same speaker community for different purposes.

- Tajiki Arabic
- Uzbeki Arabic
- Shirvani Arabic
- Chadian Arabic (Baggara, Shuwa Arabic)
- Nigerian Arabic
- Khuzestani Arabic
- The Jewish variants:
 - Judeo-Arabic
 - Judeo-Iraqi Arabic
 - * Judeo-Baghdadi Arabic
 - Judeo-Moroccan Arabic
 - Judeo-Tripolitanian Arabic
 - Judeo-Tunisian Arabic
 - Judeo-Yemeni Arabic

The Creole variants:

- Nubi Creole Arabic
- Babalia Creole Arabic
- Sudanese Creole Arabic (Juba Arabic)

The Pidgin variants:

- +Maridi Arabic
- Turku Arabic

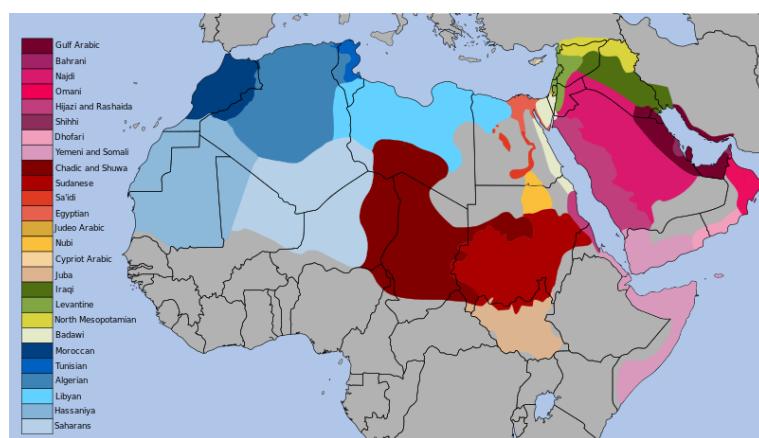
3.2 Main typological and syntactic features

3.2.1 Linguistic typology

As previously mentioned, Arabic is a Semitic language. The Semitic languages share some typical characteristics that applies to the phonology and morphology of the languages in the group. In the phonological inventory of the Semitic languages there are only six vowels, three short /a, i, u/ and their long counterparts (this number may vary considerably in the dialects). Additionally, the Semitic

Table 3.3: The number of the speakers of the Arabic dialects

Arabic varieties	Number of speakers
Levantine Arabic	21,000,000
Mesopotamian Arabic	15,000,000
Egyptian Arabic	55,000,000
Sa'idi Arabic	19,000,000
Sudanese Arabic	(?) 17,000,000
Moroccan Arabic	21,000,000
Algerian Arabic	27,000,000
Tunisian Arabic	11,200,000
Libyan Arabic	4,325,000
Maltese language	520,000
Saharan Arabic	130,000
Hassaniya Arabic	3,200,000
Gulf Arabic	5,000,000
Bahrani Arabic	300,000
Najdi Arabic	(?) 10,000,000
Hijazi/Hejazi Arabic	6,000,000
Yemeni Arabic	15,100,000
Dhofari Arabic	70,000
Omani Arabic	1,055,000
Shihhi Arabic	44,000
Chadian Arabic (Baggara, Shuwa Arabic)/Nigerian Arabic	1,100,000
Judeo-Iraqi Arabic	152,000 - 172,000
Judeo-Moroccan Arabic	260,000
Judeo-Tripolitanian Arabic	35,000
Judeo-Tunisian Arabic	46,000
Judeo-Yemeni Arabic	50,000
Nubi Creole Arabic	42,000
Babalia Creole Arabic	3,900
Sudanese Creole Arabic (Juba Arabic)	20,000

Figure 3.4: The variants of the Arabic language (source: [Wikipedia](#))

Arabic IPA Chart										
Bilabial	Labiodental	Dental	Alveolar	Palatoalveolar	Retroflex	Palatal	Vocal	Uvular	Pharyngeal	Glottal
Stop	b		t t̪ d d̪			k g q				?
Nasal	m		n							
Trill			r							
Tap or flap			r̪							
Fricative	f θ ð ð̪ s s̪ z		ʃ		ʒ	x	χ ħ	ʃ ħ		
Lateral fricative				dʒ		j				
Approximate	w									
Lateral approximants			l t̪							

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Created by John Oberlin, Kent State University. jpober@kent.edu

Figure 3.5: The Arabic consonant inventory represented by IPA characters (source: [algebra.xls.pt](#))

languages have relatively rare consonants, such as the pharyngeal fricatives. Besides, the nouns in the Semitic languages are typically inflected for two genders (masculine and feminine) and three numbers (singular, plural, and dual). The difference between genders can be observed in second and third person too (cf. [the blog entry of the Linguistlist](#)).

The various spoken forms of Arabic differ from each other in terms of their phonological, morphological and syntactic features. This section describes some features, which typically characterise MSA. However, a comprehensive comparative description is not intended here. For detailed descriptions of the phonology and morphology of Arabic see e.g. [Al-Ani \(1970\)](#) and [Watson \(2002\)](#).

Phonological level In MSA there are 6 vowels , i.e. /a, a:, i, i:, u, u:/, and 28 consonants.

As it is seen the phonemic quality of length applies to consonants as well as vowels.

Morphological level In Arabic, grammatical relations are mostly expressed by suffixes. There are, nevertheless, prepositions also available in the language.

Morphosyntactic level The nouns in Arabic are typically specified for three numbers, i.e. singular, dual and plural; three cases, i.e. nominative, accusative, and genitive; two genders, i.e. masculine and feminine; and three states, i.e. indefinite, definite, and construct. The spoken variants of Arabic may show differences. For example, dual number is not used in some spoken Arabic, e.g. in Maghrebi Arabic. In these variants the plural forms of nouns are used instead. The plural is usually indicated through suffixes. There are some instances, however, in which internal modification expresses the plural forms of nouns. In addition, the grammatical cases are usually expressed by suffixes attached to the nouns. However, in some variants, such as in Egyptian Arabic, there is no morphological case marking. Finally, the genders of Arabic nouns are sex-based ones. The definiteness of nouns are expressed by definite affixes. Besides, a definite article can also be found in Arabic.

Verbs functioning as predicates are marked for two tenses, i.e. non-past and past; four moods, i.e. indicative, subjunctive, jussive, imperative; and two voices, i.e. active and passive. Additionally, predicate verbs take agreement markers. The future tense is expressed by a prefix. The modal markers are available in the non-past only. There are two further moods in Classical Arabic: short and long energetics.

Syntactic level Two basic word order patterns can be observed in the Arabic variants including MSA. One of these patterns exhibit the VSO order in the simple declarative clauses. This order is attested in MSA and in the Syrian Arabic language. However, in the Syrian variant of Arabic, the clauses may also exhibit the SVO order, and there is no dominant between VS and SV orders. The

other attested word order is the SVO, which is available for the vast majority of the Arabic variants, e.g. Egyptian, Gulf, Iraqi, Kuwait Arabic, etc.

On the phrase level both head final and head initial strategies can be observed. The genitives (Gen), the adjectives (Adj) and the relative clauses (Rel) follow their noun heads, i.e. their order show head-initial patterns. Consequently, N-Gen, N-Adj, and N-Rel orders surface. In contrast, the numerals (Num) typically precedes the quantified nouns, i.e. they appear in Num-N order. In some variants of Arabic (e.g. in Kuwaiti and in Gulf Arabic), there is no dominant order of numeral and noun, as well as, of demonstrative (Dem) and noun. In these phrases, both the Num-N, N-Dem and the N-Num, Dem-N orders are attested.

3.2.2 Predication

The agreement between the predicate verb and its subject takes place in gender, i.e. masculine and feminine; in person, i.e. first, second, third; and in number, i.e. singular, dual, plural.

ولد الولد في مصر.
wulida l-walad fi miSr
'The boy was born in Egypt.'

Nominal and adjectival predicates in the present tense are formed without a copular verb. In the past and the future, a copula is used which gets the temporal markers. The locational predicate is encoded by the strategy used in nominal predication.

الولد مصري.
al-walad miSri
'The boy is Egyptian.'

3.2.3 Possession

There is a set of possessive pronouns in Arabic, which are used as suffixes, i.e. they are attached to the possessed noun.

يَقْرَأُ الْوَلَدُ كِتَابًا.
yaqra'(u) ('a)l-walad(u) kitaaba-h(u)
'The boy reads his book.'

3.2.4 Imperative

The imperative is restricted to the 2nd person singular and plural. The imperative is implicated by morphemes attached to the verb.

إِوْعِي رَاسِكَ
ew3a raasak
'Watch out for your head.'

In Palestinian Arabic, the morphological markers of imperative are missing. In prohibition, the combination of prohibitive, i.e. negative imperative, and normal sentential negative structures are used. In some variants, such as in Gulf Arabic, normal imperative and a special negative structure is available in prohibition.

ر	د	ذ	خ	ح	ج	ث	ت	ب	ا
راء	DAL	DAL	خاء	حاء	جم	ثاء	تاء	باء	ألف
rā'	dāl	dāl	bā'	ḥā'	ğim	tā'	tā'	bā'	'alif
r	d	d	b	ḥ	ğ	t	t	b	(a)
[r~r ⁵]	[ð]	[d]	[b]	[χ~χ]	[ħ~ħ]	[ç~ç]	[θ]	[t]	[b]
www.omniglot.com									
ف	غ	ع	ظ	ط	ض	ص	ش	س	ز
فاء	غين	عين	ظاء	طاء	ضاد	صاد	شين	سين	زاي
fa'	ğayn	'ayn	zā'	tā'	dād	şād	śīn	śīn	zāy
f	ğ	'	z	t	d	ş	ś	s	z
[f]	[χ~χ]	[?~?]	[ð ⁵]	[t ⁵]	[d ⁵ ~z ⁵]	[ç ⁵]	[ʃ ⁵]	[s ⁵]	[z ⁵]
ء	ي	و	ه	ن	م	ل	ك	ق	كاف
هبرة	باء	واو	هاء	نون	مميم	لام	كاف	قاف	قاف
hamza	yā'	wāw	hā'	nūn	mīm	lām	kāf	qāf	qaf
	[j]	[w]	[h]	[n]	[m]	[l~ɫ]	[k]	[q]	

Figure 3.6: The Arabic alphabet (source: [Omniglot](#))

3.2.5 Interrogative

The interrogative phrases in wh- (or content) interrogatives do not appear in initial position.

ما اسمك؟
ma ismuka?
'What's your name?'

arabic

3.3 Writing system, transcription

The basic Arabic Alphabet contains 28 letters written cursively. The Arabic language is written by the Arabic script called **Arabic abjad**. The alphabet only represents the consonants. The Arabic script is illustrated by Figure 3.6.

The direction of writing is from right to left, with the exception of numerals, which are written from left to right. The letters change their forms depending on their position in the word, i.e. they have different forms whether they appear at the beginning, middle or end of a word, or on their own (cf. [Omniglot](#)).

In the Qur'an, in other religious texts, in classical poetry, in books for children and foreign learners there are also vowel diacritics that are used to mark the short vowels.

Attempts to transcribe the Arabic script into Latin and/or Romanize the orthography have already been made. These attempts resulted in a number of various ways of transcription. But there is no standard transcriptional system of Arabic available.

3.4 Previous research on the language

Paper edited linguistic descriptions of the Arabic language are available (for a detailed collection see [WALS](#)). Comprehensive referential grammars of MSA such as [Cowan \(1958\)](#), [Haywood and Nahmadi \(1965\)](#), [Nasr \(1967\)](#), [Ryding \(2005\)](#) are accessible to the public.

In addition, several grammars of the spoken variants of Arabic can be found. Edited grammars of Egyptian Arabic are: [Abdel-Massih et al. \(1979\)](#), [Mitchell \(1956\)](#), [Mitchell \(1962\)](#), [Wise \(1975\)](#).

Furthermore, a study aiming at comparing the grammar of the several Arabic variants is also published (see ([Brustad, 2000](#))).

Besides, the following sources provide grammatical overviews and online courses for L2 speakers of Arabic: [LearnArabicOnline](#), [Arabic Course](#), [Learn101](#), [Arabic Keyboard](#), [Arabic Studio](#).

Research and control bodies The [Arabic Linguistics Society](#) aims at encouraging research in the field of modern Arabic linguistics and providing a forum for scholars interested in the study of Arabic. Additionally, the [Arabic Linguistics Forum](#) promotes academic and scholarly exchange on the linguistics of the Arabic language family. The forum organizes annual meetings/conferences on all areas of Arabic linguistics. Furthermore, departments of Arabic studies can be found *inter alia* at the [Georgetown University](#) (i.e. Department of Arabic and Islamic Studies), at the [University of Warwick](#) (i.e. Centre of Applied Linguistics).

3.5 Data and sources

This section provides lists of the available primer sources of Arabic.

3.5.1 Basic vocabulary

Online Arabic vocabularies and phrase lists are available, *inter alia*, at the following sites: [Atlas Tours](#), [101 Languages](#), [Arabic learning resources](#), [learn101](#).

Furthermore, the *An Crúbadán* project provides data, e.g. character trigrams, word bigrams, and word lists, etc. of certain Arabic variants: [Arabic](#) (on the basis of 11,948,481 words), [Standard Arabic](#) (1,049,780 words), [Egyptian Arabic](#) (on the basis of 1,382,233 words) (and its Latin version used in chats), [Moroccan Arabic](#) (637 words), [North Mesopotamian Arabic](#) (581,766 words), and [Tunisian Arabic](#) (1,215 words).

3.5.2 Dictionaries

Paper edition

There are Arabic–English dictionaries edited, e.g.:

- A [Modern Arabic–English Dictionary](#) which contains c. 67,000 entries, but does not contain any phonemic transcription or POS-tags
- [Arabic-English bilingual visual dictionary](#) which contains more than 6,000 words and phrases, with Latin transcriptions and English translations but without POS-tags; ([PDF](#))

Online dictionaries

The [English–Arabic Dictionary collection by Lexilogos](#) contains the following dictionaries:

- the [Reverso Dictionary](#) contains language pairs including English–Arabic; provides the words and translations and their POS-tags
- the [Reverso Context](#) provides translations in contexts

- the [Almaany Dictionary](#) provides translations and POS-tags of the words
- the [Glosbe Dictionary](#) contains the translation(s), the POS-tag(s), the transliteration, the pronunciation (that can be listened too), the declensional paradigms of the word entries
- the [The English to Arabic Dictionary](#) provides translations, Latin transcriptions and POS-tags of the words
- the [Egyptian Arabic Dictionary by Lisaan Masry](#) provides word entries, Latin transcriptions (the pronunciation can be listened), English translations, POS-tags, examples in which the word is used in clausal contexts and declensional paradigms of the Egyptian Arabic
- the [BabLa Dictionary](#) contains word translations, Latin transcriptions with pronunciation (which can be listened), contexts and POS-tags. It is also possible to search for entries by letters in this dictionary. This dictionary is easily scrapable

Further online dictionaries can also be found, e.g.:

- the [Arabic Dictionary by Lexicool](#)
- the [Arabic–English Dictionary by Systran](#)
- the [Ectaco](#);
- the [WebTranslation](#)

3.5.3 Corpora

Monolingual corpora

As of 11/01/2017, the [Arabic Wikipedia](#) contains 457,183 articles. There is also an Egyptian Arabic (arz) Wikipedia with 16,181 articles.

The [Qur'anic Arabic Corpus](#) provides morphological annotation, a syntactic treebank and a semantic ontology of the Qur'an. The corpus consists of 77,430 words of Qur'anic Arabic.

Besides, the [International Corpus of Arabic](#) (henceforth ICA) is planned to contain 100 million words, of which 79% is accomplished. The corpus contains morphologically annotated texts from numerous sources, e.g. newspapers, web articles, etc. written in different genres, e.g. literature, politics, etc. ICA represents more spoken variants of the Arabic language, as well as, the MSA language.

The [AQMAR](#) Arabic Wikipedia Named Entity Corpus & Tagger containing 74,000 tokens of 28 Arabic Wikipedia articles has been hand-annotated for named entities.

Additionally, the [300 Languages](#) Project, which is the sub-part of the *Rosetta Project*, includes Arabic data. The project aims at collecting materials of the 300 most widely-spoken languages in the world in order to build parallel corpora.

Furthermore, the following texts and Arabic translations can be used for building parallel corpora:

- the [Arabic](#) translation of the *Bible* is available online
- the *Qur'an* is in [Classical Arabic](#)
- there is a [Dictionary of the Holy Qur'an](#)
- the [Arabic](#) translation of the *Universal Declaration of Human Rights*

Bilingual corpora

English–Arabic parallel corpora are also available:

- the [QCRI Educational Domain Corpus](#), which is an open multilingual collection of subtitles for educational videos and lectures. The current release of the corpus (v1.4) contains 20 languages (distributed over 44,620 files).
- the parallel corpus of the [OPUS](#) project containing a (growing) collection of translated and linguistically annotated texts from the web. The data are collected from freely available online sources and the output of the project is also freely available.
- the [WIT3](#) (Web Inventory of Transcribed and Translated Talks) constituted by the TED Talks. The English-Arabic corpus contains 3.9 millions of (untokenized) words.

3.5.4 News portals

The following Arabic-language radio stations, TV stations and news portals can be found:

- Radio stations:
 - the [2ME Radio Arabic](#) that is based in Parramatta, broadcasting to Sydney, Melbourne, Hobart, Darwin, Brisbane, Adelaide, and Perth
 - the [2moro Radio](#) in Sydney, Australia
 - the [Al-Madina FM](#) Syrian radio station
 - the BBC Arabic radio station run by the BBC World Service
 - * Hadeeth as-Saa'a
 - * Tahqeeq
 - * [BBC Xtra](#)
 - * the Talking Point (Nuqtat Hewar)
 - the [Arta FM](#) Syrian community radio station established in Amudah. It produces radio programs in Kurdish, Syriac, Arabic and Armenian
- TV stations:
 - the [Al Jazeera](#)
 - the [BBC Arabic Television](#) that is a television news channel broadcast to the Middle East
 - the Talking Point (Nuqtat Hewar)
- Online news portals:
 - the [BBC](#) Arabic website, that serves as a Literary Arabic language news portal
 - the [Al Jazeera](#)
 - there is a [collection of Arabic websites](#) powered by drupal

3.6 Computational tools

This section introduces the main computational tools developed for Arabic.

3.6.1 Language identification

The [CLD2](#) provides full support for Arabic. Furthermore, another [identifier tool written in python](#) of Arabic is also available. Additionally, the [Basis technology Rosette Language Identifier](#) of Arabic can be found.

3.6.2 Tokenizer

There is an [Arabic tokenizer](#) developed at the Stanford University. Besides, another [Arabic tokenizer](#) is available. Additionally, tokenization and Parts-of-Speech (henceforth POS) tagging method is summarised by ([Habash and Rambow \(2005\)](#)).

3.6.3 Stemmer

An [Arabic stemmer](#) is developed by Shereen Khoja. In addition, the algorithm of Shereen Khoja forms the basis of another [Arabic stemmer](#).

3.6.4 Spell checker

Online spell checkers are provided by e.g. [SpellCheck.net](#) or [spellchecker.net](#). Additionally, there is a further [Online Spell Checker](#) for Arabic.

3.6.5 Phrase level and higher tools

An [Arabic toolkit by Microsoft](#) including Colloquial to Arabic Converter, Diacritizer, Named Entity Recognizer (NER), Parser, POS Tagger, SARF (morphological analyzer), Speller, and Transliterator is developed by the Microsoft Research Group. These tools are integrated into multiple Microsoft products.

Arabic morphological analyzer A [morphological analyser](#) for Arabic developed by Ken Beesley and Tim Buckwalter is available online.

Arabic part-of-speech tagger The Stanford Natural Language Processing Group provides a [POS tagger](#) for Arabic. Besides, Shereen Khoja developed a [POS tagger](#) as well.

Arabic chunker A new approach for chunking Arabic texts based on a combinatorial classification process is discussed by [Fraj and Kessentini \(2012\)](#) ([PDF](#)). Additionally, the challenges of processing the second generation of tools for Arabic (AMIRA) is addressed by [Diab \(2009\)](#) [PDF](#).

Arabic named entity recognizer An [Arabic NER tool called NERAr](#) is available online. In addition, ([Shaalan, 2014](#)) presents a survey of Arabic Named Entity Recognition and Classification ([PDF](#)). [Meselhi et al. \(2014\)](#) provides a hybrid approach to Arabic NER [PDF](#).

Arabic sentence parser Outside of the Microsoft parser, the Stanford Natural Language Processing Group provides further NLP tools for Arabic. In particular, a [parser](#), and a [word segmenter](#) are available on the site of the reserach group.

Arabic speech recognizer The [Google Cloud Speech API](#) supports many varieties and dialects of the Arabic language.

Arabic machine translator The [SDL](#) developed a rule-based machine translator for the Arabic language. In addition, the [Google translator](#) supports Arabic (for a detailed description and the supported languages see the [Wikipedia entry of Google Translate](#)).

Arabic question answering machine [Rosso et al. \(2006\) \(PDF\)](#), [Abouenour et al. \(PDF\)](#), [Bdour and Gharaibeh \(2013\) \(PDF\)](#), a.o., discuss and present modules of different Arabic Question Answering systems.

3.6.6 End-user support

The OS support for Arabic is the following:

- a [MAC OS x](#) language pack is available for Arabic
- there is a [Microsoft Windows](#) language pack
- an [Ubuntu](#) language pack is also available

The Unicode range for Arabic is 0600 — 06FF. Additionally, Arabic Presentation Forms-A (FB50 — FDFF) and Arabic Presentation Forms-B (FE70 — FEFF) can also be found.

The following sources provide online Arabic keyboards: [Arabic keyboard](#), [Lexilogos](#), [Clavier Arabic keyboard](#), [Yamli](#).

There are online resources, that automatically transcribe Arabic into the Latin script. Some of them are listed here: [Transliterating Arabic to English in One Step](#), [eiktub-TM](#), [Glosbe](#), [Arabic Transliterator](#).

Freely available OCR tools can be found provided by [NewOCR.com](#).

XeLaTeX allows to use any Arabic system fonts, i.e. there are packages that works with XeLaTeX to get Arabic texts into a document

Bibliography

Ernest T. Abdel-Massih, Z. N. Abdel-Malek, E.-S. M. Badawi, and E. N. McCarus. *A Comprehensive Study of Egyptian Arabic*. Center for Near Eastern and North African Studies, the University of Michigan, Ann Arbor, 1979.

Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. Idraaq: New arabic question answering system based on query expansion and passage retrieval. In *CLEF 2012 Conference and Labs of the Evaluation Forum*, page 2012.

Salman H Al-Ani. *Arabic phonology: An acoustical and physiological investigation*, volume 61. Walter de Gruyter, 1970.

Wafa N Bdour and Natheer K Gharaibeh. Development of yes/no arabic question answering system. *arXiv preprint arXiv:1302.5675*, 2013.

Kristen E. Brustad. *Syntax of spoken Arabic. A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. University Press, Georgetown, 2000.

David Cowan. *An Introduction to Modern Literary Arabic*. Cambridge University Press, Cambridge, 1958.

Mona Diab. Second generation amira tools for arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, 2009.

Fériel Ben Fraj and Maroua Kessentini. Combinatorial classification for chunking arabic texts. *International Journal of Artificial Intelligence & Applications*, 3(5):63, 2012.

Nizar Habash and Owen Rambow. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580, 2005.

J. A. Haywood and H. M. Nahmad. *A New Arabic Grammar of the Written Language*. Lund Humphries, London, 1965.

Mohamed A Meselhi, Hitham M Abo Bakr, Ibrahim Ziedan, and Khaled Shaalan. A novel hybrid approach to arabic named entity recognition. In *China Workshop on Machine Translation*, pages 93–103. Springer, 2014.

T. F. Mitchell. *An Introduction to Egyptian Colloquial Arabic*. Oxford University Press, Oxford, 1956.

T. F. Mitchell. *Colloquial Arabic: the Living Language of Egypt*. The English Universities Press, London, 1962.

- Raja T. Nasr. *The structure of Arabic: from sound to sentence*. Librairie du Liban, Beirut, 1967.
- Paolo Rosso, Yassine Benajiba, and Abdelouahid Lyhyaoui. Towards an arabic question answering system. In *Proc. 4th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROIV, Damascus, Syria*, pages 11–14, 2006.
- Karin C. Ryding. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press, Cambridge, 2005.
- Khaled Shaalan. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510, 2014.
- Janet CE Watson. *The phonology and morphology of Arabic*. Oxford University Press on Demand, 2002.
- Hilary Wise. *A transformational Grammar of Spoken Egyptian Arabic*. Blackwell, Oxford, 1975.

Chapter 4

Bengali (Levente Madarász)

Contents

4.1 Demography and ethnography	47
4.2 Main typological and syntactic features	49
4.3 Writing system, transcription	52
4.4 Previous research on the language	52
4.5 Data and sources	53
4.6 Computational tools	55
Bibliography	60

Introduction

The present paper aims at providing an overview of Bengali, a member of the Indo-Aryan languages, mainly spoken in Bangladesh, India and Nepal. The first part will present the reader with the demo- and ethnographic properties of Bengali speaker populations. Following the typological overview of Bengali, the available online and offline Bengali materials will be introduced. The closing section will detail the various digital resources and the available computational tools for the language.

4.1 Demography and ethnography

4.1.1 Name variants

The endonym of the language, Bangla (/baŋpla:/, বাংলা), probably originates from *Banga*, the name of a founding chief ([Online Etymology Dictionary, 2015](#)). Besides *Bengali*, the language of Bengali people is also known as *Begali*, *Banga-Bhasha*, *Bangala*, *Bangla* and *Bengali-Assamese* ([Hammarström et al., 2016](#)), identified with the *bn* ISO 639-1 and *ben* ISO 639-2 and ISO 639-3 codes ([Paul et al., 2015](#)).

4.1.2 Geographic spread

Figure 4.1 displays Bengali speaking areas around the world. Bengali is mainly spoken in Bangladesh where it is a statutory national, EGIDS level 1 language (see Figure 4.2.). Significant speaker populations also reside in India (EGIDS level 2, statutory provincial language used in West Bengal, Tripura

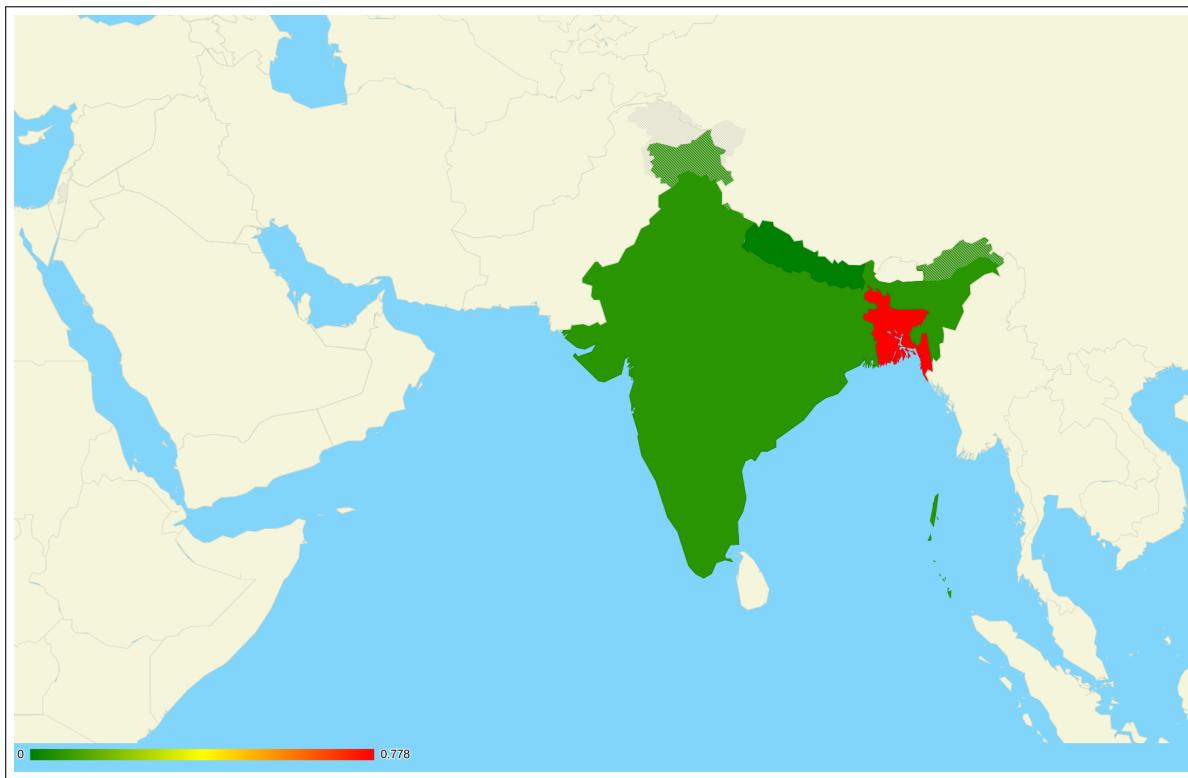


Figure 4.1: Map of Bengali speaking areas (Paul et al., 2016). Coloring indicates relative salience of the language in a given country; near 0% is indicated by green, while near 100% spread by red, in gray countries no Bengali speaker communities are attested.

and Assam states), while smaller communities populate Nepal (EGIDS level 5, dispersed language) and Singapore (EGIDS level 6b, threatened language).

4.1.3 Speaker populations

According to Encyclopædia Britannica (2015), Bengali people are of diverse origin: the region of present day Bangladesh was populated by Vedda people from Sri Lanka, who were joined by Mediterranean peoples speaking Indo-European languages. During the eighth century, Arab, Turkish and Persian people began to infiltrate the area; Bengali people descend from this diverse community.

As regards the contemporary situation of speaker populations, in addition to the 106,000,000 L1 and 19,200,000 L2 Bangladeshi Bengali speakers, Bengali is spoken by approximately 82,000,000 people outside Bangladesh, most of whom resides in India (for a comprehensive overview of the different communities in India, see Banthia et al. 2001, "Distribution of 10,000 persons by langauge - India, states and union territories"). Smaller communities are also present in Nepal (21,100 speakers) and Singapore (600 speakers) Paul et al. (2016). Besides the official records, The LINGUIST List (2014) also reports speaker populations with unspecified size in the US, Malawi, Saudi Arabia, the United Kingdom, as well as in the United Arab Emirates.

4.1.4 Dialect situation

Bengali has several regional variants (see Figure 4.3). Faquire (2012) distinguishes between speech varieties produced by native Bengali speakers and speech varieties spoken by people of Mongoloid origin (e.g., *Chakma*, *Hajong*, *Tanchangya*, *Rajbangshi* and *Mal Paharia*). According to the author, the core

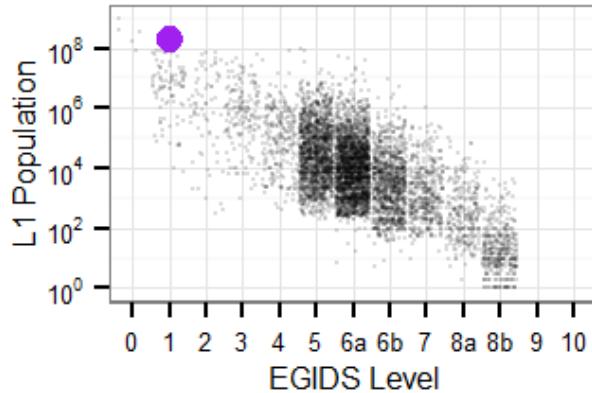


Figure 4.2: The EGIDS level for Bengali spoken in Bangladesh (source: [Ethnologue](#)). For more details about the plot, please see Section X.1 Demography and ethnography.

of the different variants is the Nadia district: the more distant is a particular region from Nadia, the more likely it is to observe variances on the levels of linguistic description (as a case in point, there are at least fifteen dialectal variants for the word 'son'). Grierson (cited in Faquire 2012) identified the following speech varieties of Bangla: *Central (Standard) Bengali*, *Western Bengali* (encompassing *Kharia Thar*, *Mal Paharia* and *Saraki*), *Southwestern Bengali*, *Northern Bengali* (including *Koch* and *Siripuria*), *Rajbanshi*, *Bahe*, *Eastern Bengali* (*East Central Bengali* with *Sylheti*), *Haijong* (or *Hajong*), *Southeastern Bengali* (*Chakma*), *Ganda*, *Vanga* and *Chittagonian* (for further details, please see Grierson 1903).

4.2 Main typological and syntactic features

Genetically, Bengali belongs to the Indo-Aryan Eastern zone branch of the Indo-European family (Hammarström et al., 2016). Below a linguistic sketch will be provided.

4.2.1 Linguistic typology

Phonological level. Bengali displays a relatively rich phonemic inventory consisting of 29 consonants (including a number of retroflex consonant articulations) complete with a vowel inventory of 7, the various combinations of which yield as many as 25 diphthongs (Chatterji, 1926, p. 415–416). For this reason the consonant to vowel ratio of the language is average (UCLA Language Materials Project, 2006). Bengali maintains contrast between voiced and unvoiced plosives, however there is no distinction in the case of fricatives. Bengali utilizes a fixed word-initial stress system (Dryer and Haspelmath, 2013; UCLA Language Materials Project, 2006), and according to Mahanta (2008), standard colloquial Bengali displays a non-iterative vowel harmony. Bengali is not a tone language and vowel length is not contrastive feature in its inventory. The syllable structure of the language adhered to by most speakers is (C)V(C).

Morphological level. Bengali is a highly inflectional language with 160 different verbal forms, 26 nominal forms and 26 pronominal forms. Bengali also utilizes reduplication for the marking of distributive numerals. Both prefixation and suffixation are attested (Bhattacharya et al., 2005; Dryer and Haspelmath, 2013; UCLA Language Materials Project, 2006).



Figure 4.3: Languages of Bangladesh (Paul et al., 2015).

Morphosyntactic level. Nouns are inflected for case (nominative, accusative, genitive, locative), as well as number/measure (the latter being a special class of suffixes attached to words of measure, e.g., nine, indicating the semantic class of measured objects). Finite Bengali verb forms inflect for person (first, second, third), tense (past, present, future), aspect (simple, progressive, perfect), mood (indicative, conditional, imperative), and honor (intimate, familiar, formal), however they do not inflect for number. In Bengali, non-finite verbs do not inflect for either person, tense, aspect, honor, or number. Bengali incorporates a marked system of pronominal reference that allows the expression of various degrees of familiarity, politeness, as well as spatial location (Thompson, 2012, p. 68). Grammatical gender is not attested. (UCLA Language Materials Project, 2006).

Syntactic level. Bengali is a head-final language whose main word order is SOV (UCLA Language Materials Project, 2006). While languages like English rely on prepositions, Bengali use postpositions that follow the noun-phrase; in addition, the postpositions are not a closed word class, but are either nouns in the locative case, or perfect participle verb forms (Thompson, 2012, p. 103). Below a simplistic overview of the four basic sentence patterns are described. For further details, the reader is advised to consult with the comprehensive descriptive grammar of Thompson (2012).

4.2.2 Predication

Simple predicative sentences follow the SOV order described above. The subject is very often a noun or pronoun, while the predicate can either hold of a verb plus any objects or locatives required by the verb, or a complement (Thompson, 2012, p. 185).

- (18) ækjon namkora jadukor jadu dækhaben.
 one.CL famous magician magic show.3H.FUT
 ‘A famous magician will show his magic.’

4.2.3 Possession

In Bangla, the possessor is always in genitive case, i.e., it is followed by a genitive suffix (c.f., *jutā-ṭa* ‘the shoe’ and *jutā-ṭa-r* ‘the shoe’s’). The formation of complex possessive constructs can either be done by the recursive concatenation of possessor constructs or by the usage of attributive pronouns (*amar, tomar* ‘my’) (Thompson, 2012, p. 128):

- (19) mayér biyér sômôykar dukhana curi.
 mother.GEN wedding.GEN time.GEN two.CL bracelet
 ‘two bracelets from the time of the wedding of my mother’

4.2.4 Imperative

Present tense imperative verb forms are identical to that of declarative sentences, the only difference being the omission of personal pronouns. Imperatives can be formed from active/agentive, existential and copular verbs. Imperative verbs can also be in their future tense, and in case of negative imperatives, future tense is obligatory (e.g., *na asuk!* ‘Let him not come!’). Besides second person imperatives, third person imperatives are also available in Bengali (instrumental for blessings, cursing, encouragement, etc.) (Thompson, 2012, p. 210–212).

- (20) alur cas suru kôrun na!
 potato cultivation start do.2H.PR.IMP PRT.reinforce!
 ‘Start growing potatoes!'

4.2.5 Interrogative

Bengali uses both yes-no and content interrogative clauses. In case of yes-no questions, the word order (subject, object, verb) does not differ from the word order found in indicative clauses, however yes-no questions contain the question marker *ki* inserted after the subject. Wh- (or content) interrogatives in Bangla all start with *k-* and can be pronouns, adjectives or adverbs (Thompson, 2012, p. 201–203). The example below illustrates the usage of the content interrogative *ke* ‘who’.

- (21) kɔthat bôleche ke?
 word.CL say.3.PR.PERF who
 ‘Who said that?’

4.3 Writing system, transcription

Native Bengali script is situated in the Unicode range [0980–09FF](#). Bengali people use an *abugida* alphabet (in this system consonants and vowels in a syllable are treated as clusters, i.e., each letter denotes a consonant with an inherent vowel) written from left to right. The script itself is of Brahmic origin and it is also used in other Indian languages (e.g., in Sylhet and Assamese). In this system, there are no upper or lower case variants, there are thirty-five consonant letters and eleven independent vowel letters (see Figure 4.4). For a thorough review, see [ScriptSource \(2015\)](#) and [Ager \(2016\)](#).

There are numerous ways to romanise the abugida script of Bengali, each having its advantages and disadvantages. Throughout this chapter, and especially in the segment with the romanised linguistic examples, a unique transliteration scheme was used (described in details in the introductory chapters of [Thompson 2012](#), p. xxiii). Transliteration conventions are orthographically-correct transformations whereby the original script is recoverable (e.g., National Library at Kolkata romanisation system, ITRANS, Harvard-Kyoto scheme, [ALA-LC Romanization Tables](#), etc). The other prominent way of romanizing scripts is through transcription. Transcriptions aim at being phonetically accurate, thus reproducing the original pronunciation of words (e.g., the Wiki transcription scheme). For a comparison of available transliteration/transcription methods, visit the Wikipedia site entitled [Romanisation of Bengali](#).

4.4 Previous research on the language

The present section provides a list on the available reference grammars of the Bengali language: [Mithun and Van der Wurff \(2009\)](#); [Radice \(2008\)](#); [Klaiman \(1998\)](#); [Radice \(1994\)](#); [Ray et al. \(1966\)](#), [Hudson \(1965b\)](#); [Dimock \(1964\)](#); [Beames \(1894\)](#).

For an exhaustive literary list on the available reference grammars, the reader is pointed to the [Bengali entry of OLAC](#). For a typological overview, visit the [Bengali section of WALS](#).

4.4.1 Journals

The list below provides an overview of Bengali journals dealing with topics related to linguistics:

- [Pratidhwani the Echo](#) (ISSN: 2278-5264, IF: 6.28) – multilingual journal centering around the humanities and social sciences
- [Journal of Bengali Studies](#) (ISSN: 2277-9426, IF: 4.956) – journal dedicated to the study of the history and culture of Indic Bengali people (general humanities journal)
- [International Journal of Humanities and Social Science Studies \(IJHSSS\)](#) (ISSN: 2349-6711, IF: 0.275) – bi-lingual (English–Bengali) journal on humanities and social sciences
- [Arts Faculty Journal](#) (ISSN: 1994-8891, IF: N/A) – journal published by the Arts Faculty at the University of Dhaka (last published in 2011)
- [Asian Journal of Humanity, Art and Literature \(AJHAL\)](#) (ISSN: 2312-2021, IF: N/A) – journal with a broad focus, encompassing issues related to Bengali and linguistics alike
- [BRAC University Journal of Humanities and Social Sciences \(BUJHSS\)](#) (ISSN: 1015-6836, IF: N/A) – journal dedicated to the exploration of the social sciences and humanities

Institute	Location		Ranking					
	city	country	QS world	ARWU	CWUR	CWTS	THE	webometrics
Harvard University	Cambridge	US	3	1	1	1	6	1
University of Washington	Seattle	US	59	15	27	10	32	7
University of Oxford	Oxford	UK	6	7	5	13	2	10
University of Texas at Austin	Austin	US	67	44	32	52	46	17
University of Chicago	Chicago	US	10	10	8	82	10	18
SOAS University of London	London	UK	252	N/A	N/A	N/A	401–500	964
Banaras Hindu University	Varanasi	India	701+	674	375	N/A	N/A	1242
University of Kalyani	Kalyani	India	N/A	N/A	N/A	N/A	N/A	1621
University of Calcutta	Kolkata	India	651–701	N/A	931	1383	601–800	1686
University of Burdwan	Burdwan	India	N/A	N/A	N/A	N/A	N/A	2328
Jadavpur University	Kolkata	India	N/A	N/A	921	572	501–600	2632
University of Karachi	Karachi	Pakistan	701+	N/A	N/A	N/A	801+	2673
University of Chittagong	Chittagong	Bangladesh	N/A	N/A	N/A	N/A	N/A	2810
Assam University	Assam	India	N/A	N/A	N/A	N/A	N/A	2860
Vidyasagar University	Midnapore	India	N/A	N/A	N/A	N/A	N/A	2940
Presidency University Kolkata	Kolkata	India	N/A	N/A	N/A	N/A	N/A	3814
Tripura University	Suryamaninagar	India	N/A	N/A	N/A	N/A	N/A	4125
West Bengal State University	Berunanpukuria	India	N/A	N/A	N/A	N/A	N/A	6726
Patna University	Patna	India	N/A	N/A	N/A	N/A	N/A	9316
Tilka Manjhi Bhagalpur University	Bhagalpur	India	N/A	N/A	N/A	N/A	N/A	9678
Rabindra Bharati University	Kolkata	India	N/A	N/A	N/A	N/A	N/A	12490
Pabna University of Science and Technology	Pabna	Bangladesh	N/A	N/A	N/A	N/A	N/A	14522
Cotton College	Guwahati	India	N/A	N/A	N/A	N/A	N/A	17044
Nagar College	Nagar	India	N/A	N/A	N/A	N/A	N/A	17275
Scottish Church College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	17609
Malda College	Malda	India	N/A	N/A	N/A	N/A	N/A	17793
Serampore College	Serampore	India	N/A	N/A	N/A	N/A	N/A	18239
Bankura Christian College	Bankura	India	N/A	N/A	N/A	N/A	N/A	18777
Bethune College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	18859
Lady Brabourne College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	18889
Ranchi University	Ranchi	India	N/A	N/A	N/A	N/A	N/A	19214
Shri Shikshayatan College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	19235
Surendranath College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	20471
Cooch Behar Panchanan Barma University	Cooch Behar	India	N/A	N/A	N/A	N/A	N/A	22802
Asian University of Bangladesh	Dhaka	Bangladesh	N/A	N/A	N/A	N/A	N/A	22945
Gurudas College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	23133
Bhairab Ganguly College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	24223
Ramananda College	Bishnupur	India	N/A	N/A	N/A	N/A	N/A	24275
Bankura Sammilani College	Bankura	India	N/A	N/A	N/A	N/A	N/A	24884
Darjeeling Government College	Darjeeling	India	N/A	N/A	N/A	N/A	N/A	25269
City College	Kolkata	India	N/A	N/A	N/A	N/A	N/A	25338

Table 4.1: List of academic units dealing with the Bengali language, supplemented with various measures of academic excellency.

- [Dhaka University Journal of Linguistics](#) (ISSN: 2075-3098, IF: N/A) – linguistics journal published by the Department of Linguistics at the University of Dhaka (last published in 2009)
- [The Asiatic Society of Bangladesh \(Humanities\)](#) (ISSN: 1015-6836, IF: N/A) – Bangladeshi journal publishing articles on general issues in the humanities (last published in 2014)

4.4.2 Research and control bodies

Table 4.1 summarizes the various academic institutes with a department of Bengali or a similar unit.

4.5 Data and sources

4.5.1 Basic vocabulary

The *An Crúbadán* project offers a [package of Bengali resources](#) (such as character trigrams, word bigrams and word frequency tables) compiled from 4030 documents, with a total of 8,255,819 words. Based on the Bengali Bible, a word frequency table is also available on the [academic site of Levente Madarász](#). (Login and password: DLD.) In addition, the [Bengali WordNet](#) consists of 39,563 words and 36,345 synsets.

Learning Bengali

For the acquisition of the language and its native writing system, the following materials are highlighted: [Lambert and Cantab \(1953\)](#); [Hudson \(1965a\)](#) and [Nasrin and Van Der Wurff \(2015\)](#).

4.5.2 Dictionaries

Paper-based Bengali dictionaries

Table 4.2 provides an overview of the major paper-based dictionaries, indicating the number of entries, as well as their length in pages.

Publication	Entries	Pages
Chaki 2006	40,000	1272
Biswas et al. 2003	>35,000	1992
Dev 1961	40,000	800
Banerjee and West 1958	21,000	510
Central Institute of Indian Languages 2011	12,000	536
Guha 2007	N/A	2500
Das 2011	N/A	2128
Ghosh and Lahiri 2003	N/A	1990
Aich and Ganguly 1999	N/A	1451
Ahmed 1993	N/A	1207
Ālī" et al. 1994	N/A	878
Anonymous 2015	N/A	540
Sykes 1874	N/A	299
Olsen 1967	N/A	216

Table 4.2: Major paper-based Bengali–English and English–Bengali dictionaries ordered by the number of their entries.

Online Bengali dictionaries

Table 4.3 provides an overview of the available Bengali–English as well as English–Bengali online dictionaries. Dictionaries that are mere wrappers of other engines are not included in the list.

4.5.3 Corpora

Monolingual

The largest monolingual corpus of Bengali is provided by the Bengali Wikipedia. As of 11/01/2016, the [Bengali Wikipedia](#) consisted of 472,846 pages. The [EMILLE/CIIL corpus](#) (consisting of approximately 92,799,000 words) contains monolingual, parallel and annotated corpora for different South Asian languages, among others, Bengali. A slightly different approach is manifested in the [SHRUTI Bengali Continuous ASR Speech Corpus](#). The corpus contains 7383 sentences produced by 34 speakers of the standard Bengali colloquial language of the West Bengal region of India. For mining Bengali text, a further approach might be the crawling of Bangladeshi websites, ending with a .BD top-level domain (TLD).

Bilingual

The aforementioned *EMILLE/CIIH corpus* provides a 200,000 word English, Bengali, Hindi, Punjabi, Gujarati and Urdu multi-parallel section. Besides, *The Indic multi-parallel corpus* is also available for Bengali; the corpus contains the translation of approximately 2000 Wikipedia sentences in Bengali, Hindi, Malayalam, Tamil, Telugu and Urdu languages (Birch et al., 2011). Bengali Microsoft and Ubuntu localization files are good candidates for building parallel corpora. The list below contains further candidates for bilingual corpus building:

- **Bengali Bible** in .HTML format, made available on [Wordproject®](#)
- **The Quran** in .XML format on [OPUS](#), compiled by the Tanzil project
- **The Book of Mormon** in .PDF format, available on the web page of [The Church of Jesus Christ of Latter-Day Saints](#)
- **The Universal Declaration of Human Rights** in .TXT format on the [web page of the The Unicode Consortium](#)

Online Bengali news portals

The most important Bengali news outlets are reviewed in Table 4.4. To give an impression on the quality of the individual sites, post lengths and posting frequencies were measured. Mean post lengths are indicated in characters (not counting whitespaces) and are calculated from a random sample of 10 post bodies (titles are not included in the figures, artifacts, such as code chunks, were removed). Since most sites do not sport a list of their material ordered by release date, post frequencies for individual portals were measured by noting down the release time of the first 10 posts appearing on the main site, starting from page top. The dates were then ordered from oldest to latest, and the difference of individual steps were calculated and converted to minutes. From this set of 9 difference values, mean and standard deviation values as well as ranges were calculated. The measurements were carried out between 04:30 AM and 06:00 AM, Bangladesh Standard Time (BST).

4.6 Computational tools

4.6.1 Language identification

For language identification purposes, the [Compact Language Detector 2 \(CLD\)](#) (161) is available for Bengali. [Polyglot3000](#) and the [langid.py LangID tool](#) are also compatible with Bengali (2015).

4.6.2 Tokenizer

MorphAdorner developed by Philip R. Burns supports Bengali (2013) and is [available online](#). Rami Al-Rfou's [polygot](#) also supports tokenization (2015). (The above tools are also capable of language identification.)

4.6.3 Stemmer

[Dolamic](#) (2010) created a light stemmer for removing number, gender and case suffixes from Bengali nouns and adjectives. Another stemmer developed by [Ganguly](#) (2014) is available [on GitHub](#). Further

	a	ā	i	ī	u	ō	au
ka	ক	କା	ি	ି	ୁ	୦	ାଉ
kha	খ	ଖା	କି	ି	ୁ	୦	ାଉ
ga	গ	ଗା	ଯି	ଯି	ୁ	୦	ାଉ
gha	ଘ	ଘା	ଜି	ଜି	ୁ	୦	ାଉ
ñia	ঙ	ଙା	ତି	ତି	ୁ	୦	ାଉ
ca	চ	ଚା	ବି	ବି	ୁ	୦	ାଉ
cha	ছ	ଛା	ବି	ବି	ୁ	୦	ାଉ
ja	জ	ଜା	ବି	ବି	ୁ	୦	ାଉ
jha	ঝ	ঝା	ବି	ବି	ୁ	୦	ାଉ
ñia	ঞ	ঞା	ବି	ବି	ୁ	୦	ାଉ
ta	ট	ଟା	ତି	ତି	ୁ	୦	ାଉ
tha	ঠ	ঠା	ତି	ତି	ୁ	୦	ାଉ
da	ড	ଡା	ତି	ତି	ୁ	୦	ାଉ
dha	ঢ	ঢା	ତି	ତି	ୁ	୦	ାଉ
na	ণ	ଣା	ତି	ତି	ୁ	୦	ାଉ
ta	ত	ତା	ତି	ତି	ୁ	୦	ାଉ
tha	থ	ଥା	ତି	ତି	ୁ	୦	ାଉ
da	দ	ଦା	ତି	ତି	ୁ	୦	ାଉ
dha	ধ	ଧା	ତି	ତି	ୁ	୦	ାଉ
na	ন	ନା	ତି	ତି	ୁ	୦	ାଉ
pa	প	ପା	ପି	ପି	ୁ	୦	ାଉ
pha	ফ	ଫା	ପି	ପି	ୁ	୦	ାଉ
ba	ব	ବା	ପି	ପି	ୁ	୦	ାଉ
bha	ভ	ଭା	ପି	ପି	ୁ	୦	ାଉ
ma	ম	ମା	ପି	ପି	ୁ	୦	ାଉ
ya	য	ଯା	ପି	ପି	ୁ	୦	ାଉ
ra	ର	ରା	ପି	ପି	ୁ	୦	ାଉ
la	ଲ	ଲା	ପି	ପି	ୁ	୦	ାଉ
śa	শ	ଶା	ପି	ପି	ୁ	୦	ାଉ
ṣa	ষ	ଷା	ପି	ପି	ୁ	୦	ାଉ
sa	স	ସା	ପି	ପି	ୁ	୦	ାଉ
ha	হ	ହା	ପି	ପି	ୁ	୦	ାଉ
ya	ঘ	ঘା	ପି	ପି	ୁ	୦	ାଉ
ra	ড	ଡା	ପି	ପି	ୁ	୦	ାଉ
tha	ঢ	ঢା	ପି	ପି	ୁ	୦	ାଉ

Figure 4.4: Bengali alphabet obtained from Ager (2016).

Dictionary	Developer	Year	Size	Quality				Scrapeability	
				IPA	POS	TRANSLITERATION	AUDIO	LIST	QUERY
Bengali to English Dictionary and Translation	BDWord	2009-2016	143,442	×				×	×
English & Bengali Online Dictionary & Grammar	English-Bangla	2016	>65,000	×				×	×
Ankur's English to Bengali dictionary	Ankur	2006	38,899	×					×
English Bengali Dictionary online	Glosbe	2011-2016	20,103	×					×
English Bengali Dictionary	Shyam Krishnan	2011	14,000					×	
Shabdkosh	Maneesh Soni	2003-2016	N/A	×		×		×	×
Samsad Bengali-English Dictionary	Shishu Sahitya Samsad	2004	N/A	×		×			×
Online Bangla Dictionary	Abdullah Ibne Alam	2015-2016	N/A	×					×
ALDictonary	Adept Leal Software	2015-2016	N/A	×					×

Table 4.3: Summary table of Bengali–English as well as English–Bengali online dictionaries. Signs in the Quality multicolumn indicate whether a page have IPA transcription for the entries, whether the part-of-speech of a given entry is listed and whether transliteration is available. Signs in the Scrapeability multicolumn indicate whether a word listing is available with hyper links to the entries, and whether an URL-query is possible.

Portal	Ranking		Quality		
	Daily unique visitors	Rank in India	Mean post length (N=10)	Update frequency	Video content
Prothom Alo	346,000	5	3343.1±1900.3 [1457-6792]	20.44±31.56 [0-100]	×
ManabZamin	107,491	N/A	2419.9±2079.84 [526-7391]	38.78±37.36 [8-131]	
Kaler Kantha	86,000	10	4047±2045 [1457-7642]	1.78±2.17 [0-6]	×
Bangladesh Protidin	62,593	51	2070.9±2042.98 [281-7346]	104.63±155.31 [0.27-447.97]	×
Ittefaq	61,100	28	2447.3±3886.73 [257-13099]	42.56±97.64 [0-302]	
Noya Diganta	45,000	56	5162.7±4380.04 [1829-16485]	N/A	
Jugantor	44,152	N/A	3416.5±2837.53 [815-9955]	47.91±83.27 [3.33-265.83]	×
Shamokal	29,350	91	3030.5±1801.74 [1200-7454]	12.11±20.68 [1-66]	
Janakantha	20,562	N/A	6039.3±1734.52 [3402-8749]	N/A	×
Bhorer Kagoj	18,999	N/A	1029±410.69 [509-1915]	N/A	×
Inqilab	16,687	N/A	985.2±499.76 [522-2212]	13.78±25.97 [3-82]	
Alokito Bangladesh	15,970	174	3332.3±1450.46 [1083-5828]	N/A	
Sangram	11,145	N/A	5160.6±3143.46 [961-9761]	N/A	
Manobkantha	10,750	247	2919.6±2377.2 [971-9273]	99±136.24 [10-425]	
Amader Shomoy	9,400	N/A	2576.6±1739.78 [649-6732]	N/A	
Shokaler Khabor	5,615	527	4817.7±2779.66 [1564-8953]	N/A	
Jai Jai Din	4,400	468	2575.6±2582.14 [913-9461]	49.89±45.07 [0-138]	
Sangbad	3,918	N/A	3303.1±2622.33 [835-9883]	27.33±44.03 [0-111]	×
Ajker Patrika	2,800	1710	2339.8±2839.26 [454-9620]	70.7±52.9 [4.33-180.65]	
Dinkal	1,106	N/A	960.2±478.4 [420-1992]	17.44±22.61 [0-74]	
Protidiner Sangbad	1,100	1889	1720.1±1134.07 [585-3919]	N/A	×
Ajkaler Khobor	685	3224	2911.7±3380.7 [693-11262]	87.33±205.78 [0-634]	×

Table 4.4: Bengali news portals ordered by the number of their daily unique visitors. In addition, the table also shows how individual sites are ranked in Bangladesh (the country with the most Bengali speakers). As an indicator of ‘quality’, the mean character length of a random sample of posts and mean post lengths are indicated (see Section 4.5.3). The table also highlights if a site hosts video content. N/A values in the Update frequency column indicate that a given outlet does not include the publication time of its posts.

articles on Bengali stemming: [Ganguly et al. \(2012\)](#); [Sarkar and Bandyopadhyay \(2008\)](#); [Islam et al. \(2007\)](#).

4.6.4 Spell checker

Open source spell checkers include [HunSpell](#) and [Mozilla dictionaries](#) (the latter is available for both West Bengal and Bangladeshi Bengali). The [Stars21](#) free online service also supports Bengali.

4.6.5 Phrase level and higher tools

In this section a collection of references are provided, each dealing with higher-level computational tools for Bengali:

- **Bengali part-of-speech tagger:** [Sarkar and Gayen 2012](#); [Ekbal et al. 2008b](#); [Ekbal and Bandyopadhyay 2008b](#); [Dandapat et al. 2007](#); [Ekbal et al. 2007](#); [Dandapat and Sarkar 2006](#)
- **Bengali named entity recognizer:** [Hasanuzzaman et al. 2009](#); [Ekbal and Bandyopadhyay 2008a](#); [Ekbal et al. 2008a](#); [Ekbal and Bandyopadhyay 2007](#)
- **Bengali chunker:** [De et al. 2011](#); [Bandyopadhyay et al. 2006](#)
- **Bengali morphological parser:** [Dasgupta and Ng 2006](#)
- **Bengali sentence parser:** [Ghosh et al. 2010](#); [Mursheed 1998](#)
- **Bengali question answering system:** not available
- **Bengali speech recognizer:** [Ali et al. 2013](#); [Banerjee et al. 2008](#)
- **Bengali machine translator:** [Francisca et al. 2011](#); [Islam et al. 2010](#); [Dasgupta et al. 2004](#); [Vijayanand et al. 2002](#); [Google 2006](#) and [Yandex Translator](#)

4.6.6 End-user support

OS support

- **Mac OSX** (as of 2016-10-28): [No OS-level support](#)
- **Microsoft Windows** (as of 2016-10-28): [No language packs are available](#)
- **Ubuntu** (as of 2016-10-28): [Partial translation](#) (80% untranslated)

Browser compatibility

Among the web browsers, Bangla is supported by Google Chrome, [Mozilla Firefox](#), [Internet Explorer](#), and [Opera](#). As of 2017-01-10, Safari does not support the Bengali language.

Bibliography

- S. Ager. Bengali – omniglot, 2016. URL <http://www.omniglot.com/writing/bengali.htm>. [Online; accessed 04-Nov-2016].
- S.K. Ahmed. *Joy Advanced Learner's Dictionary: English – Bengali – English*. Gyankosh Prokashony, 1993.
- N. Aich and P.N. Ganguly. *Progressive English-Bengali Dictionary: English-to-Bengali & English*. Indian Progressive Publishing Company, 1999.
- R. Al-Rfou. polyglot, 2015. URL <http://polyglot.readthedocs.org/en/latest/index.html#>.
- M. Ālī", M. Manirujjāmāna, J. Tareque, and Bāmlā Ekādemī (Bangladesh). *Bangla Academy Bengali-English dictionary*. The Academy, 1994.
- M. Ali, M. Hossain, M. N. Bhuiyan, and et al. Automatic speech recognition technique for Bangla words. *International Journal of Advanced Science & Technology*, 50, 2013.
- Anonymous. *A Dictionary English-Bengali-Hindouistani, in the Roman Character, with Walker's Pronociation*. BiblioBazaar, 2015.
- S. Bandyopadhyay, A. Ekbal, and D. Halder. Hmm based pos tagger and rule-based chunker for Bengali. In *Proceeding of the NLPAI Machine Learning Competition*, 2006.
- H.C. Banerjee and M.P. West. *The new method English-Bengali dictionary, explaining the meaning of about 21,000 items within a vocabulary of 1,490 words*. Orient Longmans, 1958.
- P. Banerjee, G. Garg, P. Mitra, and A. Basu. Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.
- J.K. Banthia, J.L. Modi, P. Bansod, P. Hazarika, S.K. Rakesh, Rajasthan India. Directorate of Census Operations, Assam India. Director of Census Operations, Bihar India. Director of Census Operations, Madhya Pradesh India. Director of Census Operations, India. Office of the Registrar General & Census Commissioner, et al. *Census of India, 2001: Assam. Paper*. Census of India, 2001: Paper. Controller of Publications, 2001. URL <https://books.google.hu/books?id=i8EUAQAAQAAJ>.
- J. Beames. *Grammar of the Bengali language, literary and colloquial*. Oxford Oriental Series. Oxford, Clarendon Press, 1894.
- S. Bhattacharya, M. Choudhury, S. Sarkar, and A. Basu. Inflectional morphology synthesis for bengali noun, pronoun and verb systems. In *In Proceedings of the national conference on computer processing of Bangla (NCCPB*, 2005.

- L. Birch, C. Callison-Burch, M. Osborne, and M. Post. Indic multi-parallel corpus, 2011. URL <http://homepages.inf.ed.ac.uk/miles/babel.html>.
- S. Biswas, S. Bhattacharya, and S. Sengupta. *Samsad English-Bengali Bengali-English Dictionary*. Laurier Books Limited, 2003.
- P. R. Burns. Morphaditioner v2.0, 2013. URL <https://devadortner.northwestern.edu/maserver/>.
- Central Institute of Indian Languages. *Longman-Cil: English-English-Bengali Dictionary*. Pearson Education, 2011. URL <https://books.google.hu/books?id=EqG08sF0QaYC>.
- J.B. Chaki. *Dev's students' favourite dictionary: English-to-Bengali and English*. Dev Sahitya kutir, 2006.
- S. K. Chatterji. *The origin and development of the Bengali language*. The Origin and Development of the Bengali Language. Calcutta University Press, 1926.
- S. Dandapat and S. Sarkar. Part of speech tagging for Bengali with hidden Markov model. In *Proceeding of the NLPAI Machine Learning Competition*, 2006.
- S. Dandapat, S. Sarkar, and A. Basu. Automatic part-of-speech tagging for Bengali: An approach for morphologically rich languages in a poor resource scenario. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 221–224. Association for Computational Linguistics, 2007.
- P.C. Das. *A Complete English-Bengali Dictionary: With British & American Pronunciation, Illustrations in Bengali & English, Synonyms, Antonyms, Phrasal, Verbs and Usage*. Education Foundation, 2011.
- S. Dasgupta and V. Ng. Unsupervised morphological parsing of Bengali. *Language Resources and Evaluation*, 40(3-4):311–330, 2006.
- S. Dasgupta, A. Wasif, and S. Azam. An optimal way of machine translation from English to Bengali. In *Proceedings Of 7 th International Conference On Computer and Information Technology*, pages 648–653, 2004.
- S. De, A. Dhar, S. Biswas, and U. Garain. On development and evaluation of a chunker for Bangla. In *Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on*, pages 321–324. IEEE, 2011.
- A.T. Dev. *Concise dictionary: English to Bengali & English*. S.C. Mazumder, 1961.
- Edward Dimock. *Introduction to Bengali, part i [microform] / Edward Dimock, Jr. and Others*. Distributed by ERIC Clearinghouse [Washington, D.C.], 1964. URL <http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED012811>.
- L. Dolamic. Bengali light stemmer, 2010. URL <http://members.unine.ch/jacques.savoy/clef/BengaliStemmerLight.java.txt>.
- M. S. Dryer and M. Haspelmath, editors. *Language Bengali*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL http://wals.info/languoid/lect/wals_code_ben.

- A. Ekbal and S. Bandyopadhyay. A hidden Markov model based named entity recognition system: Bengali and Hindi as case studies. In *Pattern Recognition and Machine Intelligence*, pages 545–552. Springer, 2007.
- A. Ekbal and S. Bandyopadhyay. Bengali named entity recognition using support vector machine. In *IJCNLP*, pages 51–58, 2008a.
- A. Ekbal and S. Bandyopadhyay. Part of speech tagging in Bengali using support vector machine. In *Information Technology, 2008. ICIT'08. International Conference on*, pages 106–111. IEEE, 2008b.
- A. Ekbal, R. Haque, and S. Bandyopadhyay. Bengali part of speech tagging using conditional random field. In *Proceedings of Seventh International Symposium on Natural Language Processing (SNLP2007)*, pages 131–136, 2007.
- A. Ekbal, R. Haque, and S. Bandyopadhyay. Named entity recognition in Bengali: A conditional random field approach. In *IJCNLP*, pages 589–594, 2008a.
- A. Ekbal, R. Haque, and S. Bandyopadhyay. Maximum entropy based Bengali part of speech tagging. *A. Gelbukh (Ed.), Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal*, 33:67–78, 2008b.
- Encyclopædia Britannica. Bengali people, 2015. URL <http://www.britannica.com/EBchecked/topic/60782/Bengali>.
- A. B. M. R. K Faquire. On the classification of varieties of Bangla spoken in Bangladesh. *BUP Journal*, 1(1):136, 2012.
- J. Francisca, M. Mia, and M. Rahman. Adapting rule based machine translation from English to Bangla. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(3):334–342, 2011.
- D. Ganguly. Bengali stemmer, 2014. URL <https://github.com/gdebasis/BengaliStemmer>.
- D. Ganguly, J. Leveling, and G. J. F. Jones. Dcu@ fire-2012: rule-based stemmers for Bengali and Hindi. In *FIRE 2012 Workshop*, 2012.
- Aniruddha Ghosh, Amitava Das, and Sivaji Bandyopadhyay. Clause identification and classification in Bengali. In *23rd International Conference on Computational Linguistics*, page 17, 2010.
- G.P. Ghosh and A.K. Lahiri. *Everyman's dictionary: English-Bengali*. Ramakrishna Pustakalaya, 2003.
- Google. Google translate, 2006. URL <http://translate.google.com/>.
- George Abraham Grierson. *Linguistic survey of India*, volume 5. Office of the superintendent of government printing, India, 1903.
- C. Guha. *The Modern Anglo-Bengali Dictionary*. The Modern Anglo-Bengali Dictionary. Kalpaz Publications, 2007.
- H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 2.z – Bengali, 2016. URL <http://glottolog.org/resource/languoid/id/beng1280>. [Online; accessed 3-November-2016].

- M. Hasanuzzaman, A. Ekbal, and S. Bandyopadhyay. Maximum entropy approach for named entity recognition in Bengali and Hindi. *International Journal of Recent Trends in Engineering*, 1(1): 408–412, 2009.
- D. F. Hudson. *Teach Yourself Bengali*. The English Universities Press Ltd, 1965a.
- D. F. Hudson. *Teach Yourself Bengali*. The English Universities Press Ltd, 1965b.
- Z. Islam, N. Uddin, M. Khan, and et al. A light weight stemmer for Bengali and its use in spelling checker, 2007. manuscript.
- Z. Islam, J. Tiedemann, and A. Eisele. English to Bangla phrase-based machine translation. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, 2010.
- M.H. Klaiman. *A Bengali Reference Grammar*. Oxford University Press, Incorporated, 1998. ISBN 9780195081060. URL <https://books.google.hu/books?id=ZsJmkQEACAAJ>.
- H. M. Lambert and M. A. Cantab. *Introduction to the Devanagari Script for students of Sanskrit, Hindi, Marathi, Gujarati, and Bengali*. Oxford University Press, 1953.
- M. Lui and T. Baldwin. langid.py, 2015. URL <https://github.com/saffsd/langid.py>.
- S. Mahanta. *Directionality and locality in vowel harmony: With special reference to vowel harmony in Assamese*. 2008.
- B. Mithun and W. Van der Wurff. *Colloquial Bengali: The Complete Course for Beginners*. Colloquial Series. Routledge, bilingual edition, 2009. ISBN 9780415261197.
- Md Manzoor Murshed. Parsing of Bengali natural language sentences. In *Proceedings of ICCIT, Dhaka*, pages 185–189, 1998.
- M. B. Nasrin and W. A. M. Van Der Wurff. *Colloquial Bengali*. Colloquial series. Taylor & Francis, 2015.
- V.B. Olsen. *Muslima Bāmlā Imrejī abhidhāma: A Muslim Bengali-English Dictionary*. Pakistan Co-operative Book Society, 1967.
- Online Etymology Dictionary. Word origin and history for bengal, 2015. URL <http://dictionary.reference.com/browse/bengal>.
- L. M. Paul, G. F. Simons, and D. F. Charles. Ethnologue: Languages of the World, 2015. URL <http://www.ethnologue.com>.
- L. M. Paul, G. F. Simons, and D. F. Charles. Bengali - ethnologue, 2016. URL <https://www.ethnologue.com/language/ben>. [Online; accessed 21-November-2015].
- W. Radice. *Bengali: A Complete Course for Beginners*. Teach Yourself Languages Series. Hodder Education Group, 1994. ISBN 9780340552575. URL <https://books.google.com.au/books?id=EySkQgAACAAJ>.
- W. Radice. *Teach Yourself Bengali*. Book + 2CD's TY: Complete Courses. McGraw-Hill, 3 edition, 2008. ISBN 0071546944,9780071546942.

- Punya Sloka. Ray, Muhammad. ‘Abd al Hayy, and Lila. Ray. *Bengali language handbook*. Center for Applied Linguistics [Washington], 1966.
- K. Sarkar and V. Gayen. A practical part-of-speech tagger for Bengali. In *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*, pages 36–40. IEEE, 2012.
- S. Sarkar and S. Bandyopadhyay. Design of a rule-based stemmer for natural language text in Bengali. In *IJCNLP*, pages 65–72. Citeseer, 2008.
- ScriptSource. Scriptsource - writing systems, computers and people, 2015. URL <http://scriptsource.org>.
- J. Sykes. *English and Bengali Dictionary, for the Use of Schools*. Rипol Classic, 1874.
- The LINGUIST List. Multitree: A digital library of language relationships, 2014. URL <http://multitree.org/>.
- H. R. Thompson. *Bengali*. London Oriental and African Library. John Benjamins Publishing, 2012.
- UCLA Language Materials Project. Bengali, 2006. URL <http://lmp.ucla.edu/Profile.aspx?LangID=84&menu=004>.
- K. Vijayanand, S. I. Choudhury, and P. Ratna. Vaasaanubaada: automatic machine translation of bilingual Bengali-Assamese news texts. In *Language Engineering Conference, 2002. Proceedings*, pages 183–188. IEEE, 2002.

Chapter 5

Hindi (Levente Madarász)

Contents

5.1 Demography and ethnography	67
5.2 Main typological and syntactic features	69
5.3 Writing system, transcription	72
5.4 Previous research on the language	73
5.5 Data and sources	73
5.6 Computational tools	76
Bibliography	79

Introduction

This chapter provides an overview of the Hindi language, mainly spoken in South Asia, especially in India. Following a language historical overview and an account on the speaker communities as well as dialectal variants, the study introduces the writing system utilized by the language, the available reference grammars and different corpora. The final section details digital resources and computational tools.

5.1 Demography and ethnography

5.1.1 Name variants

The word ‘Hindi’ (हिन्दी) stems from Persian, and signifies a native of India—more precisely: a Musalman individual. Originally, Non-Musalman Indians were referred to as ‘Hindu’ (Grierson cited in Rai, 1984). As a reference for the lingua, it was used as early as the 14th century, designating the language spoken in the North-Western part of present-day India. Alternate names include Modern Hindi, High Hindi, Nagari Hindi, Literary Hindi, Standard Hindi, Khari Boli, Khadi Boli and Hindustani (Hammarström et al., 2015). The ISO identifiers of the language are *hi*, *hin* and *hin* (ISO 639-1, 639-2 and 639-3 respectively) (SIL International, 2015).

5.1.2 Geographic spread

Figure 5.1 displays Hindi speaking areas around the world. Hindi is mainly spoken in north India where it is a statutory national, EGIDS level 1 language. Significant speaker populations also reside

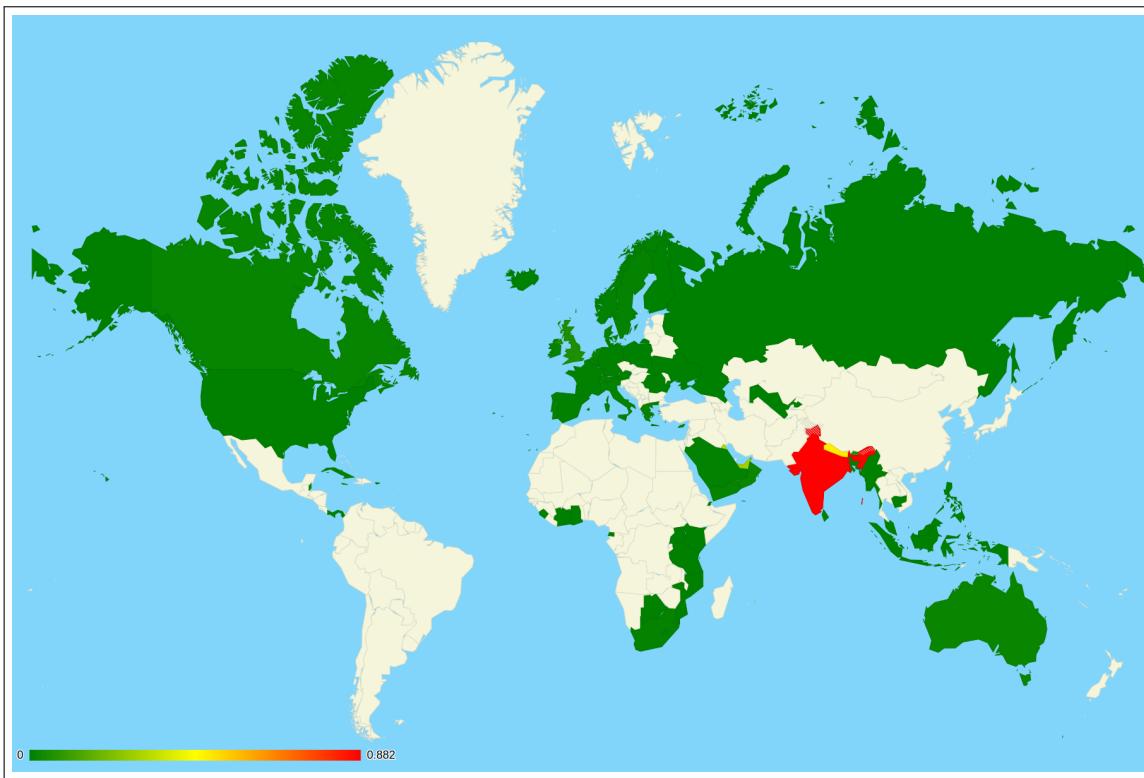


Figure 5.1: Map of Hindi speaking areas ([Joshua Project, 2016](#)). Coloring indicates relative salience of the language in a given country; near 0% is indicated by green, while near 100% spread by red, in gray countries no Hindi speaker communities are attested.

in Nepal (EGIDS level 3 language, used for wider communication), in Singapore (EGIDS level 5, dispersed language) and in the KwaZulu-Natal province of South Africa (EGIDS level 5, dispersed language) ([Paul et al., 2015](#)). While no EGIDS status is provided, the [Joshua Project \(2016\)](#) points out a series of other locations where Hindi speakers make up a part of the population. Table 5.1. provides a detailed overview of those countries where Hindi L1 speakers exceed 5,000.

5.1.3 Speaker populations

Modern Hindi—or as it is often referred to, Hindi—belongs to the Hindustani macro-language, a descendant of the Western Hindi branch of the neo-Aryan dialects spoken in the midland region of India often referred to as the Hindi belt ([Chand, 1944](#)). Since, Sanskrit influenced Modern Hindi and Urdu, a variant of Hindustani that is influenced by Persian-Arabic traditions, tend to coincide in colloquial usage, the estimation of Hindi speaker populations is problematic. Furthermore, any precise estimation is flanked by governmental efforts to mimic linguistic homogeneity: as [Deo \(2017\)](#) observes it is common practice to subsume minority languages under the regional standard determined by political boundaries. To make the matters even more complicated, the political and socioeconomic motives of post-independent India prompted the misclassification of several independent neo-Aryan languages (such as Braj, Awadhi, Kanauji, Bundeli, Bagheli and Bangaru) as dialects of Hindi. As an outcome, the inhabitants of the Hindi belt can be divide into five categories ([Khubchandani, 1997](#)):

- Bilingual users who view their own primary speech as a substandard variety of the prestigious Modern Hindi.

- Bilingual users who use a neo-Aryan variety in their intimate rural milieu and Modern Hindi in modern settings.
- Bilingual users of a neo-Aryan variety and Modern Hindi who relegate Modern Hindi to written use.
- Illiterate monolingual users who only speak a misclassified neo-Aryan variety (but nevertheless report their language as Hindi).
- Monolingual users of Modern Hindi.

Accordingly, available estimates are somewhat discordant. As of 2016-10-10, the results of the linguistic survey of the most recent 2011 Census of India are not available yet. In the 2001 Census of India, 422,048,642 speaker returned Hindi as their mother tongue, however 164,129,007 of these people actually spoke other neo-Aryan dialects ([Government of India, 2001](#)). This number is expected to rise as the population of the country produced a 17.7% decadal growth ([Government of India, 2011](#)).

In addition, 162,642,730 people worldwide use the aforementioned Urdu Hindustani variant, which is often considered linguistically identical with Hindi ([Paul et al., 2016](#)). Census records of other countries also uncover significant Hindi speaking communities: according to the results of the 2011 census a total of 77,600 L1 Hindi speakers and a further 490,000 L2 speakers reside in Nepal, Singapore houses 13,100 users (2010 census), while South Africa (2003 census) incorporates a 361,000 Hindi speaking community ([Paul et al., 2015](#)).

In the absence of up to date governmental figures, it is worth mentioning the work carried out by various Christian denominations with the purpose of quantifying the presence of gospel among various ethno-linguistic groups. The data frames published by [Joshua Project \(2016\)](#) based on data collected from the International Mission Board of the Southern Baptist Convention list all speaker communities of Hindi around the globe. According to these measures, Hindi is spoken by 664,346,400 as a primary language and another 519,209,200 people as a secondary language (see Table 5.1. for a detailed overview, and Figure 5.1 for a map of population normed speaker ratios).

5.1.4 Dialect situation

Due to the fact that Hindi is used both as an umbrella term covering Western and Eastern neo-Aryan languages (see Section 5.1.3.) and also as a reference to an individual language (i.e., Modern Hindi) ([Deo, 2017](#), p. 7), the literature on the true dialectal varieties of the language is virtually nonexistent ([Paul et al., 2015](#)).

To obtain a detailed overview of the linguistic details of neo-Aryan languages subsumed under Hindi, the reader is pointed to the seminal work of [Masica \(1993\)](#).

5.2 Main typological and syntactic features

Like many other languages in the midland region of India, Modern Hindi developed from Shauraseni Apabhraṃśa, the mother of all neo-Aryan dialects. Khari Boli, the vernacular out of which Modern Hindi evolved, belongs to the Western Hindi branch of Indo-Aryan languages ([Chand, 1944](#)). Khari Boli, or Hindustani, eventually developed two literary registers: Urdu, incorporating more Persian loans in its lexicon, and Modern Hindi, in which the loans are replaced with that of Sanskrit origin ([Masica, 1993](#), p. 27).

Country	Primary	Secondary	Grand Total	Population ratio
India	657,008,000	499,241,700	1,156,249,700	88.19%
United Arab Emirates	2,551,000		2,551,000	27.86%
Kuwait	1,224,000		1,224,000	31.45%
Canada	464,000		464,000	1.29%
South Africa	432,000		432,000	0.79%
United States	343,000		343,000	0.11%
Yemen	294,000		294,000	1.10%
France	273,000		273,000	0.41%
Netherlands	222,000		222,000	1.31%
Germany	189,000		189,000	0.23%
Congo, Democratic Republic of	171,000		171,000	0.22%
Saudi Arabia	160,000		160,000	0.51%
Myanmar (Burma)	132,000		132,000	0.24%
Oman	95,000		95,000	2.12%
Jamaica	93,000		93,000	3.41%
Indonesia	89,000		89,000	0.03%
Portugal	80,000		80,000	0.77%
Tanzania	63,000		63,000	0.12%
Sweden	60,000		60,000	0.61%
Malaysia	57,000		57,000	0.19%
Mauritius	37,000	1,053,000	1,090,000	26.80%
Russia	35,000		35,000	0.02%
Cuba	35,000		35,000	0.49%
Mozambique	31,000		31,000	0.11%
Ireland	31,000		31,000	0.67%
Austria	22,000		22,000	0.26%
Singapore	14,000		14,000	0.25%
Panama	14,000		14,000	0.36%
Sierra Leone	13,000		13,000	0.20%
Ghana	11,000		11,000	0.04%
Finland	11,000		11,000	0.20%
Swaziland	9,300		9,300	0.72%
Ukraine	9,100		9,100	0.02%
Belize	7,300		7,300	2.03%
St Vincent and Grenadines	5,900		5,900	5.39%
Kenya	5,900		5,900	0.01%
Saint Lucia	5,400		5,400	2.92%
Brunei	5,200		5,200	1.23%
...
world	664,346,400	519,209,200	1,183,555,600	16.11%

Table 5.1: Summary of Hindi speaker populations, based on data obtained by the [Joshua Project \(2016\)](#). Table is truncated so that only countries with an L1 speaker community of at least 5,000 is displayed (grand sums incorporate unlisted values). Population ratios were calculated by relying on the measures of the [World Bank \(2015\)](#).

5.2.1 Linguistic typology

Phonological level. Hindi maintains phonemic contrast between voiced and unvoiced stops and fricatives (Dryer and Haspelmath, 2013). The utilized consonant inventory is extensive as aspiration is a phonemic feature in the language. Hindi is not a tone language and it entertains no vowel harmony. The syllable structure of the language is (C)(C)V(C)(C). Stress typically falls on the heaviest syllable, and to avoid stress clash where syllables are equally heavy, stress usually falls on the rightmost syllable UCLA Language Materials Project (2006).

Morphological level. Hindi is a largely suffixing language with productive full and partial reduplication (Dryer and Haspelmath, 2013).

Morphosyntactic level. Hindi nouns show morphological marking only for number (singular or plural) and case (direct or oblique) (Singh and Sarma, 2010). Verbs are inflected for tense (past, present, future), aspect (habitual, progressive, perfective, completive), mood (imperative, subjunctive, presumptive, root conditional, condition), gender-number (feminine-singular, feminine-plural, masculine-singular, masculine-plural), person-number and voice (Singh and Sarma, 2011). Hindi also uses a sex-based two-level gender marking system (Dryer and Haspelmath, 2013).

Syntactic level. The typical word order in Hindi is Sov (Dryer and Haspelmath, 2013). Below a simplistic overview of the four basic sentence patterns are described. For further details, the reader is advised to consult with the comprehensive descriptive grammar of Kachru (2006).

5.2.2 Predication

Predicates with intransitive verbs have a simple argument structure, e.g.:

- (22) *Am pəke h̥e̥.*
mango.M.Pl ripen.PERF.M.PL be.PRES.PL
'Mangoes are ripe.'

Multivalent verbs display a more complex structure, whereby subjects are followed by the (indirect object and the) direct object (Kachru, 2006, p. 173–174):

- (23) *Mēne pita ji̥ ko cītt^hi̥ lik^hi̥.*
I AG father HON DAT letter.F
'I wrote a letter to (my) father.'

5.2.3 Possession

Possession is indicated by a construction in which the possessor noun is followed by a postposition (Kachru, 2006, p. 193).

- (24) *Ram ke do bētiyḁ̄ h̥e̥.*
Ram POSS two daughter.F.PL be.PRES.PL
'Ram has two daughters.'

5.2.4 Imperative

Depending on the intended degree of politeness, Hindi can express five direct imperative forms (for more details see (Kachru, 2006, p. 178–179)). The basic syntactic frame in all cases is the following:

- (25) *Yəh cīt^hī pət^ho!*
 this letter read.2nd.P.FAM
 'Read this letter!'

5.2.5 Interrogative

The valency patterns observed in predicative sentences are unchanged in interrogative ones (Kachru, 2006, p. 174):

- (26) *Tum səb ko k^hana pəros dogī?*
 you all DAT meal serve give.FUT.F.PL
 'Will you serve food to everyone?'

5.3 Writing system, transcription



अ	आ	इ	ई	उ	ऊ	ए		क	ख	ग	घ	ड	च	छ	ज	झ	ञ	ঞ	ঞ
a	ā	i	ī	u	ū	e		[ka]	[kʰa]	[ga]	[gʰa]	[n̥a]	[ca]	[cʰa]	[ja]	[jha]	[ñ̥a]	[ñ̥a]	[ñ̥a]
[ʌ]	[a]	[i]	[i:]	[u]	[u:]	[e]		[ta]	[tʰa]	[da]	[dʰa]	[na]	[ta]	[tʰa]	[da]	[dʰa]	[na]	[na]	[na]
প	পা	পি	পী	পু	পূ	পে		প	ফ	ব	ভ	ম	য	র	ল	ব			
pa	pā	pi	pī	pu	pū	pe		[pa]	[pʰa]	[ba]	[bʰa]	[ma]	[ya]	[ra]	[la]	[va]			
ঞ	ঞা	ঞি	ঞী	ঞু	ঞূ	ঞে		শ	ষ	স	হ								
ai	o	au	añ	aḥ	ām	r		śa	sa	sa	ha								
[ɛ:]	[o]	[ɔ:]	[aŋ]	[əh]	[ã]	[r]		[ʃa]	[sə]	[sə]	[fiə]								
ঞে	ঞো	ঞৌ	ঞঁ	ঞঃ	ঞঁ	ঞে													
পাই	পো	পাউ	পানি	পাহ	পাম্প	প্ৰ													
paɪ	po	pau	pari	pah	pamp	pr													

(a) Devanāgarī vowels and vowel diacritics (latter highlighted with red).

ক্ষ	খ্ল	গ্র	ঞ্জ	ঞ্চ	ঞ্ফ	ঞ্ড	ঞ্ঢ	
qa	ħa	ga	za	zha	fa	ra	ħħa	
[qə]	[ħə]	[gə]	[zə]	[zħə]	[fə]	[rə]	[ħħə]	
Common conjunct consonants								
শ্ব	ঞ্ব	ত্ব	দ্ব	দ্য	দ্ব	ত্ত	ঞ্জ	ঞ্জ
kṣa	ħħa	t̥ka	d̥va	d̥ya	d̥da	t̥ta	ħħħa	ħħħa
ঞ্ম	ঞ্ম	হ্য	শ্র	ত্র	ৰ্প	প্র	ঞ্ট	ঞ্ট
dma	ħħma	hya	śra	tra	rpa	pra	ħħħa	ħħħa

(b) Devanāgarī consonants, additional consonants and common consonant conjunts.

Figure 5.2: Devanāgarī script illustration and transliteration obtained from Ager (2016).

Since the 11th century, Hindi utilizes the alphasyllabary Devanāgarī writing system illustrated on Figure 5.2 (Ager, 2016). The script is written from left to right, and letters hang from a headstroke, which is usually continuous throughout a single word. The alphabet consists of 32 consonants, these contain an inherent [ə] vowel, which can be modified using one of the ten vowel diacritics. In addition,

the language contains ten vowel letters ([ScriptSource, 2016](#)). Devanāgarī is situated in the Unicode range 0900–097F ([Unicode, Inc., 2016](#)).

5.4 Previous research on the language

The earliest grammar of the Hindi language was compiled with the purpose of providing a practical study book for the Civil Service Commissioners working in British India ([Kellogg, 1876](#)). Later on, in his seminal work entitled *Linguistic survey of India*, [Grierson \(1903\)](#) ventured to sketch the linguistic map of the Indo-Aryan languages spoken in India, dedicating an entire volume to the Western Hindi branch. A contemporary description of the phonology, morphology, syntax, as well as information structural relations of the language is offered by the monograph of [Kachru \(2006\)](#), while a thorough overview of those Indo-Aryan languages that are often suggested to be the dialects of Hindi is provided by [Masica \(1993\)](#). For further resources, the reader is pointed to the [OLAC entry on Hindi](#).

5.5 Data and sources

5.5.1 Basic vocabulary

The *An Crúbadán project* offers a set of [Hindi resources written with Devanagari script](#) (consisting of character trigrams, word bigrams and word frequency tables) compiled from 1552 documents with a total of 15,715,354 words. The same team has also compiled a set of [Hindi resources written with Latin script](#). This set was compiled from 1261 documents with a total of 2,896,210 words. In addition, the [Hindi WordNet](#) consists of 63,800 words and 28,687 synsets.

Publication	Entries	Pages
Vira 1962	150,000	N/A
Kumar and Sahāya 2014	50,500	1518
Bahri 2002	35,500	460
Avasthī and Avasthī 1981	30,000	1623
Tivārī 1990	15,000	399
Singh et al. 2011	12,000	574
Tivārī et al. 1998	N/A	1406
Kapoor 2009	N/A	1104
Bulcke 2005	N/A	1032
Singh 2001	N/A	972

Table 5.2: Major paper-based Hindi–English and English–Hindi dictionaries ordered by the number of their entries.

5.5.2 Dictionaries

Paper-based Hindi dictionaries

Table 5.2 provides an overview of the major paper-based dictionaries, indicating the number of entries, as well as their length in pages.

Dictionary	Developer	Year	Size	Quality				Scrapeability		
				IPA	POS	LATIN	SCRIPT	AUDIO	LIST	
Collins English to Hindi Dictionary	Collins	2016	>250,000	×	×				×	×
Hindicube.com	Comsys Technologies Pte. Ltd.	2015	>200,000						×	×
HamariWeb.com	Abrar Ahmed	2016	142,000			×				×
Universal Word - Hindi Dictionary	Pushpak Bhattacharyya	2005	136,111	×		×				
Shabdkosh	Maneesh Soni	2003-2016	15,000–99,000	×		×		×	×	×
HinKhoj Hindi Dictionary	HinKhoj InfoLabs Llp.	2007-2016	7800–>63,800	×		×		×	×	×
Hindi Dictionary	Shyam Krishnan	N/A	14,000–40,000						×	
A Practical Hindi-English Dictionary	Mahendra Caturvedi	2008	N/A	×	×		×		×	×
Hindlish.com	Wordtech Co. Ltd.	2012-2016	N/A	×	×				×	×
bab.la - English-Hindi dictionary	Andreas Schroeter and Patrick Uecker	2016	N/A		×		×		×	×
English to Hindi dictionary	Dicts.info	2003-2016	N/A			×				×
ALDictionary	Adept Leal Software	2015-2016	N/A	×						×

Table 5.3: Summary table of Hindi–English as well as English–Hindi online dictionaries. Signs in the Quality multicolumn indicate whether a page have IPA transcription for the entries, whether the part-of-speech of a given entry is listed, whether Hindi entries can be looked up with Latin script queries and whether audio content is available for the given word. Signs in the Scrapeability multicolumn indicate whether a word listing is available with hyper links to the entries, and whether an URL-query is possible.

Portal	Ranking		Quality		
	Daily unique visitors	Rank in India	Mean post length (N=10)	Update frequency	Video content
Navbharat Times	3,133,000	10	N/A	76.33±49.96 [14-139]	×
NDTV India	1,664,000	35	1693.8±931.35 [1002-4207]	52.33±45.45 [1-147]	×
OneIndia	1,381,725	107	1727.9±615.48 [773-2869]	59.33±65.39 [1-214]	×
ZeeNews India	858,000	68	1181.5±301.23 [713-1533]	30.11±34.97 [0-115]	×
Dainik Bashkar	696,000	40	1417.2±1096.23 [433-4240]	50.56±85.99 [2-276]	×
Naidunia	515,472	91	1183.4±608.4 [366-2379]	30.89±21.95 [4-68]	×
Dainik Jagran	374,500	101	1305.9±430.93 [713-2077]	90.78±103.64 [25-355]	×
Amar Ujala	167,500	189	984.7±370.89 [500-1628]	64±56.82 [1-181]	×
Patrika	145,000	269	1462.5±434.18 [894-2554]	58.67±60.85 [5.28-164.13]	×
Prabhat Khabar	34,978	1704	2305.9±1618.52 [1036-5571]	147.33±160.52 [9-450]	×
Samachar Jagat	13,600	2000	1334.2±411.44 [765-2037]	97.4±161 [4.3-505.67]	×
Pratakhali	3,774	15528	1245.9±731.85 [337-2368]	N/A	
Loktej	1,749	80970	1001.8±606.49 [370-2044]	N/A	×
Haribhoomi	237	N/A	1551.9±548.1 [670-2505]	N/A	

Table 5.4: Hindi news portals ordered by the number of their daily unique visitors. In addition, the table also shows how individual sites are ranked in India (the country with the most Hindi speakers). As an indicator of ‘quality’, the mean character length of a random sample of posts and mean post lengths are indicated (see Section 5.5.2). The table also highlights if a site hosts video content.

Online Hindi dictionaries

Table 5.3 provides an overview of the available Hindi–English as well as English–Hindi online dictionaries. Dictionaries that are mere wrappers of other engines are not included in the list.

Online Hindi news portals

The most important Hindi news outlets are reviewed in Table 5.4. To give an impression on the quality of the individual sites, post lengths and posting frequencies were measured. Mean post lengths are indicated in characters (not counting whitespaces) and are calculated from the random sample of 10 post bodies (titles are not included in the figures, artifacts, such as code chunks, were removed). Since most sites do not sport a list of their material ordered by release date, post frequencies for individual portals were measured by noting down the release time of the first 10 posts appearing on the main site, starting from page top. The dates were then ordered from oldest to latest, and the difference of individual steps were calculated and converted to minutes. From this set of 9 difference values mean and standard deviation values as well as ranges were calculated. The measurements were carried out between 03:00 AM and 04:00 AM, Indian Standard Time (IST).

5.5.3 Corpora

Monolingual

As of 2016-10-14, the [Hindi Wikipedia](#) consisted of 111,622 public articles. In addition, the [Hindi Wikibooks](#) also contained 313 units. For the purpose of speech recognition development, the ELRA Universal Catalogue contains several entries: [ELRA-U-S 0068](#) (1000 sentences spoken by 100 Hindi native speakers), [ELRA-U-S 0061](#) (recordings of syllables, frequent words, digits, phonetically rich sentences, prosody rich sentences, as well as domain-specific texts and news texts), [ELRA-U-S 0074](#) (LILA speech database, recorded with more than 2000 speakers) and the [BBC corpus](#). The same catalog also features the [ELRA-WC095](#) Tagged Hindi written corpus.

Bilingual

The various sacred texts are ideal candidates for bilingual corpus building; the Bible is made available in .html format by the [Wordproject®](#) (in addition to the Devanagari variant, a Hindi Bible version is also available with [Latin script](#)), the Quran is readily available aligned with various languages in .xml format on the site of the [OPUS corpus project](#), the [New World Translation](#) and the [Book of Mormon](#) is also available in Hindi. In addition, the Universal Declaration of Human Rights is also available in .txt format on the web page of the [Unicode Consortium](#). Further, [Ubuntu](#) and [GNOME](#) localization files are also available aligned with different languages in .xml format on the site of the OPUS corpus project. A parallel corpus of [KDE4](#) localization files consisting of 75,535 items in 92 languages is also available for Hindi.

5.6 Computational tools

5.6.1 Language identification

There are several tools that can identify the language of Hindi texts: [Shuyo](#) (2010) is a Java implementation, the language identification part of the OpenNER project (2015) is another Java implementation

and the Compact Language Detector 2 (2013) is a C++ variant. The langdetect 1.0.7 Python library also supports Hindi recognition (Danilak, 2014).

5.6.2 Tokenizer

The Python libraries entitled Indic NLP Library and HindiTokenizer both support the tokenization of Hindi texts (Kunchukuttan, 2013; Singh, 2015). MorphAdorner 2.0, a Java implementation, is also compatible with Hindi (Burns, 2013). The General Architecture for Text Engineering also encompasses a Hindi tokenizer (Cunningham, 2002).

5.6.3 Stemmer

Besides functioning as a tokenizer, HindiTokenizer also acts as a stemmer (Singh, 2015). Other options include the Python library entitled hindi_stemmer (Ramanathan and D., 2010) and the light stemmer developed by Dolamic (2009).

5.6.4 Spell checker

Besides the open-source HunSpell extension, Aspell also supports Hindi. In addition, Apache OpenOffice and Mozilla Firefox also offers Hindi spell checkers.

5.6.5 Phrase level and higher tools

This section presents a collection of Hindi trained tools and, in the absence of them, articles dealing with higher-level computational tasks:

- **Hindi part-of-speech tagger:** Bali et al. (2010), Bird (2006), Reddy and Sharoff (2011), (Cunningham, 2002)
- **Hindi named entity recognizer:** Al-Rfou (2015)
- **Hindi chunker:** Wani (2016)
- **Hindi morphological parser:** Al-Rfou (2015), Online Morphological Analyzer by IIT Bombay (authors remain undisclosed)
- **Hindi sentence parser:** Reddy (2014), Khan and Khwaja (2013), (Cunningham, 2002), Shallow Parser developed by a team of LTRC, IIIT (authors and creation date are undisclosed)
- **Hindi question answering system (only articles):** Nanda et al. (2016), Sahu et al. (2012), Sinha (2006), Kumar et al. (2005)
- **Hindi speech recognizer:** Josh (2008), Hindi Speech Recognizer tutorial
- **Hindi machine translator:** Chaudhury et al. (2013), Google (2006), Babylon Ltd. (2014), Microsoft Corporation (2016)

5.6.6 End-user support

- **Mac OSX** (as of 2016-10-13): [NO OS-level support](#)
- **Microsoft Windows** (as of 2016-10-13): [Language pack available](#)
- **Ubuntu** (as of 2016-10-13): [Partial translation](#) (76% untranslated)

Bibliography

- S. Ager. Hindi – omniglot, 2016. URL <http://www.omniglot.com/writing/hindi.htm>. [Online; accessed 10-Oct-2016].
- R. Al-Rfou. polyglot, 2015. URL <http://polyglot.readthedocs.org/en/latest/index.html#>.
- S. Avasthī and I. Avasthī. *Chambers English-Hindi Dictionary*. Allied, 1981. ISBN 9788170235491. URL <https://books.google.hu/books?id=d4srdecSTS4C>.
- Babylon Ltd. Babylon, 2014. URL <http://translation.babylon-software.com/english/to-hindi/>.
- Hardev Bahri. *Rajpal Concise English-Hindi Dictionary*. Rajpal & Sons, 2002.
- K. Bali, M. Choudhury, P. Biswas, Jha G. N., Choudhary N. K., and Sharma M. Indian language part-of-speech tagset: Hindi, 2010. URL <https://catalog.ldc.upenn.edu/LDC2010T24>.
- Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.
- K. Bulcke. *English Hindi Dictionary*. Laurier Books Limited, 2005. ISBN 9788121927192. URL <https://books.google.hu/books?id=LueaDwHAvH0C>.
- P. R. Burns. Morphadornor v2.0, 2013. URL <https://devadornor.northwestern.edu/maserver/>.
- Tara Chand. The problem of a common language for india. In *The problem of Hindustani*, pages 13–40. Indian Periodicals Ltd., 1944.
- S Chaudhury, A. Rao, and Sharma D. M. Anusaaraka: English to hindi machine translation system, 2013. URL <https://github.com/sriram-c/anusaaraka>.
- Hamish Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002. URL https://gate.ac.uk/gate/doc/plugins.html#Lang_Hindi.
- M. Danilak. langdetect 1.0.7, 2014. URL <https://pypi.python.org/pypi/langdetect?>
- Ashwini Deo. Dialects in the indo-aryan landscape. In Nerbonne J. Boberg, C. and Watt D., editors, *The Handbook of Dialectology*. Wiley-Blackwell, 2017.
- L. Dolamic. Hindi light stemmer, 2009. URL <http://members.unine.ch/jacques.savoy/clef/HindiStemmerLight.java.txt>.
- M. S. Dryer and M. Haspelmath, editors. *Language Hindi*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL http://wals.info/languoid/lect/wals_code_hin.

- Google. Google translate, 2006. URL <http://translate.google.com/>.
- Government of India. Abstract of speakers' strength of languages and mother tongues. online, 2001. URL http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.aspx.
- Government of India. Decadal variation in population since 1901. online, 2011. URL http://www.censusindia.gov.in/2011census/PCA/A-2_Data_Tables/00%20A%202-India.pdf.
- George Abraham Grierson. *Linguistic survey of India*, volume 9. Office of the superintendent of government printing, India, 1903.
- H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 2.6, 2015. URL <http://glottolog.org>.
- A. Hayden and J. Riesa. Compact language detector 2, 2013. URL <https://code.google.com/p/cld2/wiki/CLD2FullVersion>.
- S. Josh. Hindi automatic speech recognition system (version 2.0), 2008. URL <ftp://ftp.heanet.ie/mirrors/sourceforge/h/hi/hindiasr/Hindiasr/HindiASR-2.0/>.
- Joshua Project. Language - Hindi, 2016. URL <https://goo.gl/BgnFLi>. [Online; accessed 10-October-2016].
- Y. Kachru. *Hindi*. London Oriental and African language library. John Benjamins, 2006. URL <https://books.google.hu/books?id=ooH5VfLTQEQC>.
- B.N. Kapoor. *English-Hindi Dictionary*. Prabhat Prakashan, 2009. ISBN 9788173155604. URL <https://books.google.hu/books?id=4V1npomhuqMC>.
- S.H. Kellogg. *A Grammar of the Hindi Language*. AM. Press Mission Press, 1876.
- M. T. Khan and A. Khwaja. A top down approach to hindi parser without using chunker, 2013. URL <https://github.com/talha31093/Hindi-Parser--NLP->.
- L.M. Khubchandani. *Revisualizing Boundaries: A Plurilingual Ethos*. Language and Development series. SAGE Publications, 1997. ISBN 9780803993532. URL <https://books.google.hu/books?id=GcdhAAAAMAAJ>.
- Praveen Kumar, Shrikant Kashyap, Ankush Mittal, and Sumit Gupta. A hindi question answering system for e-learning documents. In *2005 3rd International Conference on Intelligent Sensing and Information Processing*, pages 80–85. IEEE, 2005.
- S. Kumar and R. Sahāya. *Oxford English-English-Hindi Dictionary*. Oxford, 2014. ISBN 9780198076407. URL <https://books.google.hu/books?id=u1sijgEACAAJ>.
- A. Kunchukuttan. Indic nlp library, 2013. URL http://anoopkunchukuttan.github.io/indic_nlp_library/.
- C.P. Masica. *The Indo-Aryan Languages*. Cambridge Language Surveys. Cambridge University Press, 1993. ISBN 9780521299442. URL <https://books.google.hu/books?id=Itp2twGR6tsC>.
- Microsoft Corporation. Bing translator, 2016. URL <https://www.bing.com/translator>.

- G. Nanda, M. Dua, and K. Singla. A hindi question answering system using machine learning approach.
- In *Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on*, pages 311–314. IEEE, 2016.
- OpeNER. Language identifier, 2015. URL <https://github.com/opener-project/language-identifier>.
- L. M. Paul, G. F. Simons, and D. F. Charles. Ethnologue: Languages of the World, 2015. URL <http://www.ethnologue.com>.
- L. M. Paul, G. F. Simons, and D. F. Charles. Urdu - ethnologue, 2016. URL <http://www.ethnologue.com/language/urd>. [Online; accessed 10-October-2016].
- Amrit Rai. *A house divided: The origin and development of Hindi/Hindavi*. Oxford University Press, 1984.
- A. Ramanathan and Rao D. D. Hindi stemmer, 2010. URL http://research.variancia.com/hindi_stemmer/.
- S. Reddy. Hindi dependency parser, 2014. URL <https://bitbucket.org/sivareddyg/hindi-dependency-parser>.
- S. Reddy and S. Sharoff. Cross language pos taggers (and other tools) for indian languages: An experiment with kannada using telugu resources. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 11–19. Asian Federation of Natural Language Processing, 2011. URL <https://bitbucket.org/sivareddyg/hindi-part-of-speech-tagger>.
- Shriya Sahu, N Vashnik, and Devshri Roy. Prashnottar: A hindi question answering system. *International Journal of Computer Science and Information Technology (IJCSIT)*, 4(2):149–158, 2012.
- ScriptSource. Devanagari (nagari), 2016. URL http://scriptsource.org/cms/scripts/page.php?item_id=script_detail&key=Deva.
- N. Shuyo. Language detection library for java, 2010. URL <https://github.com/shuyo/language-detection>.
- SIL International. Iso 639 code tables, 2015. URL http://www-01.sil.org/iso639-3/codes.asp?order=639_3. [Online; accessed 21-November-2015].
- B. Singh. *Diamond English-English-Hindi Dictionary*. Diamond Pub., 2001. ISBN 9788171824106. URL <https://books.google.hu/books?id=UR9kggorQ74C>.
- Smriti Singh and Vaijayanthi M Sarma. Hindi noun inflection and distributed morphology. In *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*, pages 307–321, 2010.
- Smriti Singh and Vaijayanthi M Sarma. Verbal inflection in hindi: A distributed morphology approach. In *PACLIC*, pages 283–292, 2011.
- T. Singh. Hinditokenizer, 2015. URL <https://github.com/TroJan/hindi-tokenizer/>.

- Udaya Narayana Singh, Biju Rani Pal, Shailendra Kumar Singh, Suchita Singh, and Janardan Mishra. *Longman-Central Institute of Indian Languages English-English-Hindi Dictionary*. Pearson Education India, 2011.
- RMK Sinha. On design of a question-answering interface for hindi in a restricted domain. In *International Conference on Artificial Intelligence, Las Vegas*, pages 319–24, 2006.
- B. Tivārī, A. Kapūra, and V. Gupta. *Comprehensive English-Hindi dictionary*. Kitābaghara Prakāśana, 1998. ISBN 9788170164135. URL https://books.google.hu/books?id=UhNLb3i_FH8C.
- U. Tivārī. *English-Hindi Dictionary*. Hippocrene Dictionaries Series. Hippocrene Books, 1990. ISBN 9780870529788. URL <https://books.google.hu/books?id=lItuQgAACAAJ>.
- UCLA Language Materials Project. Hindi, 2006. URL <http://lmp.ucla.edu/Profile.aspx?LangID=87&menu=004>.
- Unicode, Inc. Devanagari, 2016. URL <http://unicode.org/charts/PDF/U0900.pdf>. [Online; accessed 10-October-2016].
- Raghu Vira. *A comprehensive English-Hindi dictionary of governmental & educational words & phrases*. International Academy of Indian Culture, 1962.
- N. J. Wani. A crf based chunker for hindi, 2016. URL <https://github.com/ltrc/hin-chunker>.
- World Bank. Population 2015, 2015. URL <http://databank.worldbank.org/data/download/POP.pdf>. [Online; accessed 10-October-2016].

Chapter 6

Hungarian

Contents

6.1 Demography and ethnography (László Kálmán)	83
6.2 Main typological and syntactic features (László Kálmán)	87
6.3 Writing system, transcription (László Kálmán)	90
6.4 Previous research on the language (László Kálmán and András Kornai)	90
6.5 Data and sources (András Kornai)	91
6.6 News portals	92
6.7 Computational tools (András Kornai)	92
Bibliography	94

6.1 Demography and ethnography (László Kálmán)

According to 2005 estimates, Hungarian is spoken by about 10 million people in Hungary, plus by another c. 4 million in neighbouring countries and, sporadically, in other countries. This represents 0.21% of the world population (and ranks Hungarian as 71st among the world's languages).

6.1.1 Name variants

The Hungarians' ethnic autonym is *magyar* [məjɔr], which is also the Hungarian name of the language. Its Ethnologue code is ISO 639-3 (*hun*).

6.1.2 Geographic spread

Figure 6.1 shows that the speakers of Hungarian are to be found chiefly within the current borders of Hungary, with a narrow area around the country where Hungarian is spoken sporadically. In addition, there are larger areas in Transylvania where Hungarian is spoken sporadically, and two regions where Hungarian-speaking community is dense. The first one is called the *Székely Land* (Romanian: *Tinutul Secuiesc*), with the city of Târgu Mureş as its center, and covering the Romanian counties of Harghita, Covasna, and parts of Mureş. The second one is much smaller, it is located in the current Romanian county of Bacău, and is populated by the *Csángó* (Romanian: *Ceangău*) ethnic subgroup.



Figure 6.1: Areas in and around Hungary with Hungarian speakers.

Source: [The Hyperglot Blog](#).

6.1.3 Speaker populations

Table 6.1 below summarizes the number of Hungarian native speakers in the world for those countries where they are most numerous. The estimates date from around the turn of the millennium, and originate from [Ethnologue](#).

6.1.4 Dialect situation

Owing to the long history of the unified Hungarian Kingdom, the differences between Hungarian dialects are minor, and they do not affect mutual intelligibility, with the single exception of the dialect of the *Csángó* group, who live in a remote region that never made part of the country. As a consequence of its independent evolution, that dialect contains many archaisms and innovations that makes it hardly intelligible to other Hungarians.

Because of this situation, the classification of Hungarian dialects is notoriously difficult: the isoglosses corresponding to the variables represent small differences, and are intertwined in complex ways. Usually 7 to 10 dialectal areas are listed, but the most accepted view is that 8 areas can be distinguished (as shown in Figure 6.2), mainly on the basis of phonetic variables.

Western, Western Transdanubian (*nyugati*)

The most noticeable feature of this area is that the long vowels pronounced [o:], [ø:], [e:] in the standard dialects correspond to opening diphthongs [^o(:)])], [^ø(:)])], [^e(:)])] (in the Northern sub-area, the last one is also pronounced [i:])). In addition, a (lexical) distinction is made between [ɛ] and

COUNTRY	SPEAKERS (THOUSANDS)	EGIDS LEVEL	RATIO IN POPULATION
Hungary	10300	1	100
Romania	1450	2	7.3
Slovakia	521	2	1
Serbia	287	2	0.04
Austria	259	5	0.03
Ukraine	157	5	0.004
Israel	70	5	0.01
Slovenia	9.24	2	0.004

Table 6.1: Number of Hungarian native speakers,
with their ratio in the country's population.

[e] (collapsed into [ɛ] in standard Hungarian), with the latter often pronounced [œ]. There is a general tendency of shortening the long high vowels ([i:], [u:], [y:]). With the exception of compound boundaries, standard Hungarian [j] is pronounced [ʃ] after consonants; orthographic *ly* (pronounced [j] in standard Hungarian) is often pronounced [l]; the consonant [l] is often dropped when in the syllable coda (immediately following the nucleus), with the compensatory lengthening of the nucleus vowel; the word-final [n] of standard Hungarian is often pronounced [n]. Finally, the consonant [v] not only undergoes, but also triggers the regressive voicing assimilation of obstruents, unlike in standard Hungarian (e.g., *hatvan* is pronounced [hədvən], as opposed to standard Hungarian [hətvən]). The number of speakers is about 700 thousand people.

Northern Transdanubian, Transdanubian (*dunántúli*)

The most typical feature of this dialect is the shortening of the long high vowels (as was explained in the Western dialect), accompanied by the compensatory lengthening of the subsequent consonant when this happens in a non-word-final syllable. The (lexical) distinction between [ɛ] and [e] is also present. About 1.8 million people speak this dialect.

Southern Transdanubian, Southern (*déli*)

There is no single typical phonological characteristics of this dialect, but all the phenomena of the Western and the Northern Transdanubian dialects can be observed sporadically in various subregions. Standard Hungarian (spoken in Budapest and surroundings) also belongs here, and is characterized by the lack of those dialectal features mentioned in connection with the Western and Northern Transdanubian dialects. Around 4 million speakers.

Palóc

This is a relatively marked dialect area with relatively sharp geographical boundaries. Hungarians distinctly recognize Palóc speakers, but they do not have difficulties understanding them. Its most prominent property is a general tendency of illabialization: standard [y] and [œ] are often pronounced [i] and [e], respectively, and standard [ɔ] is pronounced [a] or even [a]. On the other hand, the vowel [a:] of standard Hungarian is typically pronounced [a:] or even [ɔ:]. There is also a tendency of palatalizing

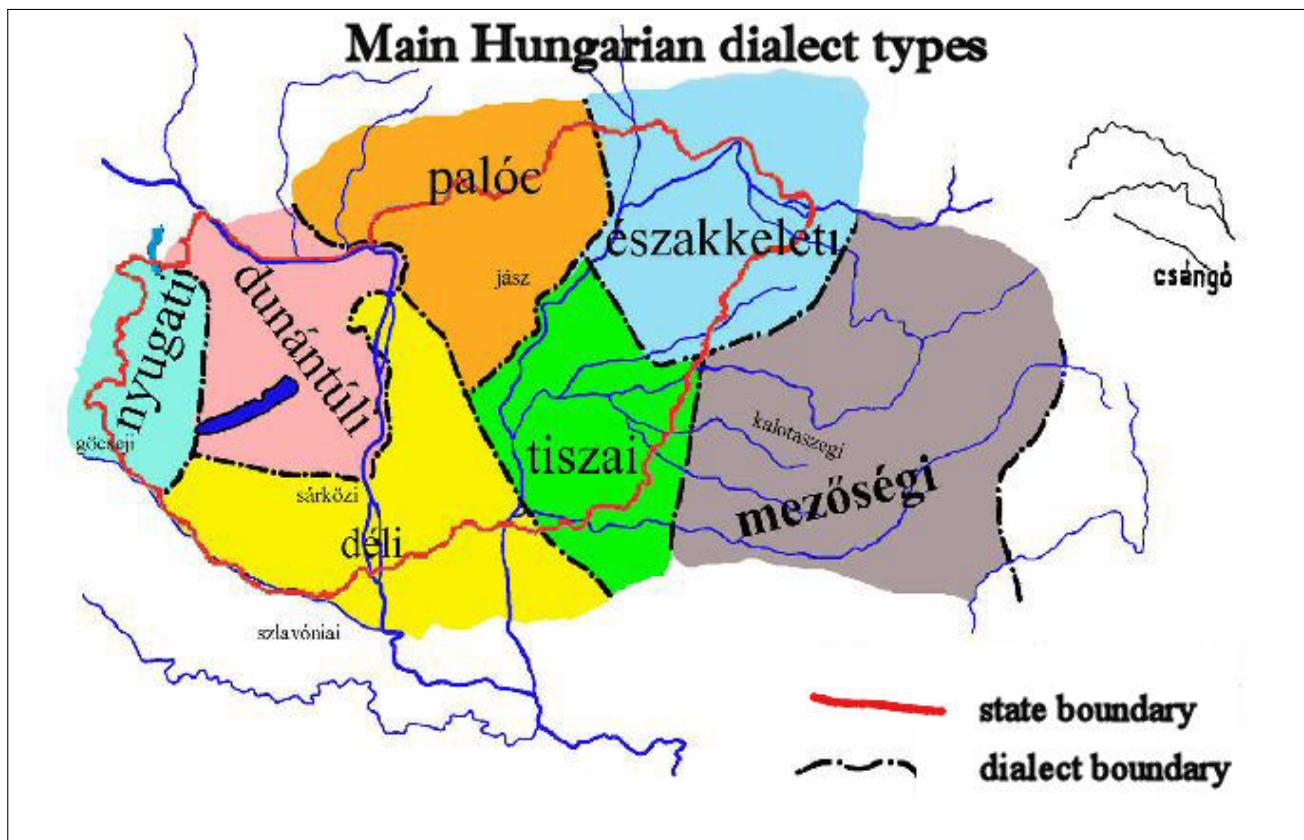


Figure 6.2: Hungarian dialects.

Source: Research Institute for Linguistics, HAS, Department of Phonetics.

the consonants [t], [d], [n] and [l] in front of the vowels [i] and [y]. Speakers: about 800 thousand people.

Danube–Tisza, Alföld, Southern Great Plains, Tisza–Körös (*tiszai*)

The distinguishing feature of this dialect is a marked (lexical) difference between [ɛ] and [e]. The latter is pronounced [œ] in a subregion of the area. In addition, in the regions bordering the Northeastern area, the long vowels [o:], [ø:] and [e:] of standard Hungarian are pronounced with closing diphthongs [o(:)u], [ø(:)y] and [e(:)i], respectively. In the same area, short vowels are pronounced as semi-long when immediately followed by a coda consonant [l], [r] or [j]. About 1.3 million speakers.

Northeastern (*északkeleti*)

This area is characterized by the diphthongization of long [o:], [ø:] and [e:], as seen in the Alföld dialect. Short vowels are even more markedly lengthened in front of an [l], [r] and [j] in the coda than in the Alföld dialect. The consonants [t] and [d] of standard Hungarian are pronounced with apical [t̪] and [d̪]. The long [e:] of standard Hungarian is realized as [e:] or [i:] (the distinction is lexical). About 2 million speakers use this dialect.

Transylvanian (*mezősségi*)

This is the most debated dialect area, because many authors further divide it into smaller areas, such as the *kalotaszegi*, the *székely* etc. subdialects. The most general characteristics of the whole region

is a tendency of pronouncing more open short vowels than in standard Hungarian. Thus standard [ɛ] may be pronounced as [æ], [o] as [ɔ], and so on. About 1.5 million speakers.

Csángó

The Csángó dialect has many lexical and morphological peculiarities as compared to standard Hungarian, and those are the main obstacles of its mutual intelligibility with other dialects. The phonological differences are not very marked. For example, in Csángó there is a three-way distinction between [s], [ʃ] and [ʃ̪], whereas in standard Hungarian only [s] and [ʃ] exist; it is lexically determined when [s] is pronounced when standard Hungarian has [s] or [ʃ]. Similarly, when standard Hungarian [ʃ] and [d] are pronounced as [dʒ] in Csángó is lexically determined. The number of speakers is estimated to 60-70 thousand people.

6.2 Main typological and syntactic features (László Kálmán)

6.2.1 Linguistic typology

Morphology

Hungarian is an *affixing* language, i.e., its words usually contain a stem plus affixes. The affixes are typically suffixes (with the exception of one or two prefixes). A stem (especially a nominal stem) may contain more than one root morpheme, i.e., compounding is very frequent. In the case of verb stems, compounding is limited to a certain type of incorporation of adverbials and nominals, called *verbal prefixes* or *preverbs*.

Morpho-phonology

Hungarian is a predominantly *agglutinative* language, i.e., the boundaries of the elements of a word form are relatively transparent, but there are many alternations in both stems and affixes that are conditioned by affixation, which is characteristic for fusional languages. One of the typical affix alternations is *vowel harmony* in terms of the front/back character of the last vowel of the stem. (In a subset of the affixes, and only in front rather than back suffixation, the round/illabial character of the last stem vowel also matters.) There are vowels *neutral* with respect to vowel harmony (namely, [i], [i:], [e:] and, to some extent, [ɛ]), which means that suffixes containing these vowels normally do not vary in terms of vowel harmony. Should a stem's last vowel be a neutral vowel, that vowel is often “invisible” to vowel harmony (or the behaviour of the stem is lexically determined). Inter-speaker and even intra-speaker variation is very frequent in this process.

Phonology

The inventory of standard Hungarian speech sounds includes 24 consonants (plus their long counterparts) and 14 vowels.

The consonants include nasals ([m], [n], [ŋ]), stops ([p], [b], [t], [d], [c], [ɟ], [k], [g]), affricates ([ts], [tʃ], [tʂ]), fricatives ([f], [v], [s], [z], [ʃ], [ʒ]), the approximants [h] and [j], the lateral [l] and the trill [r]. More or less regular alternations can be observed between voiced and voiceless obstruents, the place of articulation of alveolars (e.g., in front of palatals) and nasals.

There are seven short vowels ([ɔ], [ɛ], [i], [o], [ø], [u], [y]), which have long counterparts (although sometimes of a different quality from the short ones): [a:], [e:], [i:], [o:], [ø:], [u:], [y:]. More or less

regular alternations exist between short and long counterparts, and between the front/back pairs [ɔ]/[ɛ], [o]/[ø], [u]/[y] (in the short series) and [a:/[e:], [o:/[ø:], [u:/[y:] (in the long series). There is also an alternation between [ø] and [ɛ] in some suffixes depending on the rounded vs. illabial character of the last front vowel of the stem.

Syntax

Hungarian is traditionally considered a language with basic SVO word order (i.e., the subject precedes the finite verb, which is followed by the remaining complements). But this is a rather poor description of the actual situation. Hungarian is a so-called *topic prominent* language, which means that a more correct description of the basic word order would be TFVO, where *T* stands for *topic* (“logical subject”, about which the predication is made), irrespective of its grammatical case, and *F* stands for *focus*, an optional constituent which, when present, expresses contrast, also independently of grammatical case. It is true that, statistically, grammatical subjects very often play the role of topic (or focus), but that does not make its syntax similar to that of English or Chinese, where SVO is indeed the basic word order schema.

Hungarian uses head-final nominal phrases; subordinate clauses normally follow main clauses; the role of complements is expressed using a large number (depending on the analysis, 18–22) of case endings and/or postpositions. The difference between case endings and postpositions is mainly of a morpho-phonological character: they both occupy the last position of a nominal phrase.

In Hungarian, agreement is limited to possessor/possessed, finite verb/grammatical subject and finite verb/grammatical object structures. Hungarian is a so-called *pro-drop* language: no explicit pronominal expression is required when a grammatical subject, object or a possessor noun phrase is absent. Whenever they are missing, the agreement morphemes can be seen as pronominal elements. On the other hand, the actual morphemes used may be slightly different when the possessor or the grammatical subject are explicitly expressed.

6.2.2 Predication

In Hungarian, non-verbal predication consists of a subject and a predicate nominal (adjective or noun) in the (indicative) present tense. (Pronominal subjects need not be explicit, given that Hungarian is a *pro-drop* language.) The most neutral word order is when the subject precedes the predicate, but the inverse order is also perfectly grammatical, especially if the subject is known from the context (and does not carry main stress).

- (1) a. ('Ez a 'fickó) 'hülye.
this chap stupid
'This chap is stupid.'
- b. 'Hülye (ez a fickó).
stupid this chap
'He is stupid, this chap.'

In other tenses and moods, the presence of a form of the verb ‘be’ (or sometimes another copula, like the one meaning ‘remain’) is obligatory. The copula is unstressed, and it follows the predicate nominal in the unmarked case. The example (2b) below illustrates a marked case, when the subject occupies the contrastive focus position, therefore it has the only main stress, and it is immediately followed by the finite verb (in this case, the copula).

- (2) a. ('Ez a 'fickó) 'hülye volt/lesz/lenne/...
 this chap stupid was/will be/would be/...
 'This chap was/will be/would be stupid.'
- b. 'Ez a fickó volt/lesz/lenne/... hülye
 this chap was/will be/would be/... stupid
 'It's THIS chap who was/will be/would be stupid.'

6.2.3 Possession

Hungarian possessive constructions are characterized by a *possessive suffix* attached to the last constituent of the *possessed* noun phrase. That last constituent is usually the nominal head. The possessive affix can play a pronominal role, i.e., the possessor need not be explicit. (If, however, the possessor is missing, the noun phrase expressing the possessed entity must be introduced by a definite article.) The possessor noun phrase can stand in the *nominative* (i.e., without any case ending), in which case it has to immediately precede the possessed noun phrase, and the possessed noun phrase may not be introduced by a definite article. Alternatively, the possessive noun phrase can bear *dative* case, in which case the two noun phrases need not be adjacent, i.e., they are more independent of each other, and the possessed noun phrase is also a complete one, with an article. The possessive suffix normally agrees with the possessor in person and number. A notable exception is that, when the possessor is 3PL, then the possessive suffix must take the 3SG form if the possessor is explicitly expressed and immediately precedes the possessed noun phrase. Finally, should the possessor be expressed by a personal pronoun, that pronoun has to be preceded by a definite article.

- (3) a. *Jóska/az autója*
 Joe/the car-Poss.3SG
 'Joe's/his/her car'
- b. *a fiúk autója*
 the boys car-Poss.3SG
 'the boys' car'
- c. *Jóskának/a fiúknak az autója*
 Joe-DAT/the boys-DAT the car-Poss.3SG
 'Joe's/the boys' car'
- d. *az autójuk*
 the car-Poss.3PL
 'their car'
- e. *az (én) autóm*
 the I car-Poss.1SG
 'my car'

6.2.4 Imperative

Hungarian imperatives are formed using the (second person) forms of the *imperative-conjunctive* mood of verbs. The finite verb stands in the beginning of the imperative sentence (unless a *topicalized* constituent is also present, which must come first). In the case of complex verbs (with a *preverb*

or another incorporated element), the incorporated constituent must *follow* the verb stem (unlike in unmarked affirmative sentences).

- (4) a. *Hozd/hozzad* (fel) *a fát!*
 bring-IMPERCONJ.2SG.DEFOBJ PREVERB.up the wood
 ‘Bring (up) the wood!’
- b. *Hozz/hozzá* (fel) *fát!*
 bring-IMPERCONJ.2SG.INDEFOBJ PREVERB.up wood
 ‘Bring (up) some wood!’

6.2.5 Interrogative

Hungarian *wh*-interrogatives are formed using *wh*-words, which must stand in the so-called *focus position*, i.e., immediately precede the (unstressed) finite verb stem (or nominal predicate). The *wh*-word normally starts the entire interrogative sentence, but can be preceded by *topic* constituents. Should the sentence contain a genuine contrastive focus, that constituent is relegated to the post-verbal part (since the focus position is occupied by the *wh*-word), with one exception: the *wh*-word *miért* ‘why’ can be (immediately) followed by a contrastive focus constituent.

- (5) a. ‘*Mit hoztál?*
 what-ACC bring-PAST-2SG-INDEFOBJ
 ‘What have you brought?’
- b. (‘)*Miért Jánost hoztad?*
 why John-ACC bring-PAST-2SG-DEFOBJ
 ‘Why is it John that you brought with you?’

6.3 Writing system, transcription (László Kálmán)

Hungarian uses the Latin alphabet, with some diacritics (on vowels), namely, acute accent for the long vowels á, é, í, ó, ú, Umlaut for the two front rounded vowels ö, ü, and, quite exceptionally, a long Umlaut for their long counterpartső, ū. There are consonant symbols consisting of two letters, namely, *cs* [tʃ], *gy* [j], *ly* [j], *ny* [n], *sz* [s], *ty* [c], *zs* [ʒ]. There is one consonant symbol consisting of three letters: *dzs* [dʒ]. The long counterparts of consonants are represented by doubling the consonant letters; for consonant symbols consisting of two letters, only the first one gets doubled (e.g., *ccs*, *ggy*). Short [dʒ] and long [dʒ:] are not distinguished (both are written with *dzs*).

6.4 Previous research on the language

The most recent and most extensive reference grammar of Hungarian is a series of handbooks written in Hungarian: Kiefer (1992, 1994, 2000, 2008). More compact recent Hungarian descriptive grammars in English are Törkenczy (2002) and in Kenesei et al. (2002).

6.4.1 Research and control bodies

The Hungarian Academy of Sciences (HAS) was originally founded with the aim of “evolving” the language. The Research Institute for Linguistics of HAS has departments for theoretical linguistics,

language technology, lexicography, historical (finno-ugoric) linguistics, phonetics, psycho-, neuro-, and sociolinguistics.

The Hungarian Speech and Language Technology Platform, coordinated by the Department of Language Technology of HAS, is an umbrella organization for industry ([memoQ](#) (translation), [AITIA](#) (speech technology), [Morphologic](#), and the [Applied Logic Laboratory](#)) and university institutes (the [Human Language Technology \(HLT\)](#) research group headed by the author of this section, the Department of Telecommunications and Media Informatics of the Budapest University of Technology and Economics ([BME TMIT](#)) and the [Human Language Technology Group of the University of Szeged](#)).

6.5 Data and sources (András Kornai)

6.5.1 Basic vocabulary

An excellent lexicon and sublexicons can be found in [the “Hungarian lexical database and morphological grammar” called `morphdb.hu`](#). The reader may consult the [Hungarian WordNet](#) with its 42288 synsets available through Python 3 and C++ APIs and more command-line tools.

6.5.2 Dictionaries

The two most important monolingual dictionaries are [Bárczi and Országh \(1959–62\)](#) and [Juhász et al. \(2003\)](#) (75,000 entries). The former is available in printed version only, the latter is also [available online](#) after registration and payment. The classical Hungarian–English and English–Hungarian dictionaries (c. 120,000 entries) are [Országh \(1994a,b\)](#), also [available online](#) after registration and payment. Free online dictionaries can be found [on the web page of SZTAKI \(the Institute for Computer Science and Control of the Hungarian Academy of Sciences\)](#).

6.5.3 Corpora

Starting with the most open sources, the largest Hungarian corpus available in a downloadable format is [the Hungarian Webcorpus](#) (over 1.48 billion words unfiltered, 589 million words fully filtered). Another gigaword corpus is [the Hungarian National Corpus](#), which can only be queried online. [The Szeged Treebank](#) is a corpus annotated manually (in the spirit of Chomskyan syntax) containing 1.2 million words in 82,000 sentences. Finally, [the Hunglish Corpus](#) is a free sentence-aligned Hungarian–English parallel corpus of about 54.2 million words in 2.07 million sentences. [HuTenTen](#) is a proprietary corpus with 3.2 billion words. The publication of the Pázmány Gigaword Corpus ([Endrédy and Novák, 2013](#)) is in progress.

The greatest media corpus for Hungarian is the [National Audiovisual Archive](#). Besides, there is the very recent [emOSA Hungarian Open Speech Archive](#) designed to support the archival needs of research that relies on spoken materials (linguistics, anthropology, ethnography, etc.) with special emphasis on modern speech technologies, and in the long term, open-source automatic speech recognition. The archive focuses on Hungarian, the minority languages spoken in Hungary (eg. different Romani dialects), and neighboring and related languages (majority or minority languages and dialects). Spontaneous speech recorded in natural conditions, noisy background or with inferior equipment is preferred.

6.6 News portals

The 20 most visited sites in Hungary, according to Alexa, include the news portals [index.hu](#), [origo.hu](#), [24.hu](#), [444.hu](#), and [hvg.hu](#), blog services ([blog.hu](#) and [blogspot.hu](#)) and the weather site [időkép](#) which is much more popular than the more accurate National Meteorology Service [met.hu](#). A FAQ site [Gyakorikérdések](#)<http://www.gyakorikerdesek.hu/> also appears in the top.

6.7 Computational tools (András Kornai)

There have been three language processing pipe-lines for Hungarian: the latest [e-magyár](#) (modul names starting with em) obsoletes, in most regards, [magyarlánc](#) developed at the Szeged University and the hunstar tool-chain developed at [MOKK Centre for Media Research and Education](#) at the Budapest University of Technology and Economics and later at [HLT](#). [e-magyár](#) consists of the tokenizer [emToken](#) (but see the older [huntoken](#) as well), the morphological analyzer [emMorph](#), the lemmatizer [emLem](#), the POS tagger [emTag](#), two syntactic analyzers (dependency, [emDep](#), and constituent [emCons](#)), the chunker (shallow parser) [emChunk](#), and the named entity recognizer [emNer](#). [hunspell](#), reported in most languages of this report, remains the spell checker of LibreOffice, OpenOffice.org, Mozilla Firefox 3 & Thunderbird, Google Chrome, and it is also used by proprietary software packages, like Mac OS X, InDesign, memoQ, Opera and SDL Trados.

6.7.1 Language identification

Since Hungarian is hard to confuse with other languages, standard LID tools such as [TextCat](#) are easy to train for the language. Most (semi)commercial LID tools such as the ones offered by [Xerox](#), [Basis Technology](#), and [Google](#) support Hungarian out of the box.

6.7.2 Tokenizer

[emToken](#), a.k.a. [quntoken](#) is a quex-based version of the freestanding [huntoken flexer](#). The tokenizer in [Magyarlánc](#) is a Hungarian-adapted version of [morphadornor](#).

6.7.3 Stemmer

Stemming and morphological analysis are generally performed by the same tool for Hungarian (the equivalent of the Porter stemmer, which creates linguistically unmotivated stems, does not exist for Hungarian). [emMorph](#) is the open source result of the homologization of the Magyarlánc stemmer (built on [morphdb.hu](#)) and [hummorph](#). The morphological annotation is described (in Hungarian) [on the home page of Péter Rebrus](#).

6.7.4 Spell checker

The [hunspell](#) spell-checker is standard, and [aspell](#) also processes Hungarian.

6.7.5 Phrase level and higher tools

As we already mentioned in the introduction, the latest NLP pipe-line for Hungarian is [e-magyár](#) that covers the most recent tools and in most regards obsoletes [magyarlánc](#) and hunstar.

The most popular POS tagger for Hungarian is [purepos](#), a.k.a. [emTag](#). [Magyarlánc](#) has Hungarian-trained versions of the [Stanford Log-linear Part-of-speech tagger](#), the [Bohnet dependency parser](#) (the Hungarian modell is now called [emDep](#)), and the [Berkeley constituency parser](#) ([emCons](#)). [Hun*](#) also includes a [POS tagger](#).

NER taggers, NP-chunkers, and, more generally, sequential taggers (using Maximum Entropy Learning and Hidden Markov Models) include [HunTag3](#) (a.k.a. emTag and emNer) and [HunTag](#).

Speech technology for Hungarian has been developed at BME TMIT mentioned in [subsection 6.4.1](#). [Ádám Varga et al. \(2015\)](#) achieve remarkable results in real-time, resource-limited close captioning of live broadcast news (BN) and conversation speech. The lowest word error rate (17%) for Hungarian BN was shown by [Tóth and Grósz \(2013\)](#). Recall the recent initiative [emOSA](#) mentioned in [subsection 6.5.3](#) as well.

Besides the companies mentioned in [subsection 6.4.1](#), Hungarian machine translation, especially an end-to-end neural architecture ([Cho et al., 2014](#)) is intensively studied at the Directorate-General for Translation of the European Commission.

6.7.6 End-user support

- MacOS fully supports Hungarian since v10.7 (Lion), and MacBooks can be ordered with Hungarian keycaps from Apple.
- Microsoft Windows also has a Hungarian language pack for Win7 and later, and major hardware vendors supply Hungarian keycaps.
- For Linux, most distributions like [Ubuntu](#) have a Hungarian language pack available

Bibliography

Géza Bárczi and László Országh, editors. *A magyar nyelv értelmező szótára [Monolingual dictionary of the Hungarian language]*. Akadémiai Kiadó, Budapest, 1959–62.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014. URL <http://www.aclweb.org/anthology/D14-1179>.

I. Endrédy and A. Novák. More effective boilerplate removal – the goldminer algorithm. *POLIBITS*, 48:79–83, 2013.

József Juhász, István Szőke, Gábor O. Nagy, and Miklós Kovalovszky, editors. *Magyar értelmező kéziszótár [Abridged monolingual dictionary of Hungarian]*. Akadémiai Kiadó, Budapest, 2 edition, 2003.

István Kenesei, Robert M. Vago, and Anna Fenyvesi. *Hungarian*. Routledge, London/New York, 2002.

Ferenc Kiefer, editor. *Mondattan [Syntax]*, volume 1 of *Strukturális magyar nyelvtan [Structural Grammar of Hungarian]*. Akadémiai Kiadó, Budapest, 1992.

Ferenc Kiefer, editor. *Fonológia [Phonology]*, volume 2 of *Strukturális magyar nyelvtan [Structural Grammar of Hungarian]*. Akadémiai Kiadó, Budapest, 1994.

Ferenc Kiefer, editor. *Morfológia [Morphology]*, volume 3 of *Strukturális magyar nyelvtan [Structural Grammar of Hungarian]*. Akadémiai Kiadó, Budapest, 2000.

Ferenc Kiefer, editor. *A szótár szerkezete [Structure of the lexicon]*, volume 4 of *Strukturális magyar nyelvtan [Structural Grammar of Hungarian]*. Akadémiai Kiadó, Budapest, 2008.

László Országh. *Angol–magyar nagyszótár [English–Hungarian dictionary]*. Akadémiai Kiadó, 1994a.

László Országh. *Magyar–angol nagyszótár [Hungarian–English dictionary]*. Akadémiai Kiadó, Budapest, 12 edition, 1994b.

L. Tóth and T. Grósz. A comparison of deep neural network training methods for large vocabulary speech recognition. In *Text, Speech, and Dialogue*, volume 8082, pages 36–43, 2013.

Miklós Törkenczy. *Practical Hungarian Grammar*. Corvina, Budapest, 2002.

Ádám Varga, Balázs Tarján, Zolltán Tobler, György Szaszák, Tibor Fegyó, Csaba Bordás, and Péter Mihajlik. Automatic close captioning for live hungarian television broadcast speech: A fast and

resource-efficient approach. volume 9319 of *Lecture Notes in Computer Science*, pages 105–112, 2015. doi: 10.1007/978-3-319-23132-7_13. URL <http://citeweb.info/20152609163>.

Chapter 7

Indonesian (Dávid András Imre)

Contents

7.1 Demography and ethnography	97
7.2 Main typological and syntactic features	98
7.3 Writing system, transcription	103
7.4 Previous research on the language	103
7.5 Data and sources	103
7.6 Computational tools	105
Bibliography	109

Introduction

The present chapter aims at providing an introduction of Indonesian, mainly spoken in Indonesia. After a brief review of the demo- and ethnographic situation of the language and its speaking community, the typology, the writing system and the most notable researches about the grammar of the language will be outlined. The second part of this paper provides an account on the various available offline and online data sources of Indonesian, and an outlook at the available tools for the processing of the language.

7.1 Demography and ethnography

7.1.1 Name variants

Indonesian language has only a few name variants and there is no real difference between endonyms and exonyms. The most common name of the language is simply *Indonesian*, which is the English exonym. The Indonesian equivalent of the English name is *Bahasa Indonesia*, which means “the language of Indonesia”, sometimes abbreviated to Bahasa, but one must use this carefully as this can denote the Malay language as well (abbreviated from Bahasa Malaysia). ISO codes for the Indonesian language are the following: *id* ISO 639-1, *ind* ISO 639-2, and ISO 639-3.

7.1.2 Geographic spread and Speaker populations

Indonesian language is widespread. There are around 40 million people who speak Indonesian as a first language, and around 170 million who speak Indonesian as a foreign language. Being the

official language of Indonesia, most Indonesian speakers can be found there, with Java being the most populous, giving home to half of the Indonesian population out of the ca. 260 million total. Indonesian people live in nearby regions in Oceania, Asia and other Pacific countries as well, such as Malaysia, where the second largest Indonesian population can be found. Taiwan, Hong Kong and South Korea are the Asian countries, where larger Indonesian speaking groups live. There is only one European country with significant Indonesian population: the Netherlands. In the USA, only a tiny fraction of Indonesian people can be found. However, there is also a large group of Bahasa people in Saudi Arabia.

The eighteen Indonesian populations with more than 5000 people are listed in the table below excluding Indonesia. This information along with the respective sources can be found on the relevant [Wikipedia article](#) about Indonesian people.

Country	Number of speakers	Country	Number of speakers
Malaysia	2,500,000	Australia	86,196
Netherlands	1,800,000	South Korea	50,000
Saudi Arabia	1,500,000	Philippines	43,871
Singapore	200,000	Qatar	39,000
Taiwan	161,000	Japan	30,567
Hong Kong	102,000	Canada	14,300
United States	101,270	United Kingdom	9,624
United Arab Emirates	100,000	New Caledonia	7,000
Suriname	90,000	Macau	6,269

Table 7.1: Number of Indonesian speakers per country

7.1.3 Dialect situation

As Indonesian is a standardized dialect of the Malay language, it has no dialects according to [Ethnologue](#). But the [entry on Indonesian language](#) on AWL (About World Languages) distinguishes between several dialects that differ mostly in their phonemic inventory and to some degree in their vocabulary. More on this issue can be found in [Conners and Klok \(2016\)](#)

7.2 Main typological and syntactic features

7.2.1 Linguistic typology

The following set of features are those, that describe Indonesian language the best, including differentiators, according to the [entry on Indonesian language](#) on Ethnologue.

There are two social registers, a formal and an informal one. On sentence level, the main characteristics of Indonesian language is word order: following the subject-verb-object order is a differentiator amongst other Austroasiatic languages. Indonesian uses prepositions and heavily relies on affixes: prefixes, suffixes, circumfixes and infixes alike. Sentence initials are usually nouns, except in the presence of quantification.

There are three noun classifiers: strictly singular nouns, proper nouns and countable nouns, but there are no articles. Indonesian does not use grammatical genders except for very few cases where there is such distinction, but most probably those are either the creation of modern times (one example

being the noun “flight attendant”), or absorbed from other languages like Sanskrit. Pronouns are not a distinct part-of-speech, but a subset of nouns. Verbs are not inflected for person or number, and tense is not marked either, but there is a complex system to indicate some aspects, like voice or mood, or completeness of action, where there are three cases: completed, started but not completed, completed.

There are 19 consonants, 6 vowels and 3 diphthongs. Indonesian language is not tonal, which is also a differentiator.

Further information with several references can be found on [the entry on Indonesian](#) on the UCLA Language Materials Project site.

7.2.2 Predication

The main features of predicate phrases are the same as in English, noun phrases are followed by verb phrases. However there are minor differences between the word order in constituents. Noun phrases for example follow a different word order: in Indonesian, the modifier follows the noun, not the other way around. The English expression ‘my **book**’ translates to ‘**buku** saya’, or ‘nice **house**’ translates to ‘**rumah** bagus’. This is a striking difference and applies to any modifier that a noun can have. Speaking more generally about syntactic structure, there are more interesting differences as well: in Indonesian, not only a noun phrase and a verb phrase can constitute a sentence, but a noun phrase together with an adverbial phrase, adjective phrase or with another noun phrase can be a sentence as well: ‘Dompetnya di atas meja’ means ‘his wallet is on the table’, ‘Adam sakit semalam’ means ‘Adam was sick last night’, ‘Orang yang di sana tadi malam Peter’ means ‘The man from last night is Peter’. In English, the adverbial phrase, adjective phrase and the second noun phrase are transformed into a verb phrase by being the object of the verb ‘to be’. More on predicate phrases can be found in Sneddon (1996).

7.2.3 Possession

Possession is expressed through possessive nouns or pronouns. Possessors follow the head word, and in the resulting expression, the possessive noun still can be the head of the noun phrase (however this is not the case when a noun is placed after another to modify its meaning), but because of the same grammatical rules, out of context ambiguity can happen.

- (27) *nama toko itu*

‘the name of that shop’

The head of a possessive noun phrase can also take a possessor:

- (28) *mobil teman Marni*

‘Marni’s friend’s car’

There are no possessive pronouns in Indonesian, that is, a personal pronoun together with a noun phrase constitutes a possessive phrase.

- (29) *buku saya*

‘my book’

- (30) *buku anda*

‘your book’

It is also possible to use inflections instead of pronouns. This is used in informal contexts. The two examples above are equivalent to:

- (31) *bukuku*

‘my book’

- (32) *bukumu*

‘your book’

More on possessive constructions can be found in Sneddon (1996).

7.2.4 Imperative

Imperative sentences usually follow the same SVO word order as in English. However, some words undergo specific transformations when they are in interrogative context. The most transparent case is that of transitive verbs. In Indonesian, transitive verbs can have prefixes and suffixes which guarantee the SVO order, expressing general agents and objects. Some examples:

- (33) *Datang ke sini!*

‘Come here!’

(intransitive verbs remain the same)

- (34) *Lihat foto ini!*

‘Look at this photo!’

(transitive verbs drop *meN-* prefix if they refer to a specific object)

- (35) *Membaca sekarang!*

‘Read now!’

(but they retain the prefix in pseudo-intransitive cases)

- (36) *Menabung uang untuk hari tuamu!*

‘Save money for your old age!’

(and they also retain the prefix in general cases)

Negative imperatives are formed with the help of the word *jangan*, which means “do not” in English. In this case, the *meN-* prefix of transitive verbs is fully optional.

- (37) ***Jangan*** merokok di sini!

‘Don’t smoke here!’

There are several words that can be used to soften imperatives, they appear at the beginning of sentences. *Silakan* indicates, that someone should do something for his/her own benefit, while *tolong* indicates that someone should do something for the speakers benefit.

- (38) ***Silakan*** masuk.

‘Please come in.’

- (39) ***Tolong*** pergi ke toko depan.

‘Please go to the shop across the road.’

Requests are formed with the verbs *(me)minta* and *(me)mohon*.

- (40) ***Minta*** air.

‘Can I please have some water?’

- (41) ***Mohon*** undangan dibawa.

‘Please bring your invitation.’

More on imperative mood can be found in Sneddon (1996).

7.2.5 Interrogative

Simple yes-or-no questions usually has no different word order than the statement corresponding to that question. The interrogative mood is indicated solely by tone.

- (42) *Sri* sudah pulang.

‘Sri has gone home.’

- (43) *Sri* sudah pulang?

‘Has Sri gone home?’

However, there are several words that indicate that the sentence they occur in are questions. For example questions regarding one’s ability or possibility to do something start with the word *boleh*. It is very similar to the English word “can” or “may” in questions.

- (44) *Boleh* saya masuk?

‘May I come in?’

The word *apa* can be placed at the beginning of a sentence to make it overtly interrogative. In this case, the word has no other meaning than indicating interrogation. It is mostly used in formal

written context. Instead of *apa*, the postfix *-kah* can also be attached to a word to indicate a question. In this case the inflected word will begin the sentence. To sum up, the same question can be formed in three different ways:

- (45) *Dia sudah makan?*

‘Has she eaten?’

- (46) ***Apa dia sudah makan?***

‘Has she eaten?’

- (47) ***Sudahkah dia makan?***

‘Has she eaten?’

Information questions are formed with the use of question words (the English wh- words). The list of Indonesian wh- words is the following:

apa - what

siapa - who

(di/ke/dari)mana - where (at/to/from)

(yang) mana - which

kapan, bila - when

apabila, bila(mana) - when

bagaimana - how

berapa - how many, how much

keberapa - which number

mengapa, kenapa - why

untuk apa - why, what for

Information questions usually follow the same word order as the corresponding statements, but in cases where the question word is a predicate of a non-verbal clause, they usually precede the subject.

- (48) *Anda membaca apa?*

‘What are you reading?’

(the question word *apa* is in the object position)

- (49) ***Bagaimana jalan itu?***

‘What is that road like?’

(the question word *bagaimana* precedes the subject; *jalan* means “road”)

More on interrogative mood and questions can be found in [Sneddon \(1996\)](#).

7.3 Writing system, transcription

Throughout its history, Indonesian have witnessed many writing systems (e.g., [the Javanese script](#), and [the Korean Hangul script](#)). As of now, eversince the Dutch have introduced Latin script for writing, these archaic forms mostly disappeared.

The first Latin letter script in use was the [Van Ophuijsen](#) or *Ejaan Lama* Spelling System (i.e., Old Script). During the time of the independence movements, the standardization process of the Indonesian language started, which resulted in the language now known as Indonesian language, or Bahasa Indonesia. One output of this movement was a new writing system, the [Republikan Spelling System](#) (*Ejaan Republik* or RSS), also known as Soewandi Spelling after the minister of education of that time. It was more of a symbolical step than a practical one, since there were only few differences. For instance, in the previous system the vowel /u/ was written with <oe>, and the glottal stop, which was denoted by the diacritic <'> is now marked by the letter <k>.

RSS was replaced in 1972 by the [Enhanced Indonesian Spelling System](#) also known as the Perfected Spelling System. The aim of the new spelling system was to leverage the gap between the Malay spelling and the Indonesian one. With this reform, the roots of the original Dutch spelling system is only transparent in proper names. The list of changes can be found on the relevant Wikipedia article. It is advised to study the changes in historical order. The recent spelling system with the respective pronunciation can be found on [Omniglot](#).

7.4 Previous research on the language

There are numerous publications referenced throughout this report, which can be found in the bibliography at the end of this document. For a collected list of those researches and publications that are not mentioned in this article, including corpora, lexical resources, linguistic researches, dictionaries and more, the reader should visit the [the OLAC resources](#) site and the [the WALS resources](#) list. A more specialized list of publications can be found on the [website](#) of the Faculty of Computer Science of University of Indonesia Iho. The aim of this latter site is to collect all publications related to computational modelling of the Indonesian language.

7.5 Data and sources

7.5.1 Basic vocabulary

[Mylanguages.org](#) aims at teaching basic level knowledge of several languages and cultures, including Indonesian. The site categorizes words by part-of-speech (i.e., [nouns](#), [verbs](#), [adjectives](#), etc) and by topics (e.g., [people](#), [food items](#), [objects](#), etc) as well. Some of the pages have audio aid to show the correct pronunciation. It also has [a list](#) of the most commonly used phrases that consists of expressions for basic interactions, like greetings, introductions, seeking help and so on. Besides offering a vocabulary, the site also teaches the foundations of Indonesian grammar and culture.

Covering the basics of Indonesian, [Studiindonesian.com](#) also offers a collection of everyday [vocabulary](#), which is divided into several topics.

[Expat.or.id](#) targets those who would like to move to Indonesia or spend extended amount of time there. Besides describing the fundamental parts of Indonesian culture, business and more, it also has a [collection](#) of basic Indonesian expressions with audio aid. The site is founded by the [Indonesia Australia Language Foundation](#).

It also worth mentioning that there is an Indonesian WordNet called [WordNet Bahasa](#). The whole project can be downloaded from the website of the [NTU Computational Linguistics Lab](#), which collects WordNets in different languages. A WordNet is a lexical database that groups words by semantic relations, especially synonymy, and provides descriptions, as well as usage examples of words and the respective place of these in the structure of the net. Besides being an advanced thesaurus, it is also used in text analysis. About the creation of the Indonesian WordNet, one should read [Noor et al. \(2011\)](#). The article is available [online](#).

The GitHub user *ardwort* has a repository entitled [freq-dist-id](#), which is probably the largest Indonesian word frequency list one can find. It was constructed by scraping the Indonesian Wikipedia, Twitter, Kompas, an Indonesian news site, and Kaskus, an Indonesian forum.

7.5.2 Dictionaries

The most popular online dictionary for Indonesian is [Kamus.net](#). According to its own description it is “dedicated entirely to the Bahasa Indonesia language, Kamus.net provides instant translations to thousands of words, featuring dictionary definitions in English and Indonesian from several respected resources”.

Another reliable online dictionary is [Bab.la](#), which also uses user based verifications in translations.

[Kateglo.com](#) is an Indonesian dictionary, thesaurus and glossary application. It offers descriptions of more than 70,000 words and dictionary entries of more than 190,000. The word descriptions are in Indonesian. There is also [an API](#), but it is in a very initial stage. It supports *XML* and *JSON*. The dictionary most referenced by the OLAC entry of Indonesian is [Sederet.com](#). Besides functioning as a dictionary, it is also able to translate between English and Indonesian, using statistical methods. It also provides plugins for Firefox and Google Chrome.

All of the dictionaries mentioned above are English-Indonesian, Indonesian-English ones.

Kamus.net and Bab.la can be scraped using a word list of either Indonesian or English words, creating unique URLs, then parsing the HTML structure of the site, and fetching the translation from the HTML element that contains it.

7.5.3 Corpora

Besides the available newsportals and the previously mentioned Kaskus forum, the largest monolingual corpora one can find is [the Indonesian Wikipedia](#). According to the English Wikipedia [article](#), the Indonesian Wikipedia has over 382,777 articles, and it is the fifth fastest growing Wikipedia. To download the contents of the Indonesian Wikipedia, one should visit [the latest Wikidump](#) page.

The An Crúbadán project aims to create corpus for minority languages (in this sense, minority is about documentation not speaker population). Their site offers a [downloadable corpus](#) consisting of words from 1660 documents.

Another GitHub user, desmond86, created a repository called [Indonesian-English-Bilingual-Corpus](#) that is a bilingual corpus consisting of English and Indonesian Languages.

7.5.4 News portals

Indonesia has almost 100 million internet users and because of the quickly developing infrastructure, this number is rising. Even though there is internet censorship in several parts of Indonesia, there is a very stable online community. Indonesia has several popular news sites, some of which are present

on the list of most popular websites in Indonesia compiled by [Alexa.com](#). The greater news sites are the following:

- [Antaranews.com](#) is a news site providing realtime news from Indonesia and around the world.
- [Bisnis.com](#) is a business news portal.
- [Detik.com](#) is a popular Indonesian news site established in 1998, featuring local and worldwide news.
- [Garudamagazine.com](#) is a lesser-known Indonesian news site.
- [Hariansib.co](#) is a minor Indonesian news site.
- [Jawapos.com](#) was the most read news site in 2014.
- [Kompas.com](#) is a leading Indonesian news site established in 1995.
- [Mediaindonesia.com](#) is a relatively young news portal.
- [Pikiran Rakyat](#) is a magazine about “the thoughts of people”. This is the online version of a paper-based magazine.
- [Republik.co.id](#) is a popular news site hosting news in several topics.
- [Sinarharapan.co](#) is a Jakarta based news site, focusing on local issues.
- [Suara Pembaruan](#) is a medium sized news site focusing on recent issues, local and global alike.
- [Tempo.com](#), in its own words, an “in depth and most trusted news portal in Indonesia”.
- [Tribunnews](#) is the fifth most visited website in Indonesia.
- [Waspada.co.id](#) is a lesser-known news site hosting local and global news.

All of the above sites are linked from [Dailyindonesia.com](#). The site also has a collection of Indonesian blogs.

Even though it is not a news site, it is worth mentioning that one of the most popular sites of Indonesia is [Kaskus](#), which is a forum and online warehouse and marketplace. It was established in 1999, and according to the English Wikipedia, in 2012 Kaskus had more than 4 million registered users and over 650 million unique posts.

7.6 Computational tools

7.6.1 Language identification

Because of the spatial proximity between Malay and Indonesian, distinguishing between the two can be a challenging task. The first identifier worth mentioning is [Google Translate](#). It supports Indonesian language since 2008. Besides statistical methods, the Google Translate engine also uses WordNet databases as the foundation of translation between languages. There is an Indonesian Wordnet called [Wordnet Bahasa](#), which has been deployed by Google. The translations made by Google between English and Indonesian has exceptional quality.

There are several other language identifiers as well. [TEXTCAT](#) is an N-gram based language guesser written in *Perl* that supports Indonesian besides 68 other languages. [Xerox](#) has an online language identifier, which can be implemented to websites or softwares through its REST and SOAP API. [Whatlanguageisthis.com](#) is another basic online language identifier. A GitHub project called [Compact Language Detector 2](#), written in *C*, is a free and open source solution as well.

7.6.2 Tokenizer

Building a tokenizer for Bahasa Indonesia is a not a challenging task if one takes into consideration that every word in a text is either the first one, is preceded and followed by a space, or followed by a sentence closing symbol or comma. There are some higher level tools mentioned below, which are capable of tokenizing in order to accomplish some more complex tasks. But to mention at least one tokenizer, that is not capable of anything else, besides creating an array of words described above, one example is [Sastrawi/tokenizer](#), a repository on GitHub, which is part of a greater NLP project for Indonesian language. [Morphadorner V2.0](#) is a *Java* based project, that can be run either as a command-line controlled program or hosted as a web service. The tool, including the HTTP based server, together with full documentation, are available online.

7.6.3 Stemmer

A stemmer or a stemming algorithm is a tool which reduces inflected words to their word stem, which may or may not be equivalent to the morphological root. The criteria is that syntactically related words should have the same stem. Stemming is widely used in the field of computational linguistics, mostly in areas related to information retrieval, e.g. query system or search engines.

There are many projects which aimed for creating such tools, all but one of them are open source and available on [GitHub](#):

- There is a text preprocessor project on GitHub called [Kemangi](#). Text preprocessing covers several methods that clean raw textual data before actual text mining tools can be applied. Using the described in [Adriani et al. \(2007\)](#), this *Java* based project is able to remove non-ASCII characters, to turn every character into lower-case, to remove words or expressions that can be defined by the user via regular expressions and to reduce words into stems.
- A GitHub user called apraditya has a repository, [indonesian_stemmer](#), which is a stemmer tool written in *Ruby*. The creator of the stemmer, Adinda Praditya, claims that he based his work on [Tala \(2003\)](#).
- [Sastrawi](#) is another GitHub stemmer project written in *PHP*. The tool has a well written English documentation and it can be used free of charged under the MIT license. The author of the stemmer based his work on three articles: [Asian \(2007\)](#), [Arifin et al. \(2009\)](#) and [Tahitoe \(2010\)](#).
- [Pengakar](#) by Ivan Lanin is written in *PHP* as well, and is also a free and open source stemmer on GitHub.
- [PySastrawi](#) is written in *Python* by Hanif Amal Robbani, a developer from Jakarta. Unfortunately, the documentation of the project is written in Indonesian.
- Another useful GitHub repository is [IndonesianStemmer](#), which is written in *Python* as well. The author is George-Bogdan Ivanov, who also based his work on [Tala \(2003\)](#).

- [Apache Lucene](#) is written in *Java*, and is protected by Copyright. The developers claim that it was also inspired by [Tala \(2003\)](#).

There is also a lemmatizer for Indonesian language. Lemmatizers work like stemmers, except they are aiming to find the proper morphological root of words using a dictionary. David Christiandy, a GitHub user from Indonesia, has a repository called [lemmatizer](#) that is an open source lemmatizer, which was written in *PHP* and its dictionary is an *SQL* database.

7.6.4 Spell checker

There are several spell checkers available for Indonesian. There is an Indonesian [library](#) for the widely used spell checker engine, Hunspell. There is also a [repository](#) called ardspellchecker for spell checking on GitHub based on a stem list. Its written in *Python*, and it is a *jQuery* plugin for browsers.

[Stars21](#) is an online spell checker service, providing spell checkers for 79 languages, including Indonesian. [Spellrich](#) is another free, online spell checker tool. [WebSpellChecker](#) is a tool which not only highlights misspelled words but also gives suggestions. Amongst many languages, it supports Indonesian as well, however the license costs 250\$ for a year. The developers offer a 30 day trial version.

There are spell checkers for popular text editors as well. [Libre office](#) has a spell checker library available on their website. Even though there are official language packs released for Microsoft products (see End-user support), there is an amateur language pack as well, that can be an interesting addition to list of official packs: the author of [indodic.com](#) claims that the dictionary available to download on the site can be added to Microsoft products. Spell check add-ons are also available for [Mozilla Firefox](#). There is a spell checking application available for download on [iTunes](#) for 2.99\$.

7.6.5 Phrase level and higher tools

There is Part-of-speech (POS) tagger by Andry Luthfi. His work can be found on GitHub, under the repository called [indonesianpostag](#). The tool is written in Java and uses Perl dictionaries, however, the developer lists UNIX based operation systems as a requirement. [Pebahasa](#) is the name of a repository on GitHub, which is an NLP project for Indonesian written in *Python*. Besides being capable of tokenizing, it also offers a POS-tagger implementation.

[MorphInd](#) is a morphological analyser, consisting of morphosyntactic and morphophonemic rules for Indonesian derivational or inflectional surface words. Documentation and downloads are available on the project site. It is only compatible with UNIX based operation systems.

[Idn-treebank](#) is the name of a treebank project on GitHub by an Indonesian developer, Fam Rashel. It was created manually, so it can be safely used as a basis for later POS-tagger projects or text parsers.

[Terbilang](#) is a *PHP* based number-to-word translator. It takes a number in numerical form as input and gives back the Indonesian word for that number. It was developed by an Indonesian mobile and web developer, Mulia Arifandi Nasution.

The [website](#) of the Faculty of Computer Science at the University of Indonesia also offers various tools, such as a symbolic parser, which creates tree structures of sentences, and a semantic analyser, which translates sentences to first order predicate logic using the output trees of the symbolic parser and the methods described by Richard Montague, called [Montague Semantics](#). In addition, there is a morphological analyser.

Ananta Pandu Wicaksana, a GitHub user from Indonesia has an [indonesian-news-sraper](#) project that is a news scraper, able to scrape textual data from several of the above mentioned news sites.

The full list of supported sites is the following: Kompas, Detik, Tempo, KapanLagi, TempoEnglish, Antara, Republika, Okezone, Liputan6, Viva. It is written in *Javascript*. It has several options, ranging from fetching titles, creating a log of information about published articles, fetching URLs from news, and to actually scrape the news themselves.

The [Hidden Markov Model Toolkit](#), that is a tool used for primarily speech recognition, has an extension written for recognizing Indonesian. It can be found on GitHub, under a project called [Indonesian-Speech-Recognition](#). However, the author admits that it is only a university school project, and it is not advisable to use it for serious tasks. For a recent project, developed by the Bandung Institute of Technology, one should read [Hoesen et al. \(2016\)](#). [Microsoft Translation](#) supports Indonesian language. Its enterprise solution is capable of parsing, translating and manipulating speech, either in real time or using audio files. For more information on its capabilities and pricing, visit Microsoft Translate website.

7.6.6 End-user support

Microsoft supports Indonesian language, as there are language packs for Windows and Microsoft Office as well, the latter also supports spell checking. To see the full list of available language packs for Microsoft products, one should consult the [Microsoft Support site](#).

Apple products generally support Indonesian language: an Indonesian interface is available in OSX and on most products. To see the full list of Apple products supporting Indonesian, visit the [Apple Support site](#).

There is a partial translation available for Ubuntu on [Launchpad.net](#). As of as of 01.09.16, the 48% of packages have been translated.

Bibliography

Mirna Adriani, Bobby NaziefS. M. M. Tahaghoghi Jelita Asian, and Hugh E. Williams. Stemming indonesian: A confix-stripping approach, 2007.

Agus Zainal Arifin, I Putu Adhi Kerta Mahendra, and Henning Titi Ciptaningtyas. Enhanced confix stripping stemmer and ants algorithm for classifying news document in indonesian language, 2009.

Jelita Asian. Effective techniques for indonesian text retrieval, 2007.

Thomas J. Conners and Jozina Vander Klok. Language documentation of colloquial javanese varieties, 2016.

Devin Hoesen, Cil Hardianto Satriawan, Dessi Puji Lestari, and Masayu Leylia Khodra. Towards robust indonesian speech recognition with spontaneous-speech adapted acoustic models, 2016.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. Creating the open wordnet bahasa, 2011.

James Sneddon. *Indonesian Reference Grammar*. Allen & Unwin, 1996.

Andita Dwiyoga Tahitoe. Implementasi modifikasi enhanced confix stripping stemmer untuk bahasa indonesia dengan metode corpus based stemming, 2010.

Fadillah Z Tala. A study of stemming effects on information retrieval in bahasa indonesia, 2003.

Chapter 8

Mandarin (Dávid András Imre)

Contents

8.1 Demography and ethnography	111
8.2 Main typological and syntactic features	113
8.3 Writing system, transcription	118
8.4 Previous research on the language	119
8.5 Data and sources	120
8.6 Computational tools	122
Bibliography	125

Introduction

The present chapter aims at providing an introduction of Mandarin. After a brief review of the demo- and ethnographic situation of the language and its speaking community, the typology, the writing system and the most notable researches about the grammar of the language will be outlined. The second part of this paper provides an account on the various available offline and online data sources of Mandarin, and an outlook at the available tools for the processing of the language.

8.1 Demography and ethnography

8.1.1 Name variants

The name *Mandarin* can be used to refer to at least three different concepts. The first one is the standardized version of Chinese, whose endonyms are Putonghua (in Mainland China) and Guoyu (in Taiwan). Secondly, *Mandarin* (or Northern Chinese) can be used as a collective term to refer to the collection of Northern dialect varieties. In this case, the endonyms are Beifang Fangyan, Beifanghua, Beifang Guanhua. The third concept is a historical one, *Mandarin* was the name of the lingua franca of imperial administration during the times of Ming and Qing dynasties. The main concern of this report is Mandarin in the second sense.

The term *Hanyu* has roughly the same reference as *Chinese*. *Zhongwen* in some rare contexts can be synonymous to *Hanyu*, but it is primarily used to refer to various layers of written Chinese. A rarely used term is *Zhongguohua*, which roughly means “any variety / the totality of varieties of spoken Chinese”.

ISO code for the Mandarin language is *cmn* ISO 639-3.

8.1.2 Geographic spread

The varieties of Northern dialect that constitute Mandarin are widespread in China. The most important research on geographic spread of Chinese dialects was done by Wurm et al. (1987). The name, Northern Chinese, is based on the fact, that Mandarin speaking countries are mostly located on the northern side of China. However, there are exceptions: northern countries, where Mandarin speakers are the minority. Inner Mongolia is one, though there are some Mandarin speaking parts. Sichuan is the second, where only half of the residents speak a Northern variety. Qinghai and Tibet are the two last; these regions are almost devoid of any Mandarin speaking groups.

On the south-eastern parts of China, there are five regions where are no Mandarin speaking populations at all: Shanghai, Zhejiang, Jiangxi, Fujian, Guangdong. Even though there are Mandarin speakers in Hunan and Guangxi, Mandarin speakers are not the most significant linguistic community.

There are larger Mandarin speaking groups outside China, mostly in South-east Asian countries, Australia, and the United States. The Chinese minorities worldwide mostly speak other Chinese dialects, but nowadays Mandarin becomes more and more prevalent amongst those who seek new life outside China.

8.1.3 Speaker populations

The most detailed information about Mandarin language groups can be found on the related [Ethnologue article](#). According to the information available there, there are 889 million native speakers of Mandarin in China, and 178 million, who speak it as a second language. This means that 70% of the Chinese population speaks a Mandarin dialect as their vernacular language. The total number of all Mandarin speakers of the world, native or not, exceeds 1 billion.

There are only two countries with more than a million Mandarin speakers. One is Taiwan, where there are 4,320,000 native speakers, and almost three times as many (15 million), who speak Mandarin as a second language. The other one is Singapore, with its 1,200,000 native speakers and 800,000 L2 speakers. Other countries with significant Mandarin populations are Burma (500,000), United States of America (487,000), Indonesia (460,000), Australia (336,000), Malaysia (200,000) and Hong Kong (95,000). There are several smaller Mandarin speaking groups across the globe, but they are not the dominant Chinese speaking minorities. Another problem here is that surveys not always make distinction between different Chinese dialects, so statistics based on them can lead to false conclusions.

8.1.4 Dialect situation

According to the Language Atlas of China, there are eight subgroups of Mandarin Chinese. These dialects are:

- Northeast dialect, spoken in Heilongjiang, Jilin, Liaoning, and in several parts of Inner Mongolia, close to the border of the before mentioned regions
- Zhongyuan dialect, spoken in the western parts of Jiangsu and Anhui, southern part of Shandong and Shanxi, Henan, southern end of Ningxia and Gansu, and some Qinghai territories, close to the border with Gansu, and in the southern parts of Xinjiang

- Beijing dialect, spoken in Beijing, and territories of Hebei, Liaoning and Inner Mongolia, which are north from Beijing, and in the northern end of Xinjiang
- Lanyin dialect, spoken in the northern half of Xinjiang, Gansu and in the northern half of Ningxia
- Jilu dialect, southern half of Hebei, Tianjin, and in the corer of Hebei above Beijing and Tianjin
- Jianghui dialect, spoken in the central third of Jiangsu, Anhui, and in the eastern end of Hubei
- Jiaoliao dialect, spoken in the Liaodong Peninsula and Shandong Peninsula, and small parts of the north-eastern corners of Heilongjiang
- Southwest dialect, spoken in western parts of Hubei, Chongqing, eastern parts of Sichuan, Guizhou, Yunnan, and parts of Hunan, close to the borders of the before mentioned regions

8.2 Main typological and syntactic features

8.2.1 Linguistic typology

Mandarin is an *isolating* language, which means that Mandarin lacks morphological complexity that languages like Latin or English have. In general, almost every syllable in Chinese has an associated meaning. Mandarin is a *monosyllabic* language, so to say. Every character in the Chinese scripts corresponds to only one syllable, but the converse is not true. This means, that the vast majority of words that is written with two or more characters can be broken down to compounds that are themselves words. This is of course something that is not universally accepted. There are words, which consists of more then one syllable, where one of the syllables has lost its meaning over time. Also, from a standpoint which defines 'word' as a syntactically and semantically independent unit of language, there is no reason to think of polysyllabic words as compounds. In dictionaries, if the author uses Pīnyīn transcription, it is easy to determine the opinion of the author: if there is a word that consists of multiple Chinese characters is transcribed to Pīnyīn with spaces between the syllables, he/she promotes that Mandarin is a monosyllabic language. However, textbooks for foreigners usually don't follow this fashion, and transcribes complex words without spaces. Because of this, using tools that convert Pīnyīn to Chinese Scripts can be confusing at first. More on this issue can be found in [Li and Thompson \(1989\)](#).

Another distinguishing feature of Mandarin language is that it is *topic-prominent*. This means that besides the usual grammatical and semantical relations languages usually have in their sentences, in Mandarin there can be an extra sentence constituent, the topic. Usually this defines what the sentence is about, and it is something, on which both parties in the discussion have previous knowledge. It is somehow similar to the anaphoric relationship, but while the markers of anaphoric relations are syntactically dependant on other sentence constituents, for example they are arguments of the verb phrase, the topic can be grammatically and (to some degree) semantically independent, as they neither take or are arguments. The topic is always at the beginning of a sentence, and it can be followed by a comma.

A third distinguishing feature of the language can be its unique phonological structure. Three important notions should be explained: initials, finals and tones, as these elements constitute every syllable. The initials are the consonantal beginnings of syllables. There are no sequences of consonants in Mandarin, so the initial contains at most one consonant. At most, because some syllables contain

no consonantal initial. To overcome this situation in phonological analyses, there is a zero initial as well. Together with the zero initial, there are twenty-two initials.

	Labial	Apical	Retroflex	Palatal-Alveolar	Velar
Plosives	/p/ – b	/t/ – d			/k/ – g
	/p ^h / – p	/t ^h / – t			/k ^h / – k
Nasals	/m/ – m	/n/ – n			ŋ
Affricates		/ts/ – z	/tʂ/ – zh	/tɕ/	
		/ts ^h / – c	/tʂ ^h / – ch	/tɕ ^h / – q	
Fricatives	/f/ – f	/s/ – s	/ʂ(s)/ – sh	/ç/ – x	/χ/ – h
Sonorants	/w/	/l/ – l	/ɿ~ʐ/ – r	/j/	

The finals are the second components of a syllables. There are thirty-seven finals, all of them are presented in the table below with IPA notation. However, different sources can differ in the finals they list, one can compare for example [Li and Thompson \(1989\)](#) with the [WikiBooks entry](#) on the finals in Mandarin Chinese. Please note, that the Pīnyīn spelling of some finals are subject to changes when they are in a special environment. For a comprehensive list on spelling, consult either of the before mentioned sources.

	/a/ a	/ə/ -	/o/ -		/ai/ ai	/ei/ ei	/au/ ao	/ou/ ou	/an/ an	/ən/ en	/aŋ/ ang	/əŋ/ eng
/i/ i	/iə/ ia			/ie/ ie			/iəu/ iao	/iou/ iu	/iɛn/ ian	/in/ in	/iɑŋ/ iang	/iŋ/ ing
/u/ u	/uə/ ua		/uo/ uo, o		/uai/ uai	/uei/ ui			/uan/ uan	/uən/ un	/uɑŋ/ uang	/iŋ/ ing
/y/ ü				/ye/ üe					/yɛn/ üan	/yn/ ün		

The third components are tones. Mandarin is a *tonal language*, which means that tones play a crucial role in determining the meaning of a given word: the syllable consisting of a given initial and final can constitute words with different meanings depending on the tone. Tones are sets of phonetics features: pitch contour, intensity contour and duration. Frequency levels of a given pitch contour can be represented on a scale from 1 to 5, where 1 is the lowest, 5 is the highest level. A tone can be described with a linear order of starting, ending and turning points of the pitch contour. There are four basic tones in Mandarin: a high level tone (55), a rising tone (35), a falling-rising tone (214), and a falling tone (51). Every syllable has a tone associated to it, but in context, it can change without modifying the meaning of the syllable. The rules describing the behaviour of tones are called *tone sandhi*. These rules tell how a basic tones can turn into one another. There is a fifth, derived tone as well, the neutral tone. The neutral tone is not an absolute one as the other four, because its pitch contour depends on the tone that came before it, so there are four realizations of the neutral tone.

Neither the Standard nor the Traditional script marks tones, but Pīnyīn does. There are two representation of the tones: the first is with numbers, where 1 represents high, 2 represents rising, 3 represents falling-rising, and 4 represents the falling tone. The same can be marked with diacritics as well: ̄, ́, ˇ, ˋ respectively.

More on linguistic typology, including topic prominence, syllables and their construction can be found for example in [Li and Thompson \(1989\)](#) or [Ross and heng Sheng Ma \(2014\)](#). The [Wikipedia article](#) on Standard Chinese phonology gives a good overview of phonetic features of Mandarin.

As sources on Mandarin language can differ to some degree, e.g. the differences mentioned about syllables, or they give different phonemic inventories, the reader is advised to take every information with a pinch of salt. Also it should be kept in mind, that one attribute can be true of some Mandarin dialect and false of some others.

More on the historical and cultural aspects of Pīnyīn can be found later in this chapter.

8.2.2 Predication

Simple predicative sentences usually follow the SVO order. This is sometimes questioned. More on this can be read in first grammar referenced at the end of this section. Because it is a special topic that cannot be discussed here, in this report SVO order is assumed.

In transitive verb phrases, the object(s) follow the verb. If there are multiple arguments of a verb (e.g. *give someone something*, then usually the direct object comes after the indirect one. Prepositional phrases precede the verb phrase. If there are more than one spatial or temporal preposition, the order is from the biggest to the smallest (e.g. country - city - street, or day - part of day - hour). Generally, if there is a spatial and temporal preposition as well, the temporal precedes the spatial. Adverbial phrases occur before the verb phrase, usually after the prepositional phrases, but based on context and the speaker intention regarding what he/she wants to emphasize, this can change. Phrases which express duration of time occur after the verb phrase. This is because in Mandarin, they are not prepositional phrases. Noun phrases are somewhat more simple: the head noun comes last in the construction, every word modifying it comes before that in a common-sense semantical order.

More on the structure of predicative sentences and word order in their constituents can be found in [Li and Thompson \(1989\)](#), and in [Ross and heng Sheng Ma \(2014\)](#).

Nominal predication is also possible in Mandarin. There are two cases: copular and direct nominal predication. Copular predication is of the format *X shì Y*, where *shì* is the copula. Its notable feature is that the relation between X and Y is not necessarily the strict type of specification, classification, or identification as in Indo-European languages, but a more loosely conceived topic–comment relation, with its specific construal partly dependent on the context.

Xiǎo Lǐ shì yīshēng.
'Little Li (is a) doctor.'

Direct nominal predication lacks a copula, and is of the format 'NP NP'. This construction is primarily used with noun phrases denoting quantities, nationality, personal qualities, etc. In most cases, the sentences contructed this way would be well-formed with the copula as well, moreover negation can only be formed together with the copula. Some examples:

Tā měiguórén.
(S)he (is) American.

Wǒ érzi bú shì 18 suì.
I son not - 18.
'My son isn't 18.'

Adjectival predication has two different constructions: one for scalar adjectives, and another one for absolute ones. Scalar adjectives directly serve as predicates, but are interpreted with comparative

force if used alone. For an absolute degree reading they must take some adverbial modifier, if none else than a semantically empty one: *hěn*.

Tāmen liǎ, shéi gāo?
they two who tall
'Who is taller of the two of them?'

Háizi hěn cōngmíng.
child - clever'
The child is smart.'

8.2.3 Possession

The possessive relationship of the form *someone has something* can be expressed by the verb 有/yǒu, which roughly means 'to have'. To negate it, 没/méi is used:

He has a girlfriend.
他有女朋友。
Tā yǒu nǚ péngyǒu.

I don't have a younger sister.
我没有妹妹。
Wǒ méi yǒu mèimei.

To express that something is one's belonging, the word 的/de is used, which plays the same role as the English preposition *of* when it glues together noun phrases, but possessive pronouns are compounds of the pronoun and 的:

the car of the neighbour
邻居的车
línjū de chē

(my / your / his / our) key
(我的/ 你的/ 他的/ 我们的) 钥匙
(wǒ de / nǐ de / tā de / wǒmen de) yào shi

For more on expressing possession, the interested reader is pointed towards [Ross and heng Sheng Ma \(2014\)](#), where besides the above mentioned forms, more ways of expressing possessions can be found.

8.2.4 Imperative

In Mandarin, just in almost any other language, the simplest of commands are simply uttering a verb phrase in the right context. Constructions like this behave as commands only to the direct listener:

Eat!
吃!
Chī!

Come here!

过来!

Guò lái!

A somewhat more complex construction can be made using the particles 著/zhe and 吧/ba together with a verb phrase. The previous two examples again in this form:

Eat!

吃著!

Chī zhe!

Come here!

过来吧!

Guò lái ba!

To soften a command, there are three common verbs to choose from, all of which are roughly equivalent to 'please'. The three verbs are 劳驾/láoijià, 请/qǐng, 麻烦/máfan. Out of the three, 请/qǐng is the most common, and has the less constraints. Using the pronoun after it is optional, as in the case of simple commands. The other two have some semantical constraints: they are used if the command is beneficial to someone, other than the addressee himself/herself. The pronoun referencing the addressee cannot be omitted in the case of 麻烦/máfan. Some examples:

Please, bring tea.

麻烦你来茶。

Máfan nǐ lái chá.

Please (you) sit down.

请坐上。

Qǐng zuò shàng.

Prohibitions are formed with the help of negative imperative particle 不/bù : 不要/bù yào, 不许/bú xǔ, which mean *do not*, *not allow* respectively.

Don't go out!

不要出去!

Bùyào chūqu!

You are not allowed to smoke!

你不许吸烟! Nǐ bú xǔ xī yān!

More on imperatives can be found in Li and Thompson (1989), and in Ross and heng Sheng Ma (2014).

8.2.5 Interrogative

In Mandarin, simple **yes-or-no questions** follow the exact same word order as the corresponding statement, except that the particle 吗/ma should be added to the end of the statement to form a question:

She is a Chinese person. Is she a Chinese person?

她是中国人。 她是中国人吗?

Tā shì zhōngguo rén. Tā shì zhōngguo rén ma?

He can speak Chinese. Can he speak Chinese?

他会说中文。 他会说中文吗?

Tā huì shuō zhōngwén. Tā huì shuō zhōngwén ma?

The meaning of the English expression 'whether' can be roughly expressed with 是否/shìfǒu. Given any statement, putting this before the head verb turns the statement to a yes-or-no question. The importance of this method of question forming is that this is mostly done in written Chinese.

You like her. Do you like her (or not)?

你喜欢他。 你是否喜欢他?

Nǐ xǐhuān tā. Nǐ shìfǒu xǐhuān tā?

Wh-questions are formed with the help of a question word, just like in English. The syntax of question words are somewhat simple, since the questions follow the same order as the corresponding statements (answers), and the question words are in the same position as the part of the statement answering that question word. This rule-of-thumb is hardly applicable in some cases, e.g. in the case of *how*. For more information on those cases, the reader should consult any of the referenced grammars. The following list consists of several question words:

who	谁	shéi
what	什么	shénme
when	什么时候	shénme shíhòu
where	什么地方	shénme dìfāng
why	为什么	wèi shénme
how	怎么	zěnme

More on Mandarin Interrogatives can be found e.g. in [Ross and heng Sheng Ma \(2014\)](#), where most of these examples are present.

8.3 Writing system, transcription

Chinese writing system looks even more obscure for outsiders than the system of language varieties. There are two main Chinese scripts, the Traditional Script and the Simplified Chinese script. Written Chinese consists of logograms, which represent syllables, often whole words, instead of phonemes. Logograms can be combined to form more complex symbols. There are tens of thousands of different characters, though some of them are minor graphical variants of each other. To achieve functional literacy, a Chinese person should know about 3-4000 characters. The Chinese writing system is more than 3000 years old, making it the oldest, continuously used writing system.

The Traditional script is traditional in a sense, that it does not contain any changes made after 1956, when the first set of simplified characters was released by the government of the People's Republic of China. There were attempts at creating some set of simplified symbols for people have seen it as an obstacle in the way of creating a modern China. It also aimed to increase literacy rate. The standardization process hasn't been finished in 1956. The last version of accepted characters was published in 2013.

Regarding Mandarin language, the use of the Standardized script is prominent. The government of People's Republic of China names the Simplified script as standard, and holds back the Traditional Script only for cultural usage. Of course, this is only true to the mainland, while in Hong Kong, Macau and Taiwan the use of the Traditional script is the dominant. Again, this is something that fits perfectly into the “one country, two systems” philosophy of the Chinese government.

The traditional classification of Chinese characters divides them into six groups, and this is a concept unchallenged since 200 AD, when it was first written down by Xu Shen, a philologist of the Han dynasty. The first category is pictograms, stylised drawings of objects. The second is idiograms, simple drawings referring to abstract ideas, e.g. the number three is written with three strokes. Compound ideographs are complex signs consisting of at least two characters from the previous groups, where the parts suggest the meaning of the whole, e.g. ‘truthful’ consists of ‘person’ and ‘speech’. Rebus characters are those, that were borrowed to write an another homophonous word. The fifth category, which covers 90% of Chinese characters, is also based on combining two or more characters, but in a different way, and is called phono-semantic compound characters. Every such character is made of a rebus, and a declarative part. The rebus has approximately the right pronunciation, while the declarative plays a semantic role, namely it gives some clue about the meaning. Most of the times, the declarative is the radical of the compound, which means that in a dictionary, this is the character under which phono-semantics compounds can be found. The sixth category, called derivative cognates, is the smallest and least understood category, where there are no established rules, only guesses, that the characters in this category has something to do with etymological roots. Usually, this category of characters are included in the appendixes of dictionaries.

The basis of the simplification process was structural simplification of some characters, some of which are ‘stand-alone’ logograms in a sense, that they cannot be used as constituents in compounds. The others can be used to construct complex logograms. The simplification took place in line with some basic principles: replacing characters with more simple ones that sound the same, replacing components with a more simple shape, replacing arbitrary parts with some phonetic components, adopt simpler variants. Compounds should be also simplified in the same manner as their constituents if the simplification principles can be applied. Other than this, variants of the same characters have been eliminated, and some ‘vulgar’ characters have been adopted as standardized versions of a given logogram.

There were several attempts at creating transcriptions for Chinese, but the romanization of Chinese language that has become standard is called Pīnyīn. The development of Pīnyīn started in the 1950's by a governmental committee lead by Zhou Youguang, who is often referred to as “the father of Pīnyīn”. The first draft was presented in 1956, and two years later, the first version got accepted by the People's Congress. The aim of the romanization, besides helping international communication, was to increase literacy rate. It is used to teach pronunciation of the Chinese logograms, both for children and adults and for foreigners as well.

More on Chinese writing systems and their transcriptions can be found on the [Wikipedia article](#) on Chinese language, where other valuable sources can be found. More on Pīnyīn can be found in this chapter, where the relevant sources are referenced.

8.4 Previous research on the language

A comprehensive list of previous linguistic researches can be found on the [OLAC entry](#) on Mandarin Chinese. This collection includes papers and books about researches done in numerous linguistic areas

from phonology to sociolinguistics. There are also dictionaries and corpora. Every resource that are available online is marked.

The other very valuable resource regarding researches is the entry on Mandarin Chinese of [WALS](#). The page lists several features of the Mandarin language together with the source. These features include phonology, morphology, sentence level syntax and lexical attributes.

8.5 Data and sources

8.5.1 Basic vocabulary

There are several sites aiming at teaching Mandarin, some of which also provide a list of basic words or phrases used in everyday life. [BBC Schools](#) offers basic lessons in some languages, including Mandarin. The target audience is children. Besides some introduction to the language, the vocabulary covers topics such as numbers, calendar, food and drink, culture, family members, basic activities. It also provides audio aid. [Learnchineseez.com](#) is a site offering lessons in Cantonese and Mandarin as well, including basic vocabulary that covers common topics, like numbers, everyday objects, people, clothing, body parts, animals, food, nature, and also has audio aid for every word. [Chinese-tools.com](#) is a well-developed site with the same ambitions as the before mentioned ones, but it has the largest vocabulary of the three. The vocabulary consists of four big areas: themes, people, appearance, and health, and all categories have subcategories. Theme includes topics like food items, directions, animal sounds and astrology. The topic about people consists of categories like body parts, organs, emotions and life events. Appearance includes categories such as different clothings, accessories, and beauty. Health consists of topics like illnesses, injuries, workers and departments of hospitals, and first aid.

On a side note, there is a [frequency list](#) of the 25,000 most frequent items in the Gigaword Corpus, created by [Serge Sharoff](#), an associate professor at the University of Leeds.

8.5.2 Dictionaries

[Cambridge Dictionary](#) has an online dictionary for Simplified Chinese. The drawback of this dictionary is that it only translates from English to Chinese. But for every word, it gives example sentences, in which individual words can be selected, and idioms, whose translations are also available. The dictionary can be scraped using simple HTML parsing scripts and an English word list, because for every word in the dictionary, there is a unique URL of the site for its translation.

In the earlier internet era, there were not many Chinese dictionaries, because inputting kanji characters was problematic. Newer online dictionaries, like the one on [mdbg.net](#) solves this problem, by putting an input field on the site, in which the user can draw symbols, and the system dynamically gives suggestions based on the content of the field. This dictionary also generates a unique URL for every translation, so entries can be scraped using those, however to use this method, one should have their Mandarin wordlist written in Simplified script, as the URLs contain such characters.

The dictionary on [yellowbridge.com](#) uses the same method to enable the users to search for Mandarin logograms, but there is also an option to input Pīnyīn and convert it to Chinese Script, and an English-Chinese dictionary. It also provides the Pīnyīn version of every word, and lists words that are etymologically related to a given word. This dictionary uses PHP, and there are no unique URLs for translations.

Besides those, [Google Translate](#) also can be used as a dictionary.

8.5.3 Corpora

The An Crúbadán offers Madarin corpora both for [Simplified script](#) and [Pīnyīn](#).

An excellent corpus for parallel corpus building is the The Lancaster Corpus of Mandarin Chinese, consisting of one million words. It was aimed to be a parallel corpus of FLOB and FROWN corpora, which are corpora for modern British and American English respectively. It can be downloaded from the website of the [University of Oxford Text Archive](#) in *XML* format. The detailed description of the corpus can be read in [McEnery and Xiao \(2004\)](#).

A slightly bigger corpus containing 2.5 million words is the cleaned version of Guo Jin's Chinese PH corpus by Chris Brew and Julia Hockenmaier. According to the [README file](#), the corpus was a side-result of their studies on Chinese word segmentation. The cleaning up was done mostly by removing punctuation marks, and recognition of proper names. This was somewhat necessary as the corpus is based on articles of a Chinese news site, Xinhua news. The corpus can be downloaded from the [ftp server](#) of the cognitive science department of University of Edinburgh.

The last corpus that should be mentioned is by far the largest, counting more than 30 million words. It's the fifth edition of the Chinese Gigaword corpus, created by Robert Parker, David Graff, Ke Chen, Junbo Kong and Kazuaki Maeda, maintained and distributed by the Linguistic Data Consortium. The full description of the corpus can be read in the [README file](#), which is accessible from the site of the [Chinese Gigaword corpus](#). However, this corpus is not free from people outside LDC. Charges can be found on the site.

The Linguistic Data Consortium has a treebank consisting of 1.5 million words, based mostly on online news and conversations. For more information on development and pricing, the interested reader should visit the [Chinese Treebank 8.0](#) project site.

8.5.4 News portals

Several popular news site has a Chinese version available. It is worth mentioning, that most of the articles on these sites has a corresponding English version. Because of this property of these sites, they could be used as a solid basis for corpus building.

- [The New York Times](#)
- [BBC](#)
- [People](#)

Some of the more popular Chinese news sites together with their Alexa rank:

- [news.qq](#) is the 2nd most visited site in China, “China’s largest and most used Internet service portal owned by Tencent, Inc founded in November, 1998”.
- [cctv.com](#) is the 23rd most visited site in China, “The official website of China Central Television (CCTV) provides programs, online TV programs, new TV programs and discussions”.
- [china.com.cn](#) is the 26th most visited site in China, led by the State Council Information Office, provides news, online media and online services.
- [sina.com.cn](#) is the 7th most visited site in China, hosts news, media, games and other online services.

- news.baidu.com is the 1st most visited site in China, the news site of the Chinese search engine Bing.
- news.163.com is the 45th most visited site in China, an online news site and community provider.
- news.sohu.com is the 4th most visited site of China, the news site of Sohu.com, which “operates a comprehensive business, community, wireless and other value-added services”.

Some more news site can be found in a list compiled by Chinawhisper.com, published in an article with the title [Top 10 Chinese Portal Websites](#).

8.6 Computational tools

8.6.1 Language identification

Identifying Mandarin Chinese is not a challenging task, even though characters from the two scripts of written Chinese are used in other languages, like Japanese.

The first tool worth to mention here is [Google Translate](#). It supports Simplified Chinese since 2005, and the Traditional Script since 2007.

Two other popular language guessers, [TextCat](#) and [Xerox](#), also capable of identifying Chinese. TextCat should be downloaded, it's a *Perl* library, and can be used locally. Xerox, however, only works online either by using their website or calling their services through their API.

8.6.2 Tokenizer

Tokenization is the process of breaking down a piece of text into more simple constituents, usually words. In most of the languages, where inter-word spaces are used, e.g. in English, this can be rather straightforward. But in Chinese, where there are no visible word boundaries, it can be a challenging task.

One of the possible workarounds is to base the algorithm on a dictionary. Since Mandarin does not rely on inflectional morphemes, this can be done with high accuracy. However, it is really hard to find tokenizer tools that have English documentation.

The very first tool worth mentioning is the [Stanford Word Segmente](#). It uses an underlying lexicon to achieve fast and accurate segmentation. The tool is freely downloadable together with previous versions from their site. The description of the segmenter can be read in [Chang et al. \(2008\)](#).

Chiyuan Zhang, a Boston based developer, developed a high performance word segmenter for Chinese, rmmseg-cpp. It was written purely in *Ruby*, and is freely available for download on his [GitHub](#). In his original work, only the user interface was written in Ruby, the core of the program was in *C++*. He advices to study this one if someone is more interested in the mechanics than in the performance of the tool. The original project, called RRMSeg, is downloadable from [rubyforge](#). He also created one with a UI written in *Python*, pymmseg-cpp, which is also available on his GitHub. This tool has English documentation.

[jcseg](#) on Google Code is a fast and accurate word segmentation tool that besides the Simplified Script, also supports the Traditional Script. It also uses a dictionary as its basis, but it is possible to change the default dictionary, either replacing it with a completely new one, or to add new files, as multiple dictionaries are supported. It was written in *Java*, but the very same tool can be found written in C or PHP: [friso](#) is the name of the C project, while the PHP project is called [robbe](#). Unfortunately, the documentation is written in Chinese.

For some other tools, which are also capable of word segmenting, see the Phrase level and higher tools section.

8.6.3 Conversion between Pīnyīn and Chinese Scripts

There are tools available for download, which are designed to convert between Chinese scripts and Pīnyīn. A GitHub repository simply called [pinyin](#) by a user called hotoo is such a converter, written in *JavaScript*, so it can be ran from Node.js or in browser through API, but the first one is more efficient and reliable. The very same tool, except with an additional *Python* interface, can be downloaded from a repository called [python-pinyin](#). A third one, also on GitHub, is a repository called [PinYin4Objc](#). It was written in *objective-C*. All three supports multiple Pīnyīn styles (tones written with diacritics or numbers, heteronyms, etc), both scripts (Simplified and Traditional), and are free to use.

8.6.4 Spell checker

Because of the unique nature of the Chinese scripts, both the Simplified and the Traditional, spell checking makes no real sense regarding written Chinese.

But it makes sense in the case of Pīnyīn. [Pinyin.info](#) is a site entirely dedicated to provide information about the romanization of the kanji scripts. It has a [spell checker](#) tool, which works in an interesting way: instead of checking each word, guessing what that word is and comparing their spelling, this tool only checks whether the words provided could have been written in Pīnyīn. So it accepts non-existent words if their spelling does not go against the rules of Pīnyīn.

8.6.5 Phrase level and higher tools

[Rakuten MA](#) is a word segmenter and a part-of-speech tagger for Chinese and Japanese, written purely in *JavaScript*. It comes with a pre-trained model, but there is an option to update the model or create new ones. There is a [demo page](#), and the whole project can be downloaded from GitHub.

The Stanford Natural Language Processing group, besides a tokenizer, also developed a [part-of-speech tagger](#), which is capable of tagging Chinese text. It was completely written in *Java*, but others developed wrappers for the core in different languages. All of these extensions are listed on the site of the Stanford Group. The full download comes with pre-trained models, but new models can be created on new datasets.

The Stanford NLP group has also created a syntactic parser aimed to solve the problem of word-order in translating between English and Mandarin. Instead of using only phase-structure based methods, they implemented several semantic relations as well to indicate relationship between the words. The full specification of the parser can be found in [Levy and Manning \(2003\)](#), and the tool itself can be downloaded from the [Stanford NLP website](#) dedicated for their parser, where additional information can be found as well.

[Microsoft Translation](#) supports Mandarin language. Its enterprise solution is capable of parsing, translating and manipulating speech, either in real time or using audio files. For more information on its capabilities and pricing, visit Microsoft Translate website. The Android mobile smart assistant application, [Dragon Mobile Assistant](#), is also capable of supporting Mandarin. The list of functionalities and information on pricing can be found on their website. [Siri](#), a similar application for IOS, is also capable of supporting Mandarin.

8.6.6 End-user support

Microsoft products support Chinese: Windows and Office products can be changed to have a Chinese interface, either Simplified or Traditional. However, for the Office products to display inputted characters correctly, fonts should be installed manually. Macintosh also supports Chinese localization for both scripts. Linux based operation systems also support Chinese interfaces after the required fonts have been installed.

Bibliography

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing chinese word segmentation for machine translation performance, 2008.

Roger Levy and Christopher D. Manning. Is it harder to parse chinese, or the chinese treebank?, 2003.

Charles N. Li and Sandra A. Thompson. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, 1989.

Anthony McEnery and Zhonghua Xiao. The lancaster corpus of mandarin chinese: A corpus for monolingual and contrastive language study, 2004.

Claudia Ross and Jing heng Sheng Ma. *Modern Mandarin Chinese Grammar: A Practical Guide*. Routledge, 2014.

Stephen Adolphe Wurm, Rong Li, Theo Baumann, and Mei W. Lee. *Language Atlas of China*. Longman, 1987.

Chapter 9

Persian (Nikolett Mus)

Contents

9.1 Demography and ethnography	127
9.2 The main typological and syntactic features of Persian	130
9.3 Writing system, transliteration	131
9.4 Previous research on the Persian language	132
9.5 Data and sources	132
9.6 Computational tools	134
Bibliography	136

Introduction

In the following sections, the ethnolinguistic situation (see 9.1), the main typological features (see 9.2), the writing system (and linguistic transcriptions; see 9.3), the previous research (see 9.4), the available linguistic data and sources (such as dictionaries, corpora, news portals see 9.5) as well as the computational tools (see 9.6) of the Persian language (Persian, Indo-European; Asia) are presented in detail.

9.1 Demography and ethnography

9.1.1 Name variants

Persian is the predominant modern descendant of Old Persian. The language is a southwestern Iranian language, which belongs to the Indo-Iranian branch of the Indo-European language family.

One of the endonyms of Persian is **Farsi**, which refers to the official language. However, the name Farsi is exclusively used in Iran. In Afghanistan, Persian is known as **Dari** (or **Dari-Persian**) since 1958 for political reasons. In contrast, in Tajikistan the language is known as **Tajiki**. The name *Persian* is usually applied to both variants in many sources, e.g. the ISO, the Academy of Persian Language and Literature, etc. Therefore, we refer to the language by using the name Persian in this section too.

The ISO 639-3 code of the Persian *macrolanguage*, i.e. a language containing other individual languages in the standard, is **fas**, which includes the variations spoken in Afghanistan, i.e. Dari (**prs**), and the Iranian Persian (**pes**).



Figure 9.1: The Persian speaking territories in the Middle East (source: [farsinet](http://www.farsinet.com/farsi/))

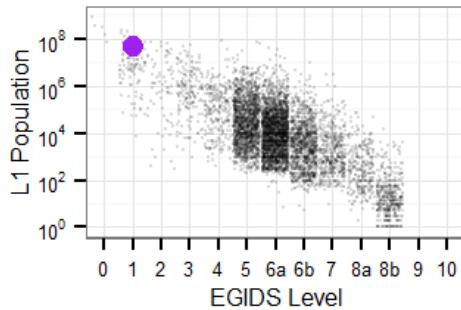


Figure 9.2: The EGIDS level for Persian spoken in Iran and in Afghanistan (source: [Ethnologue](#))

9.1.2 Geographic spread

The Persian language is spoken in Iran, Afghanistan and Tajikistan. In addition, there are speakers in other countries, such as in Uzbekistan, Bahrain, Iraq, Turkey, Kuwait, Israel, Turkmenistan, Oman, Yemen, the UAE and the USA. Figure (9.1) illustrates the Persian speaking territories in the Middle East.

9.1.3 Speaker populations

There are c. 110 million people who speak the Persian language as their mother tongue. The status of the language is 1 on the EGIDS (Expanded Graded Intergenerational Disruption Scale) scale both in Iran and in Afghanistan, which means that Persian has the strongest vitality level on the scale. Figure (9.2) illustrates the EGIDS level for Farsi and Dari, respectively. According to these data, the language is a national language used in education, work, government and mass media in these countries (cf. the Ethnologue entries for [Farsi](#) and [Dari](#)).

The Table 9.1 summarizes the number of speakers and the EGIDS level of Persian.

9.1.4 Dialect situation

There are three variants of standard Persian (i.e. the **Western Persian**, the **Eastern Persian**, and the **Tajiki**), which are based on the classic Persian literature. The number of speakers of these three

Table 9.1: The Persian speakers

Country	Number of L1 speakers	EGIDS level	Ratio of L1 speakers
Iran	45,000,000	1	58,1%
Afghanistan	7,600,000	1	24,88%
Pakistan	1,000,000	no data	0,55%
Turkey	624,000	no data	0,83%
United Arab Emirates	346,000	5	3,7%
United States	336,000	no data	0,1%
Iraq	227,000	5	0,68%
Quatar	170,000	5	7,84%
Canada	154,385	no data	0,44%
Kuwait	137,000	no data	4, 07%
Israel	135,000	no data	1,67%
Germany	97,000	no data	0,12%
Tajikistan	50,000	5	0,625%
Bahrain	48,000	no data	3,76%
Yemen	42,000	no data	0,17%
Uzbekistan	38,000	no data	0,12%
Oman	25,000	5	0,69%
Turkmenistan	13,000	no data	0,25%
Russia	3,600	no data	0,0025%

Table 9.2: The Persian dialects

Dialects	Number of speakers
Western Persian (Iranian Persian/Farsi)	47,000,000
Eastern Persian (Afghan Persian/Dari)	12,500,000
Tajiki (Tajik Persian)	7,900,000

main dialectal groups is illustrated in Table 9.2.

The Western and the Eastern varieties can be divided into further subdialects. The corresponding subdialects are illustrated in Table 9.3.

There are further local dialects of Persian. The dialectal variants spoken in Iran are the following: Abadani, Araki, Bandari, Basseri, Esfahani, Kashani, Kermani, Ketabi, Mahalhamadani, Mashadi (Meshed), Old Shirazi, Qazvini, Sedehi, Shahrudi, Kazeruni, Shirazi, Shirazjahromi, Tehrani, Yazdi.

While in Afghanistan the Darwazi, Tangshew (Tangshuri) dialects can be differentiated. The lexical similarity among Dari dialects is 86-90%. According to [Ethnologue](#), the most Afghan dialects are reportedly more similar to literary Persian than Iranian dialects are.

Table 9.3: The subdialects of the Western and the Eastern dialectal groups of Persian

Western	Eastern
Tehrani	Kabuli
Abadani	Mazari
Araki	Herati
Bandari	Badakhshi
Esfahani	Panjshiri
Karbalai	Laghmani
Kashani	Sistani
Kermani	Aimaqi
Mashhadi	Hazaragi
Qazvini	
Shirazi	
Yazdi	

9.2 The main typological and syntactic features of Persian

9.2.1 Linguistic typology

The language is accusative at the clause level, which means that the subject of the intransitive and transitive verbs (S and A respectively) are distinguished from the object (O) of the transitive verbs.

In contrast, the action nominal constructions are expressed by ergative-possessive structures (for further information see the corresponding [WALS](#) entries).

Phonological level The Persian language has six vowels (3 short vowels /a e o / and three long ones /ā ī ū /), 4 diphthongs and twenty-three consonants. For a detailed description of the phoneme system of Persian see [The Persian Online Grammar and Resources](#) of the University of Texas at Austin.

Morphological level The Persian language is a prepositional language with weak inflection expressed by suffixes. Instead the inflectional morphology, the grammatical relations are typically expressed by (productive full and partial) reduplication. In addition, there are prepositions and a postposition.

Morphosyntactic level The nouns in Persian are specified for number and case. There are two numbers in the language, i.e. singular and plural. The plural number is expressed by suffixes attached to the nouns. Different plural suffixes are used for non-human and human nouns in Persian. There are two cases, i.e. nominative and accusative cases. The nominative is the unmarked case, while the accusative is marked by a suffix. There are further prepositions, which express the oblique, e.g. local cases. The Persian nouns do not have gender, nevertheless the 3rd person singular personal pronoun exhibits gender differences. There is an indefinite article in the Persian language. Definiteness of nouns is implied by the lack of this indefinite article. Finally, possession is expressed by affixes added to the possessed nouns, which are followed by the possessor standing in genitive.

Verbs functioning as predicates in the language are specified for tense, aspect, mood, and agreement. Tenses in Persian, i.e. present, past, perfect, pluperfect, are mostly expressed by affixes. The verbal aspects can also be expressed by affixes. The tense and aspect markers appear in a mixed order on the predicate verb.

Syntactic level The basic word order of Persian simple declarative clauses is SOV. Similarly, this head-final strategy is described in phrases containing demonstrative (Dem) and noun (N), as well as, numeral (Num) and noun. Thus, Dem-N and Num-N orders surface in these phrases. In contrast, head-initial patterns can be found in the genitive (Gen), and in the adjectival (Adj) phrases, i.e. they appear in N-Gen, N-Adj orders. The relative clauses (Rel) follow their noun heads too (N-Rel).

9.2.2 Predication

The predicate verb agrees with its subject in person and number expressed through suffixes. There are two sets of verb conjugation endings. The one is used in the present tense, and the other one is used in the past.

من دروز به سینما رفتم
man diruz be sinamā raftam.
'I went to the movies yesterday.'

The nouns and adjectives may also appear as predicates in clauses. Then, a copular verb bearing the verbal grammatical meanings is always obligatory in the clauses. The nominal predicate construction is the same that is used in the so-called locational predicates.

اَنْ مَرْدُ مَعْلُمٌ مَاسِتٌ
in mard mo'allem-e māst.
'This man is our teacher.'

9.2.3 Possession

There is the possessive forms of personal pronouns used in possessive constructions.

خَانَةٌ مَا خَلِي بَزَرْگٌ اسْتَ
xāne-ye mā xeyli bozorg ast
'Our house is very large/big.'

9.2.4 Imperative

The imperative is restricted to the 2nd person singular. Usually, an imperative clause is formed by adding the prefix *bé-* to the present stem of the verb.

بَنْشِينَ
bénéšin!
'Sit!'

In the plural number, the second person plural ending is added to the imperative form. In prohibition, the combination of the normal imperative and normal negative structures is used.

9.2.5 Interrogative

The interrogative phrases in wh- (or content) interrogatives do not appear in initial position. Rather, they remain *in situ*, i.e. in the position in which a non-interrogative element fulfilling the same grammatical function appears within the clause.

کی قلم دارد؟
ki qalam dārad?
'Who has a pen?'

9.3 Writing system, transliteration

The modern Persian writing system is based on the Arabic script. The so-called *Perso-Arabic script* is an extended and modified version of the Arabic one, i.e. four letters are added to the basic Arabic script (so it contains 32 letters).

Besides, there are two methods of Latin transliteration. The one is developed by the International Organization for Standardization. It is a simplified transliteration of Persian. The other Latin transliteration is based on the Common Turkic Alphabet. It was used in Tajikistan in the 1920s and 1930s. In the late 1930s, the Latin transliteration was replaced by the Cyrillic script. For further information concerning the Persian orthography see the corresponding [Omniglot](#) and [Ethnologue](#) entries.

9.4 Previous research on the Persian language

Basic introduction and description of the language are provided by *inter alia* the [Iran Chamber Society](#) and the [Kwintessential](#).

Furthermore, there is an [online grammar of Persian](#) written and developed by the University of Texas at Austin, which provides language resources as well.

In addition, another [online grammar](#) and NLP tools of Persian written and developed by Ali [Jahanshiri](#) is also accessible. The further tools are also available on the site: a verb conjugator, a vocabulary and an additional word builder.

Research and control bodies There is an [Academy](#) of Persian Language and Literature in Iran. Furthermore, university departments and institutes can also be found, where Persian is researched, e.g. at the [School of Languages, Literatures, and Cultures of the University of Maryland](#), at the [Near Eastern Languages and Civilizations at the University of Chicago](#), at the [Persian Faculty at the Georgia Tech School of Modern Languages](#), and at the [Shiraz University](#).

9.5 Data and sources

In this section, basic vocabularies, dictionaries and language corpora of the Persian language will be provided. In addition, a collection of Persian news portals can be found here.

9.5.1 Basic vocabulary

The [300 Languages](#) subproject of *The Rosetta Project* archives and constructs a universal corpus of Persian. The project provides parallel texts and audio recordings, as well. Furthermore, the An Crubadan project provides data such as character trigrams, word bigrams, and word lists of Persian, that includes 3,424,151 words of [Iranian Persian](#), 1,834,223 words of [Persian](#) and 412,897 words of [Persian \(written with Latin script\)](#).

A thematized [English-Farsi](#) basic vocabulary is available. Further set of vocabulary lists with audio data can be found on the page of the [University of Texas at Austin](#).

9.5.2 Dictionaries

Paper editions

There are Persian–English dictionaries edited by e.g. ([Aryanpur-Kashani, 1986](#); [Haim, 1987](#); [Miandji, 2003](#)).

Online dictionaries

Additionally, the following online dictionaries can be found:

- the [English Farsi Advanced Dictionary](#) provides English translations, example sentences both in English and Persian, and Latin transcriptions of the Persian words
- a dictionary developed by [T-labs](#) (which is scrapeable) contains English translations, but no transliterations

- the [Dictionary-Farsi](#) contains English translations, POS-tags, contexts, and example sentences in Persian without transliterations
- the [Aryanpour](#) dictionary provides English translations without transliterating the Persian script
- the electronic dictionary by [Ectaco](#) provides English translation

9.5.3 Corpora

Monolingual corpora

As of 30/11/2015, the Persian [wikipedia](#) contains 475,115 articles.

Additionally, there are translations available, which can be used for building parallel corpora:

- the [Persian](#) translation of the *Bible* is available online on the site of Jehovah's witnesses
- parts of *The book of Mormon* in Persian translation are available
- there is a Persian translation of [the Quran](#)
- the [Persian](#) translation of *The Universal Declaration of Human Rights* is also available

Bilingual corpora

There is a large, open access freely available Persian corpus, which was developed in Uppsala ([Uppsala Persian Corpus: UPC](#)). UPC is a modified version of the Bijankhan corpus. The orginal Bijankhan collection contains about 2,6 millions manually tagged words with a tag set that contains 40 Persian POS tags. Sentence segmentation and consistent tokenization containing 2,704,028 tokens and annotated with 31 part-of-speech tags was added to the Bijankhan collection in the UPC.

There is also an [Uppsala Persian Dependency Treebank: UPDT](#), i.e. a dependency-based syntactically annotated corpus, which contains 6,000 sentences (151,671 tokens). The treebank data is extracted from the UPC.

9.5.4 News portals

As previously mentioned, the Farsi language is spoken in Iran, Afghanistan and Tajikistan. Besides, there are speakers in other countries. In what follows, we will concentrate on the news portals of these three main countries.

The (mass) media in [Iran](#) is controlled by censorship, that monitors mainly the printed media. The control principally spreads over the religious critics. The Iranian news portals are published in Persian, and there are also news available in English. A collection of the Iranian radio stations is provided in the [MWList](#). Additionally, the [Islamic Republic of Iran Broadcasting](#) (formerly National Iranian Radio and Television) streaming live on the internet. Outside of the Arabic broadcast, this service provides Armenian, Aztabriz, English, Kurmanci, Tajik, Turkish and Urdu live streams too. In addition, the IRIB provides [TV channels](#), such as [Al-Alam](#) or [Al-Kawthar](#). The other most significant TV channel in Iran is the [BBC Persian](#) channel.

In [Afghanistan](#), the language of the (mass) media is Dari (Farsi) and Pashto, i.e. an Eastern Iranian, Indo-European language. The media was controlled by the Taliban government until the 2000s. Nowadays, the media environment in Afghanistan is not independent from propaganda, however, the situation started to improve more markedly. A list of radio stations in Afghanistan is available on

[surfmusic.de](#). In addition, a broadcast service provided by the U.S. government called [Radio Azadi](#) (formerly Radio Free Afghanistan) provides 12 hours daily programming in Afghanistan. Furthermore, the [TOLO TV](#) is one of the most watched TV station in the country. Additionally, many global news channels, such as CNN, BBC, Sky News, DD News, Al-Jazeera, have agencies of an undertaking in Kabul. Finally, the digital media is started to grow in the country (see e.g. the [Pajhwok Afghan News](#) and the [Khaama Press](#), a.o.).

In [Tajikistan](#), the information is mainly provided by the radio and TV stations. The radio stations broadcast in Persian, Russian, Tajik, and Uzbek. Lists of radio stations streaming live in Tajikistan are provided by [Streema](#) and [radio-asia.org](#). Furthermore, the [RadioFreeEurope](#) and the [Radio Ozodi](#) are available online both in Russian and in Tajik. There is a news portal called [Tajikistan News](#) that provides news in English.

9.6 Computational tools

This section introduces the main computational tools developed for Persian.

9.6.1 Language identification

The [CLD2](#) provides support for Persian. Furthermore, there is another [identifier tool written in python](#) available on [github](#). In addition, [T-Labs also developed an identifier](#). Finally, there is the [Basis technology Rosette Language Identifier](#) of Persian.

9.6.2 Tokenizer

A [tokenizer](#) for the Persian language developed by the Test Word Tokenizer service is available. Additionally, the [tools](#) developed by MarkLogic provide full language support (such as basic full-text search, tokenization using whitespace-delimiters and punctuation) for Persian.

9.6.3 Stemmer

A Persian stemmer called *Bon* was developed by Masoud Tashakory, Mohammadreza Meybody, and Farhad Oroumchian. The Bon stemmer was tested on a collection of Persian texts. For a detailed discussion of the developing and testing the Bon stemmer see [Tashakori et al. \(2002\)](#).

Furthermore, there is a stemmer developed for Persian by using a structural approach for stemming which uses the structure of words and morphological rules of the language to recognize the stem of each word. The developers composed 33 rules to describe a structural rule-based stemmer (for more details see [Rahimtoroghi et al. \(2010\)](#)).

In addition, discussions and further plans of Persian stemming can be found on the blog called [Persian stemming](#).

9.6.4 Spell checker

The open source spell checker, [HunSpell](#), is available for Persian. In addition, there is free online spell checker provided by [jspell.com](#) for Persian.

9.6.5 Phrase level and higher tools

Persian morphological analyzer Polyglot provides a [morphological analyser](#) for Persian.

Persian part-of-speech tagger The HunPOS tagger for Persian trained on UPC is also available. Web Technology Lab developed a Persian POS tagger as well.

Persian chunker There are attempts to develop chunkers (see e.g. Kiani et al., 2009).

Persian named entity recognizer Polyglot provides possibility for Named Entity Extraction in Persian.

Persian sentence parser Ghalibaf et al. (2009) describe a semantic role labeling system for Persian (PDF). Additionally, Persian Parser Tool can be downloaded.

Persian speech recognizer Sameti et al. (2008) reviews the Nevisa Persian speech recognition engine, that is an HMM-based speaker-independent continuous speech recognition system. In addition, LingvoSoft provides a speech recognition software for Persian. Besides, the Google Cloud Speech API supports Persian.

Persian machine translator There is a rule-based Machine translator developed by SDL for Persian. In addition, the Google translator supports Persian (for a detailed description and the supported languages see the Wikipedia entry of Google Translate).

Persian question answering machine Sherkat and Farhoodi (2014) describe a hybrid approach for question classification in question answering systems. Additionally, Mollaei et al. (2012) present the architecture of question classification based on the Conditional Random Fields machine learning model (PDF).

9.6.6 End-user support

The jahansiri website provides resources for Persian including Persian verb conjugator, word builder, phonetic keyboard layout and a Tajik-Persian transcriptor. The OS support for Persian is the following:

- a Microsoft Windows language pack is available in Persian
- an Ubuntu language pack is available for Persian
- the MAC OS does not support Persian

There is no Unicode range for Persian available.

The ABBYY FineReader Engine 11 allows developers and software engineers to include OCR functionalities for Persian into their applications.

There is a XePersian Package for LATEX2 using XeTEX engine.

Bibliography

- A. Aryanpur-Kashani. *The Combined New Persian-English and English-Persian Dictionary*. Mazda Publishers, Costa Mesa, 1986.
- Azadeh Kamel Ghalibaf, Saeed Rahati, and Azam Estaji. Shallow semantic parsing of persian sentences. In *PACLIC*, pages 150–159, 2009.
- S. Haim. *English-Persian Dictionary*. French / European Publications, New York, 1987.
- S. Kiani, T. Akhavan, and M. Shamsfard. Developing a persian chunker using a hybrid approach. In *International Multiconference on Computer Science and Information Technology, 2009*, pages 227–234, 2009.
- A. M. Miandji. *Farsi-English/English-Farsi (Persian) Concise Dictionary (Bilingual edn.)*. Hippocrene Books, New York, 2003.
- Ali Mollaei, Saeed Rahati-Quchani, and Azam Estaji. Question classification in persian language based on conditional random fields. In *Computer and Knowledge Engineering (ICCKE), 2012 2nd International eConference on*, pages 295–300. IEEE, 2012.
- E. Rahimtoroghi, H. Faili, and A. Shakery. A structural rule-based stemmer for persian. In *5th International Symposium on Telecommunications*, pages 574–578, 2010.
- Hossein Sameti, Hadi Veisi, Mohammad Bahrani, Bagher Babaali, and Khosro Hosseinzadeh. Nevisa, a persian continuous speech recognition system. In *Advances in Computer Science and Engineering*, pages 485–492. Springer, 2008.
- Ehsan Sherkat and Mojgan Farhoodi. A hybrid approach for question classification in persian automatic question answering systems. In *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*, pages 279–284. IEEE, 2014.
- Masoud Tashakori, Mohammadreza Meybodi, and Farhad Oroumchian. Bon: The persian stemmer. In *EurAsia-ICT 2002: Information and Communication Technology*, pages 487–494. Springer Berlin Heidelberg, 2002.

Chapter 10

Russian (Nikolett Mus)

Contents

10.1 Demography and ethnography	139
10.2 The main typological and syntactic features of Russian	143
10.3 Writing system, transliteration	145
10.4 Previous research on the language	145
10.5 Data and sources	146
10.6 Computational tools	148
Bibliography	151

Introduction

This chapter aims at providing linguistic data on the Russian language (Indo-European, Slavic), on the one hand, and a collection of computational tools used for linguistic research of Russian, on the other. The report is organized as follows. Section 10.1 introduces the demographic and the ethnographic situation of the language, particularly as regards the geographic spread of the language, the number of its speakers as well as its dialectal variants. Then, the main typological features of Russian will be discussed in Section 10.2. The writing system, i.e. the Cyrillic alphabet and its Latin transliterational and/or transcriptional traditions will be summarized in Section 10.3. Section 10.4 surveys the previous research on the Russian language. Section 10.5 provides primary data and available sources of the language (e.g. basic vocabulary, dictionnaires, corpora, news portals, etc.). Finally, computational tools and resources that support the theoretical and technology research activities in Russian linguistics, will be presented in Section 10.6.

10.1 Demography and ethnography

The section deals with the genetic affiliation of Russian and gives an account of the literature about the demographic and ethnolinguistic situation of the Russian language.

10.1.1 Name variants and classification

The **Russian** language, also referred to as Russkij, is an East Slavic language, which belongs to the Slavic group of the Indo-European language family. Its closest relatives are the East Slavic languages, like Ukrainian, Belarusian and Rusyn. The ISO 639-3 code of Russian is **rus**.



Figure 10.1: The Russian speakers in Eurasia (source: [wikipedia](#))

10.1.2 Geographic spread

Russian is spoken mainly in Europe and Asia. Russian is the official language of the Russian Federation, the Republic of Belarus and the Kyrgyz Republic. Besides, it is also spoken in several member states of the former Soviet Union, for instance in Ukraine, Georgia, the Republic of Kazakhstan, the Republic of Tajikistan, the Republic of Estonia, and the People's Republic of China, etc. In addition, a few more groups of speakers can sporadically be found in Poland, Bulgaria, Czech Republic, Slovakia, Hungary, Albania, and former East Germany. The Figure (10.1) illustrates the Russian speaking territories in Eurasia.

Outside Europe, Russian is spoken in North America and Australia.

10.1.3 Speaker populations

While Russian is the eighth most spoken language in the world by the number of its native (i.e. L1) speakers, it is the largest native language spoken in the continent of Europe. According to the latest population Census of the Russian Federation (2010), there are 137 million Russian speakers in the Russian Federation. This number is about 82,53% of the total number of people all around the word (i.e. **166 million**) who speak Russian as their mother tongue. The EGIDS (Expanded Graded Intergenerational Disruption Scale) level of Russian in the Russian Federation is **1**, which means that the language is a national language used in education, work, mass media, etc.

In the Republic of Belarus, 8,3% of the population is Russian, which is cc. 786,000 people (see the population Census of the Republic of Belarus 2009). The EGIDS level of Russian in Belarus is **1** (national).

According to the population Census of the Kyrgyz Republic (2009), which is the third country in which Russian is an official language, 9% (cc. **520,000** people) of the population is Russian. The EGIDS level of Russian here is **1** (national) as well. The Table 10.1 shows the number of speakers and the EGIDS level of Russian in particular countries.

According to [Ethnologue](#) the number of L2 speakers is 29,945,000.

In addition, Russian is one of the six official languages of the United Nations. Furthermore, it is used as official language by other international organisations and committees, such as the International Atomic Energy Agency, the UNESCO, the World Bank, the International Monetary Fund, etc. The total amount of the Russian speakers, including L2 speakers, is **260 million** in the world.

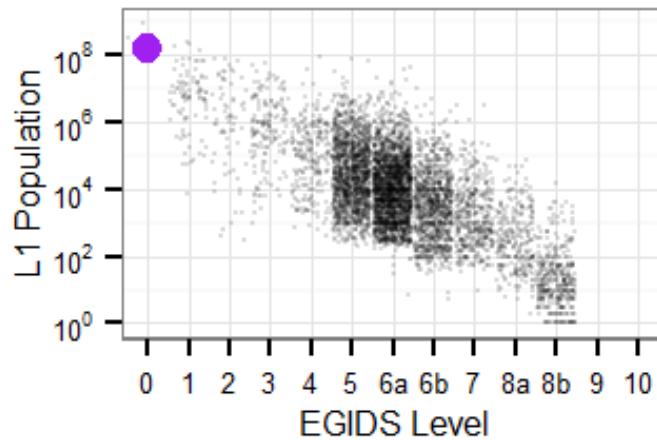


Figure 10.2: The EGIDS level for Russian spoken in the Russian Federation (source:[Ethnologue](#))

Table 10.1: The Russian speakers

Country	Number of L1 speakers	EGIDS level	Ratio of L1 speakers	Number of L2 speakers
Russian Federation	111,016,896	1	77,7%	no data
Ukraine	8,330,000	2	18,31%	no data
Uzbekistan	4,294,080	3	14,2%	no data
Republic of Kazakhstan	3,800,000	1	22,35%	15,200,000
Republic of Belarus	786,000	1	8,3%	no data
Israel	750,000	5	8,86%	no data
Moldova	541,000	3	15,2%	no data
Latvia	520,136	3	25,8%	1,390,000
Kyrgyz Republic	520,000	1	9%	no data
Azerbaijan	475,000	3	5%	no data
Georgia	372,000	3	3,72%	no data
Lithuania	344,000	3	11,64%	2,430,000
Republic of Estonia	326,235	3	25,2%	725,000
Turkmenistan	242,307	no data	5,1%	no data
Republic of Tajikistan	34,838	3	0,5%	no data
People's republic of China	15,393	5	0,00000112%	no data
Mongolia	4,000	3	0,14%	no data
Total	132,371,885			



Figure 10.3: The dialects of the Russian language (source:[Wikipedia](#))

Table 10.2: The Russian dialects

Dialectal group	Dialects	Population
Northern	Pomor	6,571
	Ladoga-Tikhvin	58,459
	Transitional groups (Onega, Lacha, Belozersk-Bezhetsk)	34,138
	Vologda	1,202,444
	Kostroma-Yaroslavl	1,940,030
Central	Western	735,000
	Eastern	15,009,850
	Chukhloma	5,411
Southern	Western	2,263,754
	Transitional group A	63,504
	Central	3,446,542
	Transitional group B	3,165,974
	Eastern	5,755,001

10.1.4 Dialect situation

The Russian language consists of three main dialectal groups, namely, the **Northern**, the **Central** and the **Southern** groups. Within them, one can distinguish further (sub)dialects. The Figure (10.3) illustrates the dialects of the three main Russian dialectal groups. In addition, the classification of the dialects are summarized in Table 10.2.

As seen, the Northern dialectal group contains the Pomor, the Olonets, the Novgorod, the VologdaKirov, and the Vladimir-Volga dialects. The Central one contains the Moscow and Tver dialects. Finally, the Southern group consists of the Orel/Oryol (Don), the Ryazan, the Tula, and the Smolensk dialects.

Despite the fact that Russian is the only official language of the Russian Federation, several indigenous (minority) languages are spoken as well. Some of them are given an official status according to the constitutions of the republics of the Russian Federation, e.g. Abaza (Northwest Caucasian)

in the Karachay–Cherkess Republic, Buryat (Mongolic) in the Republic of Buryatia, Yakut (Turkic) in the Sakha (Yakutia) Republic, Komi (Uralic) in the Komi Republic, Ossetic (Indo-European) in the Republic of North Ossetia-Alania, etc. The total number of languages which can be considered as official languages in the republics of the Russian Federation is 35. In addition, there are over 100 indigenous languages which are, however, either (seriously) endangered, e.g. Tundra and Forest Nenets (Uralic), Chukchi (Paleosiberian), Judeo-Tat (also called as Juhuri; Indo-European), etc. or extinct, e.g. Akkala Sami (Uralic), Udege (Tungusic), or Ainu (Ainu), etc. These minority languages may influence and may be influenced by the Russian language.

10.2 The main typological and syntactic features of Russian

10.2.1 Linguistic typology

Russian is a fusional language, i.e. a subtype of synthetic languages, in which several grammatical categories (such as number, case, etc.) are expressed simultaneously by one affix. In Russian, it is also common that certain grammatical meaning is expressed by the change of an internal phoneme of the root itself.

Russian is an accusative language, so the marking of the subject of intransitive and transitive verbs (S and A, respectively) is distinguished from the marking of the direct object (O) of transitive verbs. Besides, action nominal constructions exhibit an ergative-possessive pattern (cf. [Koptjevska ja-Tamm, 2013](#)).

Phonological level The Russian phoneme inventory includes 5 (or 6) vowels in stressed syllables (/a e i o u/ and in some analyses /i/) 2 (or 3) vowels in unstressed syllables (/a i u/ after hard consonants and /i u/ after soft ones). An important aspect is the reduction of the unstressed vowels. Stress, which is unpredictable, is not normally indicated orthographically (though an optional acute accent may be used to mark stress, such as to distinguish between homographic words, or to indicate the proper pronunciation of uncommon words or names; cf. [Wikipedia](#)). There are 34 consonants in the Russian language. In the consonant inventory, he so-called soft and hard sounds, i.e. consonant phonemes with palatal secondary articulation and those without, are distinguished.

Morphological level As mentioned, Russian is a fusional language that has rich morphology.

Morphosyntactic level The category of nouns is specified for two numbers: singular and plural; six (or seven) cases: nominative, accusative, genitive, dative, locative, instrumental; and three sex-based genders: feminine, masculine, and neutral (cf. [Corbett, 1991](#); [Fraser and Corbett, 1995](#); [Wade, 1992](#)). Russian expresses these categories either by prepositions or by suffixes, which are attached to the dependents in the phrases. There is no definite or indefinite article in the language. Additionally, Russian also lacks possessive affixes.

In Russian, verbs are specified for tense (present, past, future), aspect (imperfective, perfective), mood (indicative, conditional, imperative), and agreement. In the tense paradigm, it is only the past tense that is marked by affixes; the future is expressed by complex verb phrase while the present is unmarked. The language is rich in aspectual affixes. The tense and aspect affixes appear in a mixed order on the verb. The mood system in Russian covers the indicative, subjunctive, and imperative.

Syntactic level As far as syntax is concerned, the main constituents of the Russian declarative clauses appear in the SVO order; consequently, the subject occupies the clause initial position and precedes the verb, while the object is situated after the verb. Additionally, the order of the elements in possessive constructions, i.e. genitive (Gen) and noun (N), as well as the order of the noun and the relative clause (Rel) also follow the head initial pattern (hence, they appear in N-Gen, and in N-Rel order). The adjectives (Adj) and demonstratives (Dem), however, precede the nominal head (thus, their orders are: Adj-N and Dem-N). Accordingly, these phrases show typical head final patterns.

10.2.2 Predication

There is only one conjugational paradigm in the Russian language, which marks the agreement between the verb and its subject in two numbers (singular, plural) and three persons. Thus, there is no object marking on the transitive verbs.

- (50) Кошка поймала мышь.
 koshka pojma-la mysh.
 cat catch-pst.fem mouse

‘A cat caught a mouse.’

In addition, a nonverbal element, such as noun (phrase), adjective, adverbial can function as the predicate of the clause. In these predicate types, the copular verb is covert in the present tense in every person and number. These constructions express either identification or attribution.

- (51) Это кошка.
 eto koshka.
 this cat

‘This is a cat.’

The locational clauses are identical to the predicate noun and adjective constructions. Consequently, the copula is covert in the present tense.

10.2.3 Possession

In adnominal possession, the possessor is preceded by the possessed item. There is a set of possessive pronouns used to express possession.

- (52) Моя кошка поймала мышь.
 moja koshka pojma-la mysh.
 my cat catch-pst.fem mouse
- ‘My cat caught a mouse.’

10.2.4 Imperative

The imperatives typically convey commands, orders, requests, suggestions, instructions, and warnings. The Russian imperatives are used in the 2nd person singular and plural.

- (53) Поймай мышь!
 pojma-j mysh!
 catch-imp.2sg mouse
 ‘Catch (sg) a mouse!’

- (54) Поймайте мышь!
 pojma-jte mysh!
 catch-imp.2pl mouse
 ‘Catch (pl) a mouse!’

The prohibition is expressed by the combination of the regular imperative construction and the normal negation strategy.

10.2.5 Interrogative

The wh-(or content) interrogatives are used in a discourse when the speaker misses an element of a given statement and assumes that the hearer knows the required information. The interrogative phrase expresses this unknown information. In Russian, interrogative phrases are always situated clause-initially.

- (55) Что поймала кошка?
 shto pojma-la koshka?
 what catch-pst.fem cat
 ‘What did the cat catch?’

10.3 Writing system, transliteration

The Russian alphabet is based on the **Cyrillic script**. The early version of this script-type was developed in the 9th century. The Russian alphabet contains 33. Before the reform of the Russian orthography in 1918, there were four additional letters too.

There are Latin transliterations of Russian, i.e. textitromanization, also available. These transliterations conform to, for instance, scientific standards, such as the transliteration used in linguistics and science. A somewhat different transliteration is used in libraries, i.e. the GOST 7.79 (2002), or in passports, i.e. the ICAO system.

For various character charts of Russian, the corresponding [Omniglot](#) entry may be consulted.

10.4 Previous research on the language

A significant number of linguistic descriptions of the Russian language is available. There are, for example, grammars and grammatical descriptions, some of them are exemplified here (for more information see the [WALS](#) entry of the Russian language): [Unbegaun \(1957\)](#); [Borras and Christian \(1959\)](#); [Tauscher and Kirschbaum \(1963\)](#); [Stilman and Harkins \(1964\)](#); [Berneker and Vasmer \(1971\)](#); [Pulkina and Zakhava-Nekrasova \(1974\)](#); [Pulkina \(1978\)](#); [Wade \(1992\)](#); [Beyer \(2001\)](#). For further resources, the reader is pointed to the [OLAC entry on Russian](#).

The following sources provide a grammatical overview and online courses for L2 speakers of Russian: Practice Russian; Russian on-line; Master Russian; Russian Language Lessons; Russian For Everyone; Learn Russian; Online Russian lessons of the UCLA; Beginning Russian Grammar.

The series of [Studies in Slavic and General Linguistics](#) and [Studies in Russian Linguistics](#) are mainly devoted to the field of descriptive linguistics dealing with the Slavic languages, and more specifically, with Russian. Additionally, there are quite a few scientific journals dealing with Russian, such as [Russian Linguistics](#), [Journal of Slavic Linguistics](#), [Zeitschrift für Slawistik](#), etc.

Research and control bodies There are research institutes dealing with several aspects of the Russian language both in Russia: e.g. the [The Institute of Linguistics, Russian Academy of Sciences](#) in Moscow; the [Institute for Linguistic Studies Russian Academy of Sciences](#) in St Petersburg and the [Pushkin State Russian Language Institute](#) and outside of Russia: e.g. the [Russian, East European, and Eurasian Center](#) (henceforth REEC) at the University of Illinois. On the homepage of the University of Illinois, there is a collection of links to [REECs](#) and [Research resources](#), connected to Russian Studies.

10.5 Data and sources

This section contains information about the available primary sources of Russian.

10.5.1 Basic vocabulary

Online Russian vocabularies are available, *inter alia*, at the following sites: the [Russian lessons](#); the [Russian for Everyone](#); the [Learn Russian](#).

Additionally, the [An Crubadan](#) project also provides data (e.g. character trigrams, word bigrams, and word lists, etc.) of [Russian](#) on the basis of 46,212,981 words, and [Russian\(Latin\)](#) by 872,418 words. As of 30/11/2015, the Russian [wikipedia](#) contains 1,271,173 articles.

The [Swadesh List](#) is available in Russian.

10.5.2 Dictionaries

This section provides a brief overview of some available dictionaries of Russian.

Paper edition

There is a Russian–English dictionary edited by (e.g. [Howlett, 1996](#)) that includes over 120,000 words and phrases, with 190,000 translations. The entries are written in IPA and there is also a pronunciation guide.

Online dictionaries

The [Translatos.com](#) provides easily scrapable dictionaries for the following language pairs: Russian–Uzbek dictionary (\approx 108000 words and phrases), Uzbek–Russian dictionary (\approx 82000 words and phrases), Russian–Kazakh dictionary (\approx 85000 words and phrases), Kazakh–Russian dictionary (\approx 59000 words and phrases), Russian–Kyrgyz dictionary (\approx 49000 words and phrases), Kyrgyz–Russian dictionary

(≈34000 words and phrases), Russian–Turkmen dictionary (≈90000 words and phrases), Turkmen–Russian dictionary (≈93000 words and phrases), Russian–Tajik dictionary (≈9000 words and phrases), Tajik–Russian dictionary (≈42000 words and phrases).

[Lexilogos](#) provides a collection of English–Russian online dictionaries.

Further online dictionaires can be found, for instance, in the following entries:

- the English–Russian online dictionary by [Rustran](#) including POS tgas and translations
- the German–Russian–English online dictionary by [PONS](#) including translations, pronunciation (the user can listen the entries), context phrases, basic grammatical informations
- the German-Russian online dictionary of [LEO](#) providing translations, pronunciation (the user can listen the entries), POS tags
- the [bab.la](#) dictionary including translations, pronunciation (the user can listen the entries), context phrases, basic grammatical informations, POS-tags (easily scrapeable)

A further source that can be used for scraping is provided by the [LEGO](#) (Lexicon Enhancement via the GOLD Ontology).

10.5.3 Corpora

Monolingual corpora

Translations for building parallel text corpora are there available:

- the [Russian](#) translation of the *Bible* is available online on the site of Yehovah's witnesses;
- there is a [Russian](#) translation of the *Book of Mormon*;
- the *Quran* is also translated into [Russian](#);
- the [Russian](#) translaion of the *Universal Declaration of Human Rights*.

Additionally, the [Opensubtitles](#) contains Russian texts.

Bilingual corpora

The [Russian National Corpus](#) provides data of the modern Russian language containing over 350 million automatically lemmatized and POS-/grammeme-tagged words. The corpus contains morphologically annotated texts from the 1750s to the 1950s. Additionally, parallel text corpora (such as English–Russian, Russian–English, German–Russian, Ukrainian–Russian, Russian–Ukrainian, Belorussian–Russian, Russian–Belorussian, and multilingual), dialectal corpus, poetry corpus, educational corpus, and spoken Russian corpus can also be found in the Russian National Corpus.

Furthermore, the [300 Languages Project](#), which is a sub-project of *The Rosetta Project* intending to collect materials of the 300 most widely-spoken languages in the world in order to build parallel corpora, includes Russian data, as well.

The [General Internet-Corpus of Russian \(GICR\)](#) is a megacorpus with more than 15 GT created with a fully automated technology of collecting and tagging texts from Russian Internet and based on the latest achievements of computational linguistics.

There is a corpus building project called [HANCO at the Department of Slavonic and Baltic Languages and Literatures at the University of Helsinki](#). The aim of the research group is to build a corpus that contains c. 100, 000 running words, extracted from a modern Russian magazine and representing the modern Russian language.

10.5.4 News portals

A wide range of broadcast and print outlets represent the media in Russia. In total, there are 27,000 newspapers and magazines and 330 television channels in the country. The media spheres are ruled by the combination of private and state-ownership, and their political neutrality is questionable.

Three main radio stations are available in Russia, namely:

- [Radio Russia](#) whose coverage is 96.9% of the population
- [Radio Mayak](#) with 92.4% coverage
- [Radio Yunost](#) (51.0% coverage)

The radio stations broadcast music, in particular.

The most popular source of information is TV. There are 330 television channels in total from which only three channels have a nationwide outreach: the [First Channel](#), the [Rossiya](#) and the [NTV](#). Regional television is relatively popular in Russia. There is an English-language channel, i.e. [Russia Today](#). The only independent TV channel is [Dozhd](#) (Rain) that reports on corruption and human rights abuses.

Local and national newspapers are the second most popular choice, while the Internet comes third. Russians are strong users of social networks, of which [Odnoklassniki.ru](#) and [VKontakte](#) are the most popular. Additionally, Russian Internet resources provide Russian translations of the world press (e.g. [InoSmi](#), [SMI2](#), [Perevodika](#), etc.).

10.6 Computational tools

Some computational tools developed for Russian are presented in this section.

10.6.1 Language identification

The [CLD2](#) provides full support for Russian. There is a language identifier of [USB](#) as well. In addition, the following language identifiers are available online: [TextCat](#), [Polyglot3000](#), [LabsTranslated](#), and [saffsd/langid](#).

Additionally, the tools used for automatic transliteration is collected in section [10.3](#).

10.6.2 Tokenizer

Tools for tokenization as well as POS-tagging, lemmatizing and paraphrasing a word to word for Russian are provided by [NLP-Tools](#). There is also a tokenizer for the Russian language developed by [OpenCorpora](#). Furthermore, a tool developed by [Semantic Analyzer – Language intelligence](#) provides language support, i.e. tokenizer and lemmatizer, for Russian.

10.6.3 Stemmer

There is a stemmer developed for the Russian language by [Snowball](#) generated for C# and Java. This stemmer has been made adoptable for C# by the [Iveonik Systems](#) (for further information see [StemmersNet](#)).

10.6.4 Spell checker

The open source spell checker of [HunSpell](#) for Russian is available. In addition, there are free online spell checkers for Russian, such as the [translit](#) that provides transliteration and spell checker; the [Online Spell Checker for Russian](#); and the [stars21](#).

The [WebSpellChecker](#) is a tool which not only highlights misspelled words but also gives suggestions. Amongst many languages, it supports Russian, however the license is 250\$ for a year (but there is a 30 day trial version).

10.6.5 Phrase level and higher tools

The tools developed by [Giellatekno](#) provide resources for Russian, such as text analysis, paradigm generation, word generation, number word generation, and dictionaries.

Russian morphological analyzer Open source text tokenization, sentence splitting, morphological analysis, etc. are offered by the [FreeLing 3.1](#).

Russian part-of-speech tagger A language independent POS tagger called [TreeTagger](#) has been used, among other languages, for Russian.

Russian chunker The [TreeTagger](#) can also be used as chunker for Russian.

Russian named entity recognizer The [AlchemyAPI](#) supports Named Entity Extraction in Russian.

Russian sentence parser [Nivre et al. \(2008\)](#) present the first results on parsing the SYNTAGRUS treebank of Russian with a data-driven dependency parser ([PDF](#)). There is also a [Russian Sentence Parser available via github](#).

Russian speech recognizer The [LingvoSoft](#) provides speech recognition software for Russian. Additionally, the [Google Cloud Speech API](#) supports Russian.

Russian machine translator The online translator by [META](#) supports language pairs including Russian (e.g. Ukrainian, Russian, English, German, Latvian, French, etc.). Additionally, the [google translator](#) supports the Russian language (for a detailed description and the supported languages see the [Wikipedia entry of Google Translate](#)). Furthermore, the rule-based machine translator of [SDL](#) supports Russian. There is also a [machine traslator provided by Yandex, Microsoft Translator, SYSTRAN](#).

Russian question answering machine There is an [English–Russian question answering system](#).

10.6.6 End-user support

The following OS supports are available for Russian:

- a [MAC OS x](#) language pack is available for Russian;
- an [Ubuntu](#) language pack is available for Russian;
- a [Microsoft Windows](#) language pack is available in Russian;

The Unicode character range for the Cyrillic script is 0400 — 04FF (and there is an additional Cyrillic supplementary code set: 0500 — 052F).

There are online resources, which convert texts written in the Cyrillic script in Latin and vice versa automatically. Some of them are listed here: the [translit](#); the [Lexilogos](#); the [Russian Transliterate Tool](#); the [transliteration](#); the [Rusklaviatura](#); the [Lexicool](#).

The UCLA provides downloadable phonetic [keyboards](#) for PC and Apple as well as a virtual one.

The [Babel package](#) for TEX and LaTeX as well as the [polyglossia](#) for XeTeX support the Russian language and (Cyrillic) writing system.

Furthermore, several optical character recognition, i.e. OCR, softwares, such as the [ABBYY FineReader](#), the [i2OCR](#), the [Free OCR](#), etc. extract Russian texts too.

Finally, there are Russian courses via Duolingo available in different development phases: the [Russian for English speakers Duolinuo course](#) is in Incubator Phase 3 (the course is graduated from Beta); and the [Russian for Turkish speakers Duolingo course](#) is in incubation Phase 1 (the course is almost ready, but it is not yet released).

Bibliography

Erich Berneker and Max Vasmer. *Russische Grammatik*. Walter de Gruyter, Berlin, 1971.

Thomas R. Beyer. Russian. In Jane Garry and Carl Rubino, editors, *Facts About the World's Languages, An Encyclopedia of the World's Languages: Past and Present*, pages 605–607. HW Wilson, New York / Dublin, 2001.

F. M. Borras and R. F. Christian. *Russian Syntax: Aspects of Modern Russian Syntax and Vocabulary*. Clarendon Press, Oxford, 1959.

Greville G. Corbett. *Gender*. Cambridge University Press, Cambridge, 1991.

Norman M. Fraser and Greville G. Corbett. Gender, animacy and declensional class assignment: a unified account for russian. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1994*, pages 123–150. Kluwer, Dordrecht, 1995.

Colin Howlett. *Oxford Russian dictionary*. Oxford University Press, Oxford, 1996.

Maria Koptjevska-Tamm. Action nominal constructions. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

Joakim Nivre, Igor M Boguslavsky, and Leonid L Iomdin. Parsing the syntagrus treebank of russian. In *Proceedings of the 22nd International Conference on Computational Linguistics- Volume 1*, pages 641–648. Association for Computational Linguistics, 2008.

I. Pulkina and E. Zakhava-Nekrasova. *Russian: A practical grammar with excercises*. Russky Yazyk Publishing House, Moscow, 1974.

I. M. Pulkina. *A short Russian reference grammar*. Russian Language Publications, Moscow, 1978.

Galina Stilman and William E. Harkins. *Introductory Russian Grammar*. Blaisdell, Waltham, Massachusetts, 1964.

E. Tauscher and E.-G. Kirschbaum. *Grammatik der russischen Sprache*. Volk und Wissen, Berlin, 1963.

B. O. Unbegaun. *Russian Grammar*. University Press, Oxford, 1957.

Terence L. B. Wade. *A Comprehensive Russian Grammar*. Blackwell, Oxford, 1992. reprinted in 1995.

Chapter 11

Somali (Levente Madarász)

Contents

11.1 Demography and ethnography	153
11.2 Main typological and syntactic features	158
11.3 Writing system, transcription	160
11.4 Previous research on the language	160
11.5 Data and sources	161
11.6 Computational tools	162
Bibliography	164

Introduction

This chapter aims at providing an overview of the Somali language, mainly spoken in Somalia and various other countries on the Horn of Africa. Following a brief overview of the demography and ethnography of Somali's speaking community, the reader will be acquainted with the typological classification of the language. After an account on the writing system, previous researches and data sources will be presented. The closing section will detail the available computational tools for the language.

11.1 Demography and ethnography

11.1.1 Name variants

The Somali endonym is *Soomaali* (or *Aṣ-Šūmāl* (Zagórski, 2010, pg. 43)) stemming from the name of the oldest common ancestor of several Somali clans, Irir Samaale (Lewis, 1999, pg. 11). The endonym is also used to refer to the language itself: *af Soomaali*, but Somali is also known as *Af-Maxaad Tiri, Darod, Isa, Isaq, Sab, Soomaaliga* and *Somali-Aweer* (Paul et al., 2015; Hammarström et al., 2015). Somali is identified with the *so* ISO 639-1, *som* ISO 639-2 and ISO 639-3 codes (SIL International, 2015).

11.1.2 Geographic spread

As shown on Figure 11.1, outside Somalia (where Somali is a national, EGIDS level 1 lingua; see Figure 11.2), the language is spoken in Ethiopia (EGIDS level 2, provincial language spoken mainly in



Figure 11.1: Map of Somali speaking areas ([Wikipedia, 2015](#)). A map with better resolution is made by [Huffman \(2015\)](#).

the Somali and Oromia regions, as well as, areas in the Afar and Dire Dawa regions), Kenya (EGIDS level 5, dispersed language spoken in Garissa, Mandera, and Wajir counties, by the eastern border, and in northwest Luma county) and Djibouti (EGIDS level 5, dispersed language) ([Paul et al., 2015](#)). Further, a significant number of speakers (> 10,000) can also be found in Yemen, the United Arab Emirates, the United States, Saudi Arabia, the United Kingdom, Tanzania, Sweden, Canada, Norway and Denmark ([Joshua Project, 2015](#)).

11.1.3 Speaker populations

Due to various armed conflicts taking place in Somalia since 1991, the exact number of speakers is hard to estimate. According to [Paul et al. \(2015\)](#), the total population of speakers in all countries of its diaspora is 14,762,900, out of whom 6,460,000 resides in Somalia. According to the 2007 census,

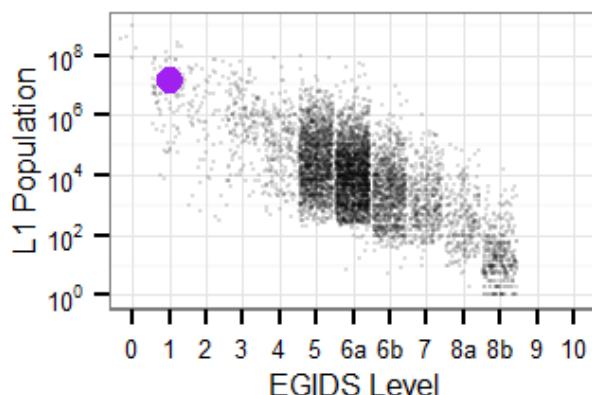


Figure 11.2: The EGIDS level for Somali spoken in Somalia (source: [Ethnologue](#)). For more details about the plot, please see Section X.1 Demography and ethnography.

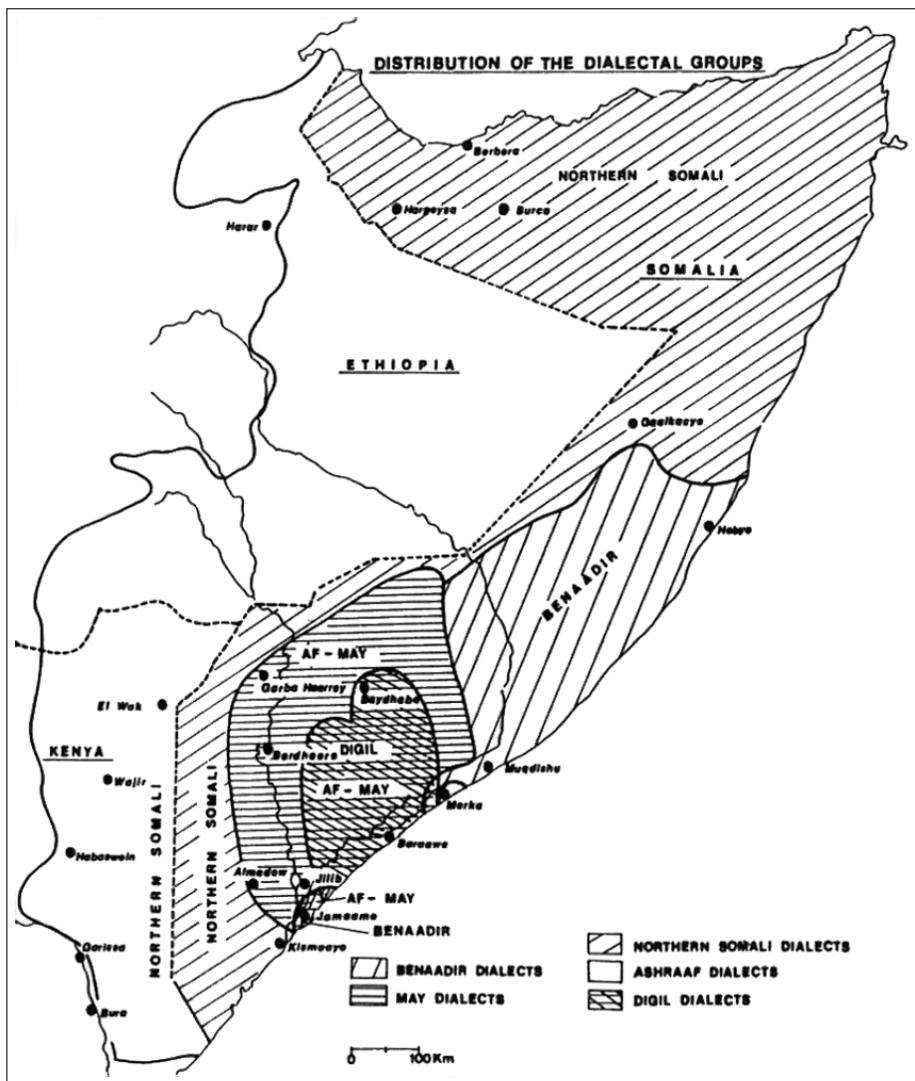


Figure 11.3: Map of Somali dialects reproduced from Lamberti (1986).

Ethiopia is the home of 4,610,000 Somali speakers (cited in Paul et al. 2015), out of whom 2,880,000 is monolingual and 95,600 speakers use it as a second language—the major clan families are the *Daarood*, the *Ogaadeen*, the *Dir*, the *Gadabuursi*, the *Hawiye*, and the *Isxaag*. Kenya host 2,386,000 Somali speakers (census data from 2009, cited in Paul et al. 2015), which includes 58,200 *Hawiye* or *Hawiyah*, 516,000 *Degodia* and 622,000 *Ogaden* clan members. Kenya also gives home to the members of the *Daarood* and *Dir* clan families. There are 297,000 Somali speakers in Djibouti (census data from 2006, cited in Paul et al. 2015) belonging to three clan families: the *Issa*, the *Gadaboursi* and the *Issaq*.

Drawing on data complied by various Christian denominations with the purpose of quantifying the presence of gospel among various ethno-linguistic groups, the Joshua Project (2015) reports different numbers. For a comparison of these figures with that of Paul et al. (2015), see Table 11.1.

11.1.4 Dialect situation

The last systematic survey of Somali dialects was carried out during the 1980s by Lamberti. The tragic events taking place in Somalia since the post-1980s era provides us with ample reasons to suspect that the available data might not be the most up to date (Kapchits, 2009; LandInfo, 2011).

country	EGIDS	Joshua Project 2015				Paul et al. 2015			
		L1 #	L2 #	L1 %	L2 %	L1 #	L2 #	L1 %	L2 %
Somalia	1	8,354,100	2,267,000	77.45%	21.02%	6,460,000	N/A	59.89%	N/A
Ethiopia	2	6,501,000	10,000	6.54%	0.01%	4,514,400	95,600	4.54%	0.10%
Kenya	5	2,943,200	861,000	6.39%	1.87%	2,386,000	N/A	5.18%	N/A
Yemen	N/A	1,041,000	N/A	3.88%	N/A	N/A	N/A	N/A	N/A
Djibouti	5	358,000	N/A	40.32%	N/A	297,000	N/A	33.45%	N/A
UAE	N/A	162,000	N/A	1.77%	N/A	N/A	N/A	N/A	N/A
USA	N/A	80,000	N/A	0.02%	N/A	N/A	N/A	N/A	N/A
Saudi Arabia	N/A	63,000	N/A	1.54%	N/A	N/A	N/A	N/A	N/A
Tanzania	N/A	54,000	N/A	0.10%	N/A	N/A	N/A	N/A	N/A
UK	N/A	54,000	N/A	0.08%	N/A	N/A	N/A	N/A	N/A
Sweden	N/A	43,000	N/A	0.44%	N/A	N/A	N/A	N/A	N/A
Canada	N/A	37,000	N/A	0.10%	N/A	N/A	N/A	N/A	N/A
Norway	N/A	36,000	N/A	0.69%	N/A	N/A	N/A	N/A	N/A
Denmark	N/A	10,000	N/A	0.18%	N/A	N/A	N/A	N/A	N/A
Finland	N/A	7,400	N/A	0.13%	N/A	N/A	N/A	N/A	N/A
Eritrea	N/A	5,900	N/A	0.42%	N/A	N/A	N/A	N/A	N/A
Italy	N/A	5,200	N/A	0.01%	N/A	N/A	N/A	N/A	N/A
Netherlands	N/A	2,900	N/A	0.02%	N/A	N/A	N/A	N/A	N/A
Sudan	N/A	1,700	N/A	0.00%	N/A	N/A	N/A	N/A	N/A
New Zealand	N/A	1,600	N/A	0.03%	N/A	N/A	N/A	N/A	N/A
Belgium	N/A	1,400	N/A	0.01%	N/A	N/A	N/A	N/A	N/A
Malta	N/A	1,000	N/A	0.23%	N/A	N/A	N/A	N/A	N/A
world		19,763,400	3,138,000	0.27%	0.04%	13,657,400	95,600	0.19%	0.00%

Table 11.1: Summary of Somali speaker populations, based on data obtained from the [Joshua Project \(2015\)](#) and the Somali Ethnologue entry ([Paul et al., 2015](#)). L1 and L2 values followed by a hash or percentage sign indicate the count and population ratio figures corresponding to the primary and secondary language users in a given country. Population ratios were calculated by relying on the measures of the [World Bank \(2015\)](#).

Based on phonological, syntactic and morphological characteristics, [Lamberti \(1986\)](#) identified five main dialectal groups: *Northern Somali*, *Benadir*, *Ashafar*, *May* and *Digil* (see Figure 11.3).

Northern Somali group

The naming of the group is somewhat misleading, as it not only refers to the language variant spoken in the northern parts of Somalia, but it embraces variants spoken in the southernmost parts of the country, as well as in territories belonging to Kenya and Ethiopia [LandInfo \(2011\)](#). The group consists of four subgroups: the subgroup called *Northern Somali* (whose speakers inhabit the Galbeed, Togdheer and Sanaag administrative regions), the subgroup called *Darood* (speakers residing in the Gedo and Bakool regions and in the Mudug, Nugal and Bari administrative districts of Somalia, as well as, in the Ogaden province of Ethiopia), *Lower Juba* (spoken in the Lower Juba administrative region) and a less consistent subgroup in northern Kenya. Northern Somali as a subgroup subsumes the *Issa*, *Gadabursi* and *Issaq* dialects. The Darood group consists of the *Majerteen*, *Dhulbahante*, *Marehan*, *Degodiya*, *Warsangeli* and *Ogaden* dialects. While Lower Juba includes the *Af-Majerteen*, *Af-Marehan*, *Af-Degodiya* and *Af-Ogaden* dialects. The Kenyan dialectal subgroup said to include the Odagen dialects called Af-Abudwaq, Afabdallah, Harti and Degodiya.

As summarized in ([LandInfo, 2011](#)), features of the Northern Somali dialects are presenting themselves on multiple levels. On the phonemic level, the key feature of the group is the presence of pharyngeal phonemes, the phonemic distinction between /q/ and /kh/ and the absence of nasals. On the morphological level, Northern Somali speakers do not combine the /-ayaa/ and /-aysaa/ progressive suffixes, the partial duplication phenomenon applies to masculine monosyllabic words in the majority of cases, there are no singular suffixes and the definite article is postfixed. For more nuanced distinctions, please refer to [LandInfo \(2011\)](#).

Benadir group

The group consists of five subgroups: *Abgal* (spoken in the Middle Shebelle region and in Mogadishu, the capital of Somalia), *Galjeel* (spoken in the southern part of the Hiraan administrative region, the northern parts of the Lower Shebelle region and in certain districts of Middle Juba), *Ajuran* (spoken in the Hiraan region and in the Dinsor, Sakow and Bu'ale districts), *Hamar* (spoken in the Hamer Weyne district) and *Biimal* (used in Mogadishu, the most southern part of the Afgoye and Qoryoley districts, and in the Merka and Jamaame districts) ([LandInfo, 2011](#)).

The Benadir dialects are generally less consistent, providing greater grammatical freedom as compared with the dialects of the Northern Somali group. The Abgal and Biimal variants of Somali have the most speakers.

Ashraf group

The Ashraf dialects are said to be used exclusively by the members of the Ashraf clan. It consists of two variants: the *Af-Shingani* variant (spoken in the Shangani district of Mogadishu) and the subgroup called *Af-Merka* (comprising of dialects spoken in the Lower Shebella district: *Af-Merka*, *Af-Gendershe* and *Af-Jilib*).

Reported distinctions in these dialects include divergent realization of intervocalic dentals (/t/~/d/) and consonants realized either as palato-alveolar or palatal (e.g., /ʃ/~/j/ in *isha* and *iyad*) ([LandInfo, 2011](#)).

May group

Lamberti only describes the *Rahanweyn* and *Rahanweyn-esque* dialects. Speakers of May inhabit the Bakool region, the south-western Hiraan region, the Jowhar district in the Middle Shebelle region, the entire Bay region, Gedo, Middle Juba, Lower Shebelle, the eastern part of Lower Juba, as well as the Merka district in Lower Shebelle. May form a dialectal continuum consisting of five variants: norther May (spoken in northern Gedo and Bakool), western May (used in Bardheere and Dinsor), eastern May (consisting of *Af-Elay* and *Af-Begedi*), Qoryoley/Jilib May and southern May (used in the Jamaame and Afmadow district) ([LandInfo, 2011](#)).

May dialects display distinguishing phonological and morphological features: they use the phoneme /ə/, lack certain pharyngeal phonemes, certain phoneme clusters display assimilation (/l/ and /t/ realizing as /ll/, /h/ and /t/ as /tt/, /h/ and /n/ becomes /nn/). On the morphological level, in May dialects, plurals are grammatically masculine, while most possessive and indicative pronouns are formed with the prefix /haan(i)-/ ([LandInfo, 2011](#)).

Digil group

The group comprises of the *Tunni* dialects (*Af-Tunni Defaraat* and *Af-Tunni Torre*—spoken in the Dhinsor, Brava and Jilib districts), the *Dabarre* dialects (*Af-Dabarre* and *Af-Oroole*—spoken primarily in the Dhinsor and Qansax Dheere districts, and along the Juba River in Middle Shebelle), the *Garre* dialects (forming a continuum with many variants—spoken in the Baidoa, Dhinsor, Bur Hakaba and Qoryoley districts) and the *Jiddu* dialects (spoken in the Qoryoley, Dhinsor, and Jilib districts).

Lamberti maintains (cited in ([LandInfo, 2011](#))) that the Digil group is the most homogeneous. As mentioned, the Garre dialect forms a continuum, which is visible from features like the preserved pre-fixed conjugations in the Bur Hakaba variant, missing from the Baidoa variant; in the same regional version the plural form /-te/, common to Garre dialects, is replaced with /-yaal/.

To sum up, *Northern Somali* can be considered as the standard version of the language; it is intelligible to *Benaadir* Somali speakers, but difficult or unintelligible to most *Maay* speakers, or the Merka and Muqdisho users of the *Af-Ashraaf* variant ([Paul et al., 2015](#)). Although not explicitly argued above, the name of the dialectal groups often coincide with the name of clan families, hinting that the emergence of main variants had not only been conditioned by spatial factors, but social ones as well.

11.2 Main typological and syntactic features

Somali is an Afro-Asiatic language, a member of the Cushitic subfamily ([Hammarström et al., 2015](#)). Despite the close religious, as well, as commercial relations with the Arab world, the Somali speaking community remained relatively intact; there are many Arabic loans in Somali, but general knowledge of Arabic is a function of ones level of education and proximity to the northern coasts ([Saeed, 1999a](#)).

11.2.1 Linguistic typology

Phonological level. As regards voicing in stops and fricatives, Standard Somali only maintains a partial voicing contrast, between /t/~/d/ and /k/~/g/. Somali is a tone language with three possible tones: High, Low and falling (High to Low) ([Saeed, 1999a](#)). Vowel harmony is also present, and vowel

length is contrastive. Somali employs a (C)V(C)(C) syllable structure (where parentheses indicate optional elements) (Project, 2006).

Morphological level. Somali is an agglutinating language. Affixation is strictly suffixal (Project, 2006; Dryer and Haspelmath, 2013). ¹

Morphosyntactic level. As summarized by the colleagues of the UCLA Language Materials Project (2006), verbs are inflected for person (first, second, third), number (singular, plural), gender (masculine, feminine), mood (indicative, imperative, subjunctive, conditional), aspect (perfect, imperfect), tense (present, past, future), and polarity (affirmative, negative); while nouns are inflected for definiteness (definite, indefinite), gender (masculine, feminine), number (singular, plural) and case (nominative, accusative, genitive, vocative).

Syntactic level. The basic word order in Somali is SOV, but the order of constituents can undergo changes as the language is discourse configurational (guided by pragmatic principles, reflecting the information structure of the utterance) (Saeed, 1999a). Below the four basic sentence patterns are described.

11.2.2 Predication

Verbless declarative sentences with the declarative marker (DM) *waa* follow the pattern below (Saeed, 1999a):

- (56) *Cali waa bare.*
 Ali DM teacher
 ‘Ali is a teacher.’

Simple predicative sentence structure (with focus on the object) follows the basic SCSOV pattern (focused element followed by focus marker) (Gutman and Avanzati, 2013):

- (57) *Nin-kii libaax-ii ay-uu dilay.*
 man-DEF.*subj* [lion-DEF.*obj* PRT-3SG.] (FOC) killed
 ‘The man killed the lion.’

11.2.3 Possession

Possessive in Somali is always attached to the end of definite nouns. The possessive suffix always precedes definite markers and agrees in gender with the noun it is attached to (Ali, 2010):

- (58) *Magac-ayg-u waa Ashkir.*
 name-POSS.1ST.SG.-SUBJ.*subj* DM Ashkir
 ‘My name is Ashkir.’

¹See the sections on the **Digil group** and the **May group** above for potential exceptions.

11.2.4 Imperative

Imperatives have dedicated verb forms distinguishing singular and plural addressees (Saeed, 1999a). The simple imperative sentence below targets several addressees:

- (59) *Ii sameeyaa.*
me+for do-IMP(PL)

‘Do it for me.’

11.2.5 Interrogative

To question nouns, Somali uses interrogative determiners (*-kée* (masculine)~*-tée* (feminine)) (Saeed, 1999a):

- (60) *nin-kée/naag-tée*
man-ID/woman-ID
‘Which man/woman?’

When verbal focus is used, yes-no questions are formed by replacing the declarative marker *waa* with the question particle *ma*:

- (61)
- | | |
|------------------------------------|--------------------------------------|
| Aabbahaa waa-n la hadlay? | Aabbahaa ma la hadlay? |
| father-your DM with spoke | father-your QPRT with spoke |
| ‘I spoke with your father’ | ‘Did I/he speak with your father?’ |

For an exhaustive overview of the Somali sentence patterns, please refer to Saeed (1999a).

11.3 Writing system, transcription

Contemporary Somalis use a variant of the Latin script (Somali Latin Alphabet), introduced in 1972 (Paul et al., 2015). The variant is customized to the peculiarities of the language (e.g., it lacks a letter denoting /p/).

To a marginal extent, Somalis also use a variant of the Arabic script (*Wadaad’s writing*) and the so-called *Osmanya* (or Somali) alphabet (situated in the Unicode range 10480–104AF, see Figure 11.4).

11.4 Previous research on the language

Somali is a language with a relatively wide-ranging linguistic coverage. Perhaps the most comprehensive reference grammar of the language was first published in 1999 by Saeed. *Somali* formulates analytic and theoretical assumptions about the language and presents them as a concise description of the lingua intended for academic audience.

Critiques of the book however cite inadequate coverage of recent findings; for an exhaustive literary review (covering the whole field up to late 2014) please refer to Nilsson (2014). Since 2014, to the

Figure 11.4: The *Osmanya* (or Somali) alphabet (Ager, 2015).

author's knowledge, no significant work has been published (except for (Goldberg, 2015; Saidat and Alenazy, 2015)). For a bibliography which provides verbal description for the entries, please refer to Green et al. (2014).

11.5 Data and sources

11.5.1 Basic vocabulary

The *An Crúbadán* project offers a package of Somali resources (such as character trigrams, word bigrams and word frequency tables) compiled from 13,490 documents, with a total of 24,648,526 words.

11.5.2 Dictionaries

Paper-based Somali dictionaries: (Ahmad, 2012; Omer, 2010; Hashi and Hashi, 1998, 1995; Korshel, 1994; Hashi and Hashi, 1993; Zorc and Osman, 1993; Farah, 1992; Zorc et al., 1991; Ohly, 1987; Luling, 1987; Castagno, 1975; Abraham, 1966; Spitler and Spitler, 1966; Abraham, 1964; De Larajasse, 1897)

Online Somali dictionaries: ([Redsea-online.com](#), 2015; Broz, 2015; Figueiredo, 2014; Jama, 2007; [WikiQaamuus](#), 2015)

11.5.3 Corpora

Monolingual

As of 11/22/2015, the [Somali Wikipedia](#) consisted of 3,638 content pages. The [Helsinki Corpus of Somali](#) is a written corpus of Somali comprising of 6,430 words extracted from running texts. The [BBC corpus](#) contains radio broadcasts in 33 languages, including Somali. [Indigenous Tweets](#) offers informal text messages ('tweets') gathered from 558 Somali Twitter users. Although the project is currently in a testing phase, the [Somali Corpus](#) will offer 3 million tagged words extracted from texts belonging to

10 different registers (e.g., literary texts, poetry and songs, newswriting, etc.). For collecting Somali texts, a further approach might be the crawling of Somali websites, identified with the .so top-level domain.

Bilingual

The Bible and the Quran are ideal candidates for bilingual corpus building: the Bible is made available in .html format by the [Wordproject®](#), while the Quran is readily available aligned with various languages in .xml format on the site of the [OPUS corpus project](#). Besides the sacred texts, the Universal Declaration of Human Rights is also available in .txt format on the web page of the [Unicode Consortium](#). Further, [Ubuntu](#) and [GNOME](#) localization files are also available aligned with different languages in .xml format on the site of the OPUS corpus project.

11.6 Computational tools

11.6.1 Language identification

The following open-source language detectors are capable of identifying Somali language: ([Shuyo](#), 2010; [OpeNER](#), 2015; [Hayden and Riesa](#), 2013). Non-open source tools include ([DataSift](#), 2012; [Translated.net](#), 2007) and ([Technology](#), 2015).

11.6.2 Tokenizer

The tool of [Burns](#) and [Al-Rfou](#) both support Somali language ([Burns](#), 2013; [Al-Rfou](#), 2015).

11.6.3 Stemmer

For the stemming of Somali words, one can rely on an (X)FST-based morphological analyzer implemented for Somali by [Ryan Joahson](#).

11.6.4 Spell checker

An open-source [HunSpell](#) checker is downloadable for Somali, and an [OpenOffice](#) extension is also available. Besides these, the Red Sea On Line Cultural Foundation also provides an [online spell checker](#).

11.6.5 Phrase level and higher tools

This section presents a collection of articles dealing with higher-level computational tools for Somali:

- **Somali part-of-speech tagger:** [Jama](#) (2016) (in his paper, Jama describes a basic POS tagger)
- **Somali named entity recognizer:** As of 2015-11-22, NERs are not available for Somali.
- **Somali chunker:** As of 2015-11-22, chunkers are not available for Somali.
- **Somali morphological parser:** [Joahson](#) (2012)
- **Somali sentence parser:** [Joahson](#) (2015)

- **Somali question answering system:** As of 2015-11-22, no QASs are available for Somali language.
- **Somali speech recognizer:** [Abdillahi et al. \(2006\)](#)
- **Somali machine translator:** [Google \(2006\)](#)

11.6.6 End-user support

- **Mac OSX** (as of 2015-11-22): No OS-level support
- **Microsoft Windows** (as of 2015-11-22): No language pack available
- **Ubuntu** (as of 2015-11-22): [GNOME translation updates](#) available

Bibliography

- Nimaan Abdillahi, Pascal Nocera, and Jean-François Bonastre. Towards automatic transcription of Somali language. In *les actes de Language Resource and Evaluation Conference (LREC)*, 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/801_pdf.pdf.
- R. C. Abraham. *Somali-English Dictionary*. University of London Press, 1964.
- R. C. Abraham. *Somali-English Dictionary*. University of London Press, 1966.
- S. Ager. Omniglot - writing systems and languages of the world, 2015. URL www.omniglot.com.
- L. A. Ahmad. *A Dictionary of Somali Verbs in Everyday Contexts*. AuthorHouse UK, 2012. ISBN 9781467881388. URL <https://books.google.hu/books?id=rK3qzbQBFVUC>.
- R. Al-Rfou. polyglot, 2015. URL <http://polyglot.readthedocs.org/en/latest/index.html#>.
- A. H. Ali. The possessives, 2010. URL http://hooyo.web.free.fr/E_chap06.html.
- Zdenek Broz. English to Somali dictionary. Ingiriisi – Soomaali qaamuus, 2015. URL <http://www.dicts.info/dictionary.php?l1=English&l2=Somali>.
- P. R. Burns. Morphadorner v2.0, 2013. URL <https://devadorner.northwestern.edu/maserver/>.
- M. Castagno. *Historical dictionary of Somalia*, volume 6. Scarecrow Pr, 1975.
- DataSift. Language detection, 2012. URL <http://app.datasift.com/source/20/language-detection>.
- E. De Larajasse. *Somali-English and English-Somali Dictionary*. K. Paul, Trench, Trübner & Company, 1897.
- M. S. Dryer and M. Haspelmath, editors. *Language Somali*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL http://wals.info/languoid/lect/wals_code_som.
- J. A. Farah. *Somali learner's dictionary*. Haan Publishing, 1992.
- R. B. Figueiredo. Somali-English-Somali online dictionary, 2014. URL <http://www.freelang.net/online/somali.php>.
- J. Goldberg. *A Brief Grammar of the Af-Somali Language*. SIL Ethiopia, 2015.
- Google. Google translate, 2006. URL <http://translate.google.com/>.
- C. Green, M. E. Morrison, N. B. Adams, E. S. Crabb, E. Jones, and V. Novak. Reference and pedagogical resources for ‘Standard’ Somali. *Electronic Journal of African Bibliography*, 15, 2014. URL <http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1017&context=ejab>.

- A. Gutman and B. Avanzati. Somali, 2013. URL <http://www.languagesgulper.com/eng/Somali.html>.
- H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 2.6 – Somali, 2015. URL <http://glottolog.org/resource/languoid/id/soma1255>. [Online; accessed 21-November-2015].
- A. A. Hashi and A. A. Hashi. *Essential English-Somali Dictionary*. Fiqi Pr Inc, 1993.
- A. A. Hashi and A. A. Hashi. *Fiqi's Somali English Dictionary*. Fiqi Educational Materials, 1995.
- A. A. Hashi and A. A. Hashi. *Essential Somali English Dictionary*. Fiqi Press, 1998.
- A. Hayden and J. Riesa. Compact language detector 2, 2013. URL <https://code.google.com/p/cld2/wiki/CLD2FullVersion>.
- S. Huffman. Languages of the Horn of Africa, 2015. URL http://www.worldgeodatasets.com/index.php/download_file/view/1538/. [Online; accessed 21-November-2015].
- M. J. Jama. Somali-English Italian online mathematical dictionary, 2007. URL http://www.dmunipi.it/~jama/alif/qaamuus/qaamuus_en.html.
- M. J. Jama. Somali corpus: state of the art, and tools for linguistic analysis, 2016. manuscript.
- R. Joahnson. Somali morphological analyser, 2012. URL <http://giellatekno.uit.no/doc/lang/som/SomaliDocumentation.html>.
- R. Joahnson. Grammatical analyser for Somali, 2015. URL <http://giellatekno.uit.no/doc/lang/som/SomaliDocumentation.html>.
- Joshua Project. Language - Somali, 2015. URL <http://goo.gl/zYQVaD>. [Online; accessed 21-November-2015].
- G. Kapchits. Lambert's maps of Somali dialects and the current socio-linguistic situation in South Somalia. Retreived from the author's website, 2009. URL kapchits.narod.ru/Lamberti-s_maps_of_dialects.doc.
- M. Korshel. *English-Somali, Somali-English Dictionary: Ingirisi Soomaali Qaamuus, Soomaali Ingirisi*. Star Publications Limited, 1994.
- M. Lamberti. *Map of Somali Dialects in the Somali Democratic Republic*. H. Buske, 1986.
- LandInfo. Somalia. language situation and dialects, 2011. URL http://www.landinfo.no/asset/1800/1/1800_1.pdf.
- I. M. Lewis. *A Pastoral Democracy: a study of pastoralism and politics among the northern Somalis of the Horn of Africa*. James Currey Publishers, 1999.
- V. Luling. *Somali-English Dictionary*. Dunwoody Press, 1987.
- M. Nilsson. A bibliography of Somali language and linguistics, 2014. URL <http://morgannilsson.se/pdf/so.bib.pdf>.
- R. Ohly. *Swahili-English slang pocket dictionary: 1500 words and phrases*, volume 31. Afro-Pub, 1987.

- A. M. Omer. *English-Somali & Somali-English One-to-one Dictionary*. One to one bilingual dictionary. IBS Books, 2010. ISBN 9781905863945. URL <https://books.google.hu/books?id=TSQ1KQEACAAJ>.
- OpeNER. Language identifier, 2015. URL <https://github.com/opener-project/language-identifier>.
- L. M. Paul, G. F. Simons, and D. F. Charles. Somali - ethnologue, 2015. URL <http://www.ethnologue.com/language/som>. [Online; accessed 21-November-2015].
- UCLA Language Materials Project. Somali, 2006. URL <http://www.lmp.ucla.edu/Profile.aspx?LangID=202&menu=004>.
- Redsea-online.com. Somali-English-Italian trilingual online dictionary, 2015. URL <http://www.redsea-online.com/modules.php?name=dictionary>.
- J. I. Saeed. *Somali*, volume 10 of *London Oriental and Africal Language Library*. John Benjamins Publishing, 1999a.
- J. I. Saeed. *Somali*. John Benjamins, 1999b.
- Ahmad M Saidat and Mamdouh A Alenazy. Thematic roles in Somali: A principles and parameters approach. *Advances in Language and Literary Studies*, 6(5):104–110, 2015.
- N. Shuyo. Language detection library for java, 2010. URL <https://github.com/shuyo/language-detection>.
- SIL International. Iso 639 code tables, 2015. URL http://www-01.sil.org/iso639-3/codes.asp?order=639_3. [Online; accessed 21-November-2015].
- A. K. Spitzer and H. Spitzer. *English-Somali Dictionary*. Pasadena, California: World-Wide Missions, 1966.
- Basis Technology. Rosette® language identifier (rli), 2015. URL <http://www.basistech.com/text-analytics/rosette/language-identifier/>.
- Translated.net. Automatic online language identifier, 2007. URL <http://labs.translated.net/language-identifier/>.
- Wikipedia. Languages of Somalia — Wikipedia, the free encyclopedia, 2015. URL https://en.wikipedia.org/w/index.php?title=Languages_of_Somalia&oldid=680178051. [Online; accessed 21-November-2015].
- WikiQaamuus. Wikiqaamuus, 2015. URL <https://so.wiktionary.org/>.
- World Bank. Population 2015, 2015. URL <http://databank.worldbank.org/data/download/POP.pdf>. [Online; accessed 10-October-2016].
- B. Zagórski. *Endonyms versus Exonyms: A Case Study in Standardization. With a List of Names of Arab Countries*. UNGNEG, 2010. URL http://ksng.gugik.gov.pl/pliki/endonyms_vs_exonyms_en.pdf.

- R. D. P. Zorc and M. M. Osman. *Somali-English dictionary with English index*. Dunwoody Press, 1993.
- R. D. P. Zorc, V. Luling, and M. M. Osman. *Somali-English dictionary*. Dunwoody Press, 1991.

Chapter 12

Spanish (Zsuzsanna Bárkányi)

Contents

12.1 Demography and ethnography	169
12.2 Main typological and syntactic features	172
12.3 Writing system, transcription	175
12.4 Previous research on the language	175
12.5 Data and sources	176
12.6 Computational tools	180
Bibliography	182

Introduction

The present chapter aims at providing a brief overview of the Spanish language (*spa* ISO 639-3). The first part will present the reader with the historical roots and the dialectal extension of the language, which is followed by a non-exhaustive survey of the reference grammars and the different available corpora and various digital resources. The closing section will list some of the available computational tools for the language.

12.1 Demography and ethnography

12.1.1 Origins

Spanish (in Spanish *castellano* or *español*) is an Indo-European West Romance language. Modern Spanish is basically a modern form of Latin as the language is the result of an uninterrupted evolution of spoken Latin which was brought to the Iberian Peninsula by the Romans during the Second Punic War (210 BC). Several pre-Roman languages might have influenced the evolution of Spanish the most important of which was Basque still spoken today in northern Spain.

After the collapse of the Roman Empire and the subsequent loss of political unity in the 5th century, Vulgar Latin fragmented into local dialects which produced a dialectal continuum in the north of the peninsula from the Catalan speaking territories in the east to the Portuguese speaking territories in the west. In the south, however, we do not find such a dialectal continuum as this territory was under Arab control at various times and with constantly changing boundaries between 711 and 1492 AD. As a result of this cohabitation, Spanish borrowed hundreds of words from Arabic. Corominas (1954) lists

over a thousand entries of Arabic origin. This historical and linguistic division between the northern and southern parts of the Iberian Peninsula is reflected in the dialectal division of present day Spanish too (see section 12.1.3).

Spanish colonization from the beginning of the 16th century took the language to the Americas, first to Central America and then gradually to South America where it is still spoken today. Interestingly, Amerindian languages had little influence on the grammar of American Spanish, but they gave a significant number of lexical items to Spanish and through Spanish to other European languages (e.g., *chocolate, jaguar*).

12.1.2 Geographic spread

Spanish is one of the most spoken languages today. It is estimated that around 400 million people speak Spanish as L1, which makes it the second language in the world (after Mandarin) by number of native speakers. Nowadays it is spoken in four continents: in America it is the official language in nineteen countries (Mexico, Guatemala, El Salvador, Honduras, Nicaragua, Costa Rica, Panama, Cuba, Dominican Republic, Colombia, Venezuela, Ecuador, Peru, Bolivia, Chile, Paraguay, Uruguay, Argentina and Puerto Rico) and is also spoken in the southern states of the United States. In Europe it is the official language of Spain. In Africa it is the official language in Equatorial Guinea and the North African cities of Ceuta and Melilla. In Asia Spanish is spoken as a minority language in Israel and the Philippines. The number of L2 users of Spanish worldwide is almost 90 million.

12.1.3 Dialect situation

While all Spanish dialects use the same written standard, spoken varieties differ from one another to varying degrees mostly in pronunciation and vocabulary and less so in grammar. Spanish can be divided into three major dialectal blocks (Piñeros, 2008): (i) Central-Northern Peninsular Spanish, (ii) terrabajense ('low lands Spanish') and (iii) terraltense ('high lands Spanish'). (i) comprises the central and northern part of Spain. This is considered the most conservative dialect of Spanish and is distinguished from the other dialects by the use of an apico-alveolar /s/ and the phonemic distinction between /s/ and /θ/, as well as the use of the pronoun *vosotros* for 2nd person plural informal. (ii) includes the southern half of Spain, the Canary Islands and the coastal areas of Latin America. The most prominent feature of this variety is the massive lenition of coda consonants. (iii) encompasses the remaining parts of the Americas. In these areas coda /s/ is generally preserved but there is no phonological distinction between orthographic 's' vs. 'z'/'ce', 'ci'.

A more detailed geographical division is made by Hualde et al. (2010) which is in accordance with the aforementioned division. They divide Spain into three major dialectal zones: (i) Central–Northern Peninsular Spanish, (ii) Andalusian and (iii) Canary. In Spanish America they speak about six major dialectal zones: (i) Caribbean (which includes the Spanish speaking Caribbean islands and the coastal regions of the surrounding countries), (ii) Central American (Mexico and the highlands of Central America), (iii) Andean (includes the highlands of Venezuela, Colombia, Ecuador, Peru, Bolivia and the north of Argentina), (iv) Chilean, (v) Paraguayan, (vi) Argentinian and Uruguayan.

12.1.4 Spanish in contact with other languages

As Spanish is spoken throughout an immense territory, in several regions it is in everyday contact with other languages and also has a significant number of bilingual speakers. In all these areas there is a degree of diglossia between the two languages which necessarily has an impact on the Spanish

country	# L1	EGIDS level	L1 ratio	# L2
Mexico	103 M	1	84.22	7.08 M
Colombia	41 M	1	84.85	70.5 K
Argentina	38.8 M	1	93.61	1 M
Spain	38.4 M	1	82.10	7.49 M
USA	34.2 M	2	10.72	15 M
Venezuela	26.3 M	1	86.48	632 K
Peru	24 M	1	79	2.06 M
Chile	15 M	1	85.13	1.55 M
Ecuador	13.5 M	1	85.77	742 K
Cuba	11.2 M	1	99.38	
Dominican Republic	8.51 M	1	81.83	52.6 K
Guatemala	7.27 M	1	46.99	2.44 K
El Salvador	6.09 M	1	96.06	
Honduras	5.9 M	1	72.86	112 K
Nicaragua	5.31 M	1	87.34	577 K
Bolivia	4.14 M	1	38.80	4.59 M
Costa Rica	4.05 M	1	83.13	81.7 K
Puerto Rico	3.54 M	1	99.70	137 K
Uruguay	3.17 M	1	93.04	75.3 K
Panama	2.55 M	1	65.99	463 K
France	444 K	3	0.67	8.4 M
Paraguay	365 K	1	5.37	4 M
Belize	174 K	3	52.43	22 K
Andorra	27.6 K	3	34.84	21.6 K
Gibraltar	14.3 K	3	47.67	K
Caribbean Netherlands	10.7 K	3	50.63	115 K
Morocco	6.6 K	7	0.02	3.41 M
Curacao	5.7 K	5	3.71	K
Trinidad and Tobago	4.1 K	5	0.31	61.8 K
Philippines	2.66 K	5	0.00	2.56 M
world	392.2 M			63.37 M

Table 12.1: Countries with the most L1 and L2 speakers of Spanish. Columns show the number of L1 speakers, the EGIDS level of Spanish, ratio of L1 speakers to the population of the country in percent, and the absolute number of L2 speakers.

of the region. In Spain the language that has the biggest impact on Spanish in terms of the number of speakers is definitely Catalan. The contact between Spanish and Galician also dates back to old times. Galician is a linguistic variety closely linked to Portuguese. The contact between Spanish and Basque is very special since this is the only living language that was spoken on the Peninsula before the Roman conquest. These two languages have been in contact from the formation of Spanish. One of the earliest examples written in Castilian Romance from the beginnings of the 11th century (*Glosas Emilianenses*) are marginalia written in a Latin codex by a bilingual monk. The same codex contains sentences written in Medieval Basque.

The contact between Spanish speaking communities and indigenous communities in America has been changing over time and region. The indigenous languages that have the largest number of speakers and have a considerable impact on Spanish spoken in the bilingual regions are: Quechua, Maya and Guarani. Quechua is spoken by 8 to 12 million people from the south of Colombia to northeast Argentina including Ecuador, Peru and Bolivia. Maya is the second biggest Amerindian language spoken by appr. 6 million people in the southern part of Mexico and Guatemala. Guarani is spoken by 5 million people mostly in Paraguay where 89% of the population is bilingual, and is also spoken in the surrounding areas of Brazil, Argentina and Bolivia. Other important languages are Aymara, Nahuatl and Mapundungu.

12.2 Main typological and syntactic features

Spanish is generally claimed to be a Subject–Verb–Object language with common variations in word order. We think (following Gutiérrez Bravo 2007) that word order in Spanish is determined by semantic features rather than syntactic functions. The unmarked word order is based on semantic roles which have the following hierarchy: Agent » Experiencer » Theme/Patient » Location. The highest NP occupies the pre-verbal position, while the remaining NPs are post-verbal. Only one NP can be pre-verbal and NPs lower in the scale than Experiencer cannot be pre-verbal. This explains why in sentences with a transitive verb and an Agent Subject 62 the unmarked word order is indeed S–V–O, while in sentences with Theme Subject like 63 the unmarked word order is V–S.

- (62) *Juan come una manzana*
 Juan eat.3RD.SG.PRES.IND. art-SG.FEM.INDEF. apple

‘John eats an apple.’

- (63) *Fracasaron las negociaciones*
 fail.3ST.PL.PAST.IND. article-PL.FEM.DEF. negotiations

‘The negotiations failed.’

Spanish is a pro-drop language, that is, subject pronouns are omitted when they are pragmatically unnecessary. It is a verb-framed language meaning that the direction of motion is expressed in the verb while the manner of motion is expressed adverbially, e.g. *salir corriendo* ‘to run out’ literally ‘go out running’.

Spanish syntax is right-branching, that is, modifying constituents tend to be placed after their head words. Adjectives in the unmarked case are placed after nouns. Pre-nominal adjectives have an expressive evaluative value. The language uses prepositions rather than postpositions or nominal

inflection. Spanish has a richly inflected verbal morphology, verbs have about fifty conjugated forms. Nominal inflection, on the other hand, is rather limited: nouns, adjectives and determiners have two numbers and two genders.

Spanish is a non-tonal language with contrastive lexical word stress which always falls on one of the last three syllables of the word (with the exception of certain verbal forms where it can fall on the 4th syllable from the right). The language has a strong tendency to open syllables. According to Quilis (1993) about 45% of Spanish syllables is open (in those varieties that do preserve coda consonants). It is very rare to have coda clusters, the second member of these clusters is always /s/ (or in Northern-Central Spanish /θ/).

12.2.1 Predication

Non-verbal predication consists of a verb (generally a copulative verb *ser* or *estar* or a pseudo-copulative verb like *quedar* ‘turn out’) and a predicate nominal. Since Spanish is a pro-drop language the pronominal subject can be omitted.

- (64) *Es abogado*
 be.1ST.SG.PRES.IND. lawyer.SG.MASC.
 ‘(He) is a lawyer.’

- (65) *Está cansada*
 be.1ST.SG.PRES.IND. tired.SG.FEM.
 ‘(She) is tired.’

The difference between the two copulative verbs in Spanish is widely studied. In a nutshell, in the first case⁶⁴, our example speaks of a permanent, essential characteristics of a person, while the second case ⁶⁵ refers to a condition that can easily change. The same subject and the same nominal predicate joined by the two copulas acquire a different meaning. In ⁶⁶ we speak of the essential characteristics of the apple: the apple is green in color and remains so even after it has ripened. In the second sentence ⁶⁷ the apple is green because it has not yet ripened. When the condition of the apple changes, it will no longer be green.

- (66) *La manzana es verde*
 ‘The apple is green.’ (essential characteristics, type of apple)

- (67) *La manzana está verde*
 ‘The apple is green.’ (state that might change)

12.2.2 Possession

As mentioned above, Spanish is right-branching, so in a possessive construction the possessor appears after the possessee, e.g., *la casa de la madre* ‘the house of the mother’.

- (68) *La casa de la madre de mi abuela*
 the house of the mother of my grandmother
 ‘My grandmother’s mother’s house.’

Possessive pronouns in Spanish have an unstressed (pre-nominal) clitic form and a long, stressed (post-nominal) form. Pre-nominal possessives are determinants which cannot co-occur with articles **la mi casa* ‘the my house’. Traditional grammars regard pre-nominal possessives as possessive adjectives. They agree in number with the possessee *mi coche* ‘my car’ vs. *mis coches* ‘my cars’. In first and second person plural long and short possessive pronouns are identical and in these cases the pronoun agrees with the possession in gender as well *nuestra casa* ‘our house’ vs. *nuestro coche* ‘our car’.

Post-nominal possessive pronouns can serve as pro-forms and replace the whole possessive construction. They agree both in gender and in number with the possession. This stressed form is normally used with the definite article.

- (69) *Tu cuarto está en el segundo piso y el mío.*
 your room is in the.SG.MASC. secondSG.MASC. floor and the.SG.MASC. mine.SG.MASC.
en el primero
 in the.SG.MASC. first.SG.MASC.

‘Your room is on the second floor and mine is on the first.’

A post-nominal possessive pronoun can have an expressive value especially in a copulative sentence 70. The long form possessive pronoun can be used in reference to an indefinite number of people or things 70 or to denote one member of a group 71.

- (70) *Amigos nuestros han visto la película*
 friends our.PL.MASC. have seen the film
 ‘Our friends have seen the film’

- (71) *Una hija mía ganó el concurso*
 a daughter my.SG.FEM. won the competition
 ‘A daughter of mine won the competition.’

12.2.3 Imperative

Formal commands (both affirmative and negative) use the present subjunctive verb form: *hable* ‘speak.2ND.SG.FORMAL.IMPERATIVE’. Affirmative 2nd person singular informal commands use the 3rd person singular present indicative form: *habla* ‘speak.2ND.SG.INFORMAL.IMPERATIVE’. Negative informal commands also use the present subjunctive form of the verb: *no hables* ‘don’t speak’ ‘speak.2ND.SG.INFORMAL.IMPERATIVE’. Affirmative 2nd person plural informal commands use the 2nd person singular informal command form + the suffix *d*: *hablad* ‘speak.2ND.PL.INFORMAL.IMPERATIVE’.

Accusative and Dative clitics are attached directly to the imperative form of the verb in affirmative commands.

- (72) *Cómpra+me+lo*

buy2ND.SG.INFORMAL.IMPERATIVE+pronoun-1ST.SG.DAT.+pronoun-3RD.SG.ACC.MASC.

‘Buy me that!’

- (73) *No me lo compres*

negation pronoun-1ST.SG.DAT. pronoun-3RD.SG.ACC.MASC. buy2ND.SG.PRES.SUBJ

‘Do not buy me that!’

12.2.4 Interrogative

Wh-questions require Subject–Verb inversion.

- (74) *¿Dónde vive Pepe?*

where live.3RD.SG.PRES.IND Pepe

‘Where does Pepe live?’

In Yes–or–No questions there might or might not be Subject–Verb inversion in Spanish. The word order without inversion is more common, in these cases the recognition of the speaker’s intention relies on intonation. Note that there is a considerable variation among Spanish dialects with regard to intonation.

- (75) *¿Es usted español?*

be3RD.PRES.IND. you.FORMAL Spanish.SG.MASC

‘Are you Spanish?’

- (76) *¿Usted es español?*

you.FORMAL be.3RD.PRES.IND. Spanish.SG.MASC

‘Are you Spanish?’

12.3 Writing system, transcription

Spanish uses Latin script with a few language-specific amendments. Word stress in the non-default case is marked by an acute accent, /n/ is signaled by a tilde over ‘n’ (ñ). Spanish uses upper quotation marks (“ . . . ”) and inverted question/exclamation marks (¿ and ¡) at the beginning of interrogations and exclamations, respectively. Braille script is also adapted for Spanish.

12.4 Previous research on the language

12.4.1 Linguistic norm

The reign of Alfonso X the Wise (1252–84) meant the first step in the formation of the linguistic norm of Castilian. Alfonso X promoted the production of books in a wide range of topics including literature,

history, law and astronomy. The orthography set by Alfonso X was used till the 16th century. The first grammar of Spanish – and of a modern European language for that matter – was Antonio Nebrija's *Grammatica* published in 1492 ([Nebrija, 1492](#)).

Finally, with the foundation of the *Royal Spanish Academy* (Real Academia Española) in the 18th century, and the publication of grammar books and dictionaries by the Academy, written Spanish gained a uniform linguistic norm. Today the *Royal Spanish Academy* is affiliated with national language academies in twenty-one other Spanish-speaking nations through the Association of Spanish Language Academies. Their aim is to promote linguistic unity within and between the various hispanophone territories.

As far as the language experts are concerned, Zsuzsanna Bárkányi has done research on Spanish at the Research Institute for Linguistics of the Hungarian Academy of Sciences.

12.4.2 Descriptive and prescriptive grammars

Spanish has been a well-studied language for centuries. There are hundreds of grammar books of different sizes and levels. The most recent publication of the Royal Spanish Academy is *La nueva gramática de la lengua española* (2009-2011). This is a comprehensive consensual work by all Spanish language academies published in three volumes which are dedicated to morphology, syntax, and phonetics and phonology.

There is also a significant number of descriptive works dealing with different aspects and different geographical and social varieties of Spanish. *Ethnologue* lists over 4 thousand entries as OLAC resources for Spanish.

12.4.3 Spanish as L2

The Instituto Cervantes, a non-profit organization, was founded by the Government of Spain in 1991 to promote Spanish language teaching as well as that of Spain's co-official languages, in addition to fostering knowledge of the cultures of Spanish-speaking countries throughout the world. Since Spanish has almost 90 million L2 users worldwide, there is a vast amount of language support material for teachers and learners of Spanish. We only mention a few as examples.

- [Instituto Cervantes](#)
- [Online Library of the Cervantes Institute](#)
- [todo ele](#)
- [zona ele](#)
- [Useful links](#) gathered by Joaquim Llisterri from Universitat Autònoma de Barcelona.

12.5 Data and sources

The *An Crúbadán* project provides Spanish resources:

- [spa resources](#) - 1272 documents, 4,485,061 words
- [spa \(Argentina\) resources](#) - 132 documents, 304,045 words
- [spa \(Bolivia\) resources](#) - 106 documents, 140,191 words

- [*spa* \(Chile\) recourses](#) - 59 documents, 75,129 words
- [*spa* \(Colombia\) recourses](#) - 72 documents, 70,033 words
- [*spa* \(Costa Rica\) recourses](#) - 131 documents, 153,466 words
- [*spa* \(Cuba\) recourses](#) - 19 documents, 22,342 words
- [*spa* \(Ecuador\) recourses](#) - 61 documents, 45,207 words
- [*spa* \(El Salvador\) recourses](#) - 22 documents, 25,587 words
- [*spa* \(Guatemala\) recourses](#) - 36 documents, 158,648 words
- [*spa* \(Honduras\) recourses](#) - 146 documents, 182,629 words
- [*spa* \(Mexico\) recourses](#) - 1101 documents, 2,843,129 words
- [*spa* \(Nicaragua\) recourses](#) - 1844 documents, 3,297,262 words
- [*spa* \(Panama\) recourses](#) - 33 documents, 26,394 words
- [*spa* \(Paraguay\) recourses](#) - 968 documents, 1,189,109 words
- [*spa* \(Peru\) recourses](#) - 1133 documents, 3,260,665 words
- [*spa* \(Puerto Rico\) recourses](#) - 842 documents, 912,031 words
- [*spa* \(Spain\) recourses](#) - 135 documents, 513,131 words
- [*spa* \(Uruguay\) recourses](#) - 681 documents, 706,152 words
- [*spa* \(Venezuela\) recourses](#) - 197 documents, 2,804,744 words

Furthermore, a [Lego](#) Spanish word list is also available.

12.5.1 Dictionaries

This section gives an overview of the Spanish dictionaries (both paper and online ones).

There are hundreds of Spanish bilingual and monolingual dictionaries of various sizes. The most widely used online dictionaries of Spanish are the following.

Diccionario de la lengua española DRAE published by the Royal Spanish Academy since 1780. The dictionary has had a free electronic version since 2001 and the online version of the 23rd (most recent) edition is available since October 2015. The search possibilities are expanded in this edition and it also allows navigation within the dictionary. The dictionary contains 93,111 entries with a total of 195,439 meanings.

Diccionario panhispánico de dudas DPD, the 1st printed edition of 2005 is available online. The goal of this dictionary is to answer questions regarding language use (spelling, lexical and grammatical concerns) from the point of view of the current norm.

[WordReference.com](#) is ranked in the top 100 most visited websites in Spain and all of Latin America. It provides free online bilingual and monolingual dictionaries: English–Spanish, Spanish–English, French–Spanish, Spanish–French, Portuguese–Spanish, Spanish–Portuguese, Spanish: definitions, Spanish: synonyms. This online dictionary offers forum discussions with the search words and thus offers examples of usage and captures up to date meanings and connotations.

There are several online etymological dictionaries of Spanish as well. They seek to present the evolution of the Spanish lexicon over time. An example is *El nuevo diccionario histórico del español CNDHE*.

Some more online dictionaries are listed below:

- [bab.la](#) Spanish–English, Spanish–German, Spanish–French, Spanish–Italian, and Spanish–Portuguese bilingual dictionaries (scrapeable)
- [Collins English–Spanish](#)
- [Oxford English–Spanish](#)
- [Cambridge English–Spanish](#)
- [Larousse Enlish–Spanish](#)

Additionally, [Google Translate](#) also supports Spanish (including a text-to-speech module). There are numerous applications for Spanish that make use of ASR: e.g., self-service systems, speech-to-text dictation software. These systems take into account the dialectal variations of Spanish and generally differentiate between Peninsular Spanish, Caribbean/Colombian Spanish, Argentinian Spanish and USA Spanish.

12.5.2 Corpora

We will give a non-exhaustive list of some of the most important written and oral corpora of Spanish that are freely searchable online.

CREA

Corpus de Referencia del Español Actual ([CREA](#)) has nearly 140,000 documents and more than 154 million word forms from texts coming from all Spanish-speaking countries. The documents were produced between 1975 and 2004. 49% of the material comes from books, another 49% from journals and newspapers and 2% is miscellaneous material. CREA has an oral subcorpus, the CREA oral.

CREA oral

Nearly 9 million word forms from transcripts of the spoken language, with over 1600 documents can be accessed. The material was obtained through various institutions as well as direct recording from the Internet and then orthographically transcribed and encoded. 50% of the oral material comes from Spain and the other 50% from America. This latter is distributed according to the traditional language areas: Andean, Caribbean, Coastal Caribbean, Chile, the United States, Mexico/Central America, and the River Plate (Argentina, Uruguay).

CORPES XXI

The Royal Spanish Academy published a new corpus of the 21st century: [CORPES XXI](#). Similarly to CREA, it is also a corpus of reference. It is a semi-open corpus meaning that 25 million new word forms are supposed to be added each year, 70% coming from Latin America and 30% from Spain. 90% of the texts correspond to written language and 10% to oral language. The latest version also allows the recovery of a sound aligned with the oral texts and it can also be searched based on grammatical category.

Corpus Valesco

The [Corpus Valesco](#) is a corpus of spoken Spanish containing transcriptions of 46 conversations divided into interventions, intonational units and words. This corpus is searchable in English too.

CORLEC

Corpus Oral de Referencia de la Lengua Espaⁿola Contemporánea ([CORLEC](#)) is a corpus of spoken Spanish prepared by a research group on the Universidad Autónoma de Madrid. It contains the transcript of 1,100,000 recorded words from texts belonging to the colloquial register.

PRESEEA

[PRESEEA](#) is a Project for the Sociolinguistic Study of Spanish from Spain and America. The goal is to coordinate sociolinguistic researchers from Spain and Hispanic America in order to make possible comparisons between different studies and materials, as well as a basic information exchange. The main aim is to create a spoken language Corpus with sufficient guarantees and rich in terms of linguistic information. The project has a free ever growing corpus.

Atlas interactivo de la entonación del español

The [Interactive Atlas of Spanish Intonation](#) provides a database of a series of audio and video materials for the study of prosody and intonation of the dialects of Spanish. It contains the following sentence types which can be compared across different dialects. All the sentence types contain a marked and an unmarked subtype.

- declarative sentences
- Yes-or-No questions
- Wh-questions
- Echo questions
- Requests and Orders
- Vocative forms

Wikipedia has 1,213,044 articles in Spanish (on November 16, 2015).

A diachronic corpus of Spanish is the *Corpus Diacrónico del Espaⁿol* [CORDE](#) which contains 236,709,914 words since the beginnings of the language till 1974.

Spanish is also represented in the [OpenSubtitles2016](#) (191,987 files, 1.3G tokens, 179.2M sentences).

12.5.3 News portals

Information about newspapers published in Spain is available at [w3newspapers](#). Among the great number of Spanish online newspapers, only a couple of them will be mentioned:

- [El País](#) – the highest-circulation daily newspaper in Spain and one of three Madrid dailies considered to be national newspapers of record for Spain (along with [El Mundo](#) and [ABC](#))
- [Boletín Oficial del Estado](#) – official information from the Spanish Government
- [Marca](#) and [AS](#) – sport newspapers
- [Cinco Días](#) and [Expansion](#) – business and finance newspapers

According to [The Statistics Portal](#), the sport focused paper, *Marca*, had the greatest readership with well over two million readers. It was followed by *El País* and *As*, another sport based paper. The data originated from a survey carried out between April 2014 and March 2015.

Information about radio stations and TV channels in Spain is available at [Spain Live Radio Stations](#) and [TV channels from Spain](#), respectively (with links).

[Wikipedia](#) provides a list of the newspapers published in Mexico as well.

12.6 Computational tools

12.6.1 Language identification

Several language identification tools are successful in identifying Spanish. The Compact Language Detector 2, the stand-alone [saffsd/langid.py](#), the Rosette Language Identifier as well as the TextCat and the Guess Language language guessers. Polyglot 3000 also quickly recognizes Spanish and the NRC System for Discriminating Similar Languages ([Goutte et al., 2014](#)) is also very successful with Spanish.

12.6.2 Tokenizer

The Test Word Tokenizer supports Spanish although the only information it takes into account is white space. The Stanford NLP has a Spanish package, and Ivory also supports tokenization in Spanish. The Apache Lucene Core provides Java-based indexing and search technology, as well as spellchecking, hit highlighting and advanced analysis/tokenization capabilities supporting Spanish.

12.6.3 Stemmer

Snowball is a language in which stemmers can be exactly defined, and from which fast stemmer programs in ANSI C or Java can be generated. It is available for all major Romance languages including Spanish.

12.6.4 Spell checker

There is a number of online spell checkers available for Spanish – we will list a few as examples – and word processing softwares are also equipped with a package for Spanish.

- [Spanish checker](#)
- [jspell](#)
- A spell checker for [Mexican Spanish](#)
- [Stilus](#) can be used besides checking spelling and grammar for refining the style of the text too.

12.6.5 Phrase level and higher tools

In this section we mention some of the most well-known tools for Spanish without aiming to provide an exhaustive list of such tools.

Verb conjugators

The internet abounds with grammar support for the Spanish language. We can find a number of verb conjugators like [Spanish Verb Conjugations](#), [Verbix](#) or [Lingolex](#), just to mention a few.

Morphological analyzers

The SMM developed by [Mahlow and Piotrowski \(2009\)](#) is a morphological analyzer for Spanish which allows to analyze Spanish word forms concerning inflection, derivation, and compounding, it is able to deal with cliticized forms too.

The [SpanMorph](#) is an open-source morphological analyzer for Spanish.

The COES developed by Rodríguez and Carretero (1994) is a complete environment which allows the user to deal with Spanish morphological problems such as formal specification of Spanish morphology, word tagging and dictionary generation.

MAHT is a morphological analyzer for the Spanish language that is mainly based on the storage of words and its morphological information, leading to a lexical knowledge base that has almost five million words.

Taggers

The Stanford NLP is compatible with POS tagging in Spanish (via NLTK in Python).

The Spaghetti tagger is just a simple recipe for Spanish POS tagging using the CESS corpus (Martí et al., 2007) with NLTK's implementation of bigram and unigram taggers. (For NLTK see Bird et al. 2009.)

The Tree Tagger is a language independent part-of-speech tool for annotating text with POS and lemma information developed at the University of Stuttgart. Tree Tagger is successfully applied for Spanish.

The Petra POS Tagger is a Spanish tagger written in C++ that assigns a POS tag to each token of a given sentence. This tagger has the special feature that it is prepared to tag bilingual texts, enhancing the precision of the tag process.

SEPE developed by Jiménez and Morales (2002) is a part-of-speech tagging system specially designed to tag Spanish texts using small linguistic resources.

There is ongoing research to adapt Priberam's question answering system to Spanish (Amaral et al., 2008).

Sentence parsers

The Stanford NLP has a Stanford Parser module which supports Spanish.

The MaltParser developed at the Universitat Pompeu Fabra is trained for Spanish using the IULA Spanish LSP Treebank, which contains more than 42,000 sentences and almost 590,000 tokens.

ASR

As for Automatic Speech Recognition, Spanish is among the best investigated languages in this respect. Voxeo is available for several varieties of Spanish (Mexico, Spain, Argentina, Chile and Colombia) with a spectrum of applications. Google recently launched OK Google for Spanish.

12.6.6 Computer support

Spanish is supported under any operation system (Mac OSX, Microsoft Windows, LINUX/UNIX).

Bibliography

Carlos Amaral, Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, José Pina, and Cláudia Pinto. Priberam’s question answering system in qa@clef 2008. *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, pages 337–344, 2008.

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

Joan Corominas. *Diccionario crítico etimológico de la lengua castellana*. Madrid: Gredos, 1954.

Cyril Goutte, Serge Leger, and Marine Carpuat. Varieties and dialects. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages*, page 139–145, August 2014.

Rodrigo Gutiérrez Bravo. Prominence scales and unmarked word order in spanish. *Natural Language and Linguistic Theory*, 25:235–271, 2007.

J. I. Hualde, A. Olarrea, A. M. Escobar, and C. E. Travis. *Introducción a la lingüística española*. Cambridge University Press, 2010.

Héctor Jiménez and Guillermo Morales. Sepe: A spanish texts pos tagger. *Computational Linguistics and Intelligent Text Processing*, 2276:250–259, 2002.

Cerstin Mahlow and Michael Piotrowski. Smm: Detailed, structured morphological analysis for spanish. *Journal Polibits (Computer science and computer engineering with applications)*, 39:4–48, 2009.

Antonia Martí, Mariona Taulé, Lluís Márquez, and Manuel Bertran. Cess-ece: A multilingual and multilevel annotated corpus. www.lsi.upc.edu/~mbertran/cess-ece/publications, 2007.

Antonio Nebrija. *Gramática castellana*. Salamanca, 1492.

C. Eduardo Piñeros. *Estructura de los sonidos del español*. Upper Saddle River, New Jersey: Pearson Prentice Hall, 2008.

Antonio Quilis. *Tratado de fonética y fonología españolas*. Madrid: Gredos. Biblioteca Románica, 1993.

Santiago Rodríguez and Jesús Carretero. A formal approach to spanish morphology: the coes tools. <http://www.datsi.fi.upm.es/~coes/publications/sepln.pdf>, 1994.

Chapter 13

Swahili (Levente Madarász)

Contents

13.1 Demography and ethnography	185
13.2 Main typological and syntactic features	190
13.3 Writing system, transcription	192
13.4 Previous research on the language	192
13.5 Data and sources	193
13.6 Computational tools	196
Bibliography	198

Introduction

The present chapter aims at providing an overview of Swahili, a Bantu language, spoken mainly in Tanzania. The first part will present the reader with the historical roots and the dialectal variations of the language, which is followed by a round-up on the writing system, the reference grammars and the different available corpora. The closing section will detail the various digital resources and the available computational tools for the language.

13.1 Demography and ethnography

13.1.1 Name variants

The endonym of Swahili speaking people stems from the Arabic word *sawahili*, literally meaning “coast-dwellers” (Harper, 2016).

As Radulesk (2008) points out, there is no clear historical evidence about the ethnic origins of Swahili, nevertheless it is often argued that the history of the language is related to the Arabs and the Persians who moved to the East African coast. Besides Swahili, the language is also known as *Kiswahili*, *Kisuaheli*, *Kiswaheli*, *Suahili*, *Kisuahili* and *Arab-Swahili* (Hammarström et al., 2015), identified with the *sw* ISO 639-1 and *swa* ISO 639-2 and ISO 639-3 codes (Paul et al., 2015).

It must also be noted, though not detailed in this chapter, that Swahili is also a macrolanguage with two branches, the respective ISO 639-3 codes of which are the following: *swc* (Congo Swahili) and *swh* (Coastal Swahili). For more details, visit the respective [Ethnologue](#) site.

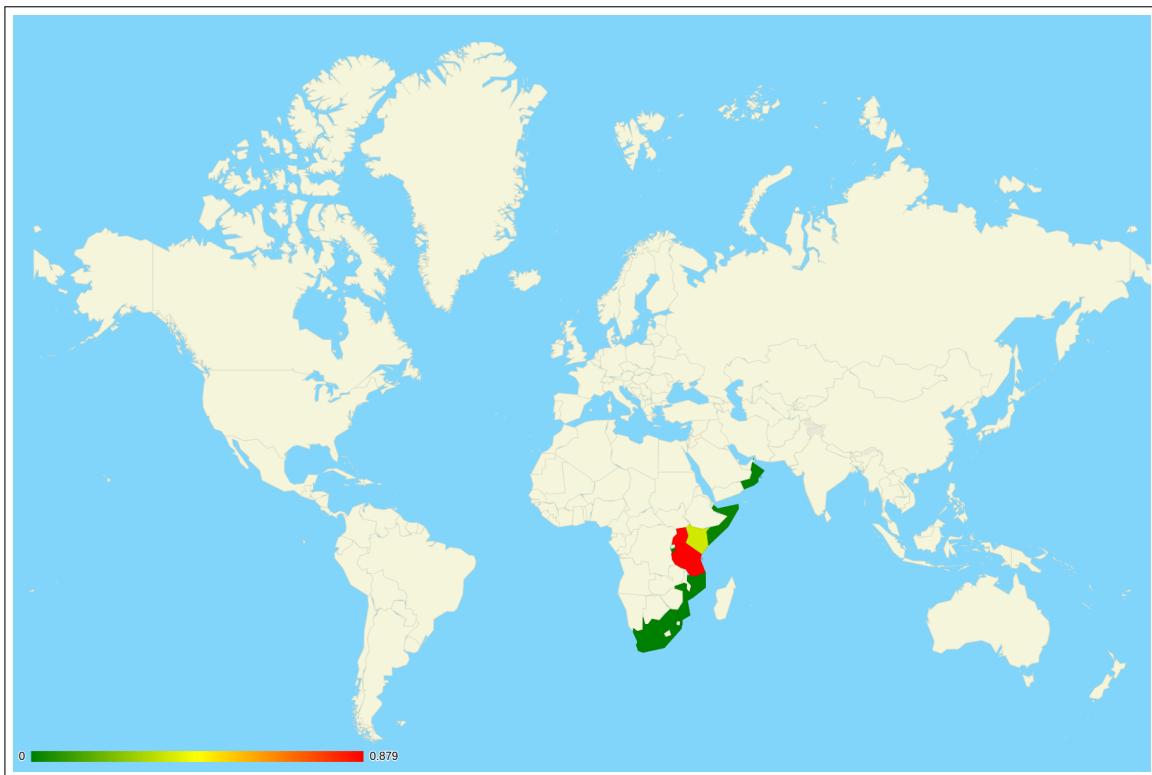


Figure 13.1: Map of Swahili speaking areas (Paul et al., 2016). Coloring indicates relative salience of the language in a given country; near 0% is indicated by green, while near 100% spread by red, in gray countries no Swahili speaker communities are attested.

13.1.2 Geographic spread

Figure 13.1 displays Swahili speaking areas around the world. Swahili is mainly spoken in Tanzania where it is a *de facto* national, EGIDS level 1 language. Significant speaker populations also reside in Uganda (EGIDS level 1, statutory national working language), Kenya (EGIDS level 1, statutory national language) and in Burundi (EGIDS level 3 language, used for wider communication, primarily, in Muslim contexts and for commerce). Smaller speaking populations can be found in Somalia, Oman, Mozambique and in South Africa; the level of the language in these countries is uniformly EGIDS 5 (dispersed) (Paul et al., 2015).

13.1.3 Speaker populations

While Swahili is only spoken as a first language by approximately 15,000,000 people, it is the official lingua franca of the [East African Community \(1999\)](#). As such, to some extent it is spoken in every member state (i.e., Republics of Kenya, Uganda, the United Republic of Tanzania, Republic of Burundi and Republic of Rwanda). Based on the data gathered by Paul et al. (2016), Table 13.1 summarizes the available official figures of Swahili speakers.

In addition, based on data collected from the International Mission Board of the Southern Baptist convention, the [Joshua Project \(2016\)](#) cites further speaker populations that are summarized in Table 13.2.

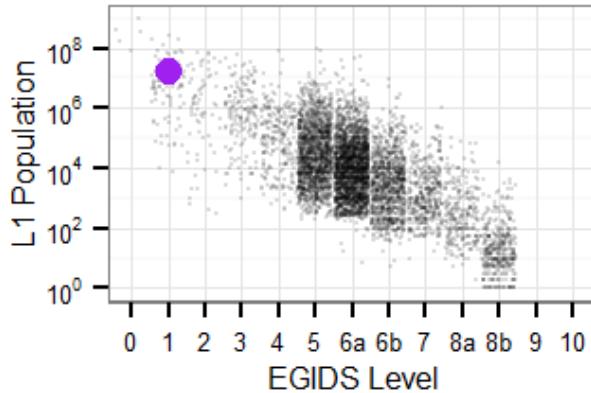


Figure 13.2: The EGIDS level for Swahili spoken in Tanzania (source: [Ethnologue](#)). For more details about the plot, please see Section [X.1 Demography and ethnography](#).

Country	Primary	Secondary	Grand Total	Population ratio
Tanzania	15,000,000	32,000,000	47,000,000	87.90%
Uganda	313,000	33,800,000	34,113,000	87.40%
Kenya	111,000	16,500,000	16,611,000	36.07%
Somalia	NA	NA	164,000	1.52%
Oman	NA	NA	22,000	0.49%
Mozambique	NA	NA	10,000	0.04%
Burundi	NA	NA	6,360	0.06%
South Africa	NA	NA	2,000	0.00%
world	15,424,000	82,300,000	97,928,360	1.33%

Table 13.1: Summary of Swahili speaker populations, based on data obtained by [Paul et al. \(2016\)](#). Population ratios were calculated by relying on the measures of the [World Bank \(2015\)](#).

13.1.4 Dialect situation

Swahili can be divided into two main variant groups: coastal dialects and inland variants. Since Swahili evolved from the former group, most of its L1 users speak one of the island variants. Inland Swahili variants, on the other hand, are mostly used by L2 speaker groups and usually embrace a substratum from the L1 of the respective group ([Polomé, 1967](#), p. 19). Further, while differences between the members of the coastal group result from geographic distance (dialectal variation), inland variation can best be explained by social factors (sociolectal variation). (Unless indicated otherwise, the entries in this subsection are based on [Polomé 1967](#).)

Coastal variants

Bravanese Variant spoken by the inhabitants of Barawa (or Brava) in the Lower Shebelle region of Somalia. The most characteristic feature of the dialect is the tone utilized in verb inflections ([Goodman referenced in Polomé, 1967](#)). The endonym of the variant is Cimiini, but it is also known as Mwini, Mwiini, Chimwiini, Af-Chimwiini and Barwaani ([Hammarström et al., 2015](#)).

Country	Speakers	Population ratio
Saudi Arabia	381,000	1.21%
Canada	329,000	0.92%
Germany	240,000	0.29%
Italy	125,000	0.21%
United States	59,000	0.02%
Democratic Republic of Congo	48,000	0.06%
Zambia	42,000	0.26%
Reunion	25,000	2.96%
Sudan	23,000	0.06%
Rwanda	17,000	0.15%
Mayotte	8,000	3.76%
Greece	6,100	0.06%
Madagascar	5,100	0.02%
Malawi	3,500	0.02%
Libya	2,200	0.04%
New Zealand	1,500	0.03%
Comoros	1,100	0.14%
Denmark	1,000	0.02%

Table 13.2: Summary of speaker communities based on non-governmental data gathered by the [Joshua Project \(2016\)](#). Population ratios were calculated by relying on the measures of the [World Bank \(2015\)](#).

Bajuni Variant spoken at the southern coasts of Somalia (namely north of Pate, up to Fuma and in Kismayo). In Bajuni, the Bantu */t/ is affricated to be [tʃ] (e.g., *tatu* ‘three’ is realized as *chachu*), in addition, obstruents, e.g., the voiced dental fricative [ð] is replaced by [z] before /i/ as in /maðziwa/ ‘milk’ (originally *idhiva*). Bajuni and Mrima said to share 72% of their lexicons, Bajuni and Amu share 85% of their lexicons, while Bajuni and Mvita share 78% of theirs ([Paul et al., 2016](#)). Bajuni is also known as Bajun, T’ik’uu, Tikulu, Tukulu, Gunya, Mbalazi and Tikuu ([Paul, 2009](#)).

Pate Dialect spoken on Pate Island close to the northern coast of Kenya. Its treatment of Bantu obstruents is similar to that of Bajuni.

Siu Variant spoken in Siyu on Pate Island. Its treatment of Bantu obstruents is similar to that of Bajuni. Siu is also known as Siyu ([Hammarström et al., 2015](#)).

Amu Variant spoken on Lamu Island which is part of the Lamu Archipelago of Kenya. Amu reportedly shares 79% of its vocabulary with Mrima, 85% with Bajun and 86% with Mvita ([Paul et al., 2016](#)).

Matondoni Variant spoken on the northwestern coast of Lamu Island ([Hammarström et al., 2015](#)).

Mvita Variant spoken on Mombasa Island, at the southern part of Kenya’s coast on the Indian Ocean. In this dialect Bantu */c/ is reflected as a voiceless dental stop [t] in general, and [tʰ] after /n/ in particular (e.g., *bichi* ‘unripe’ manifest as *biti*). Additionally, Bantu */j/ is reflected as [d]

after /n/ (e.g., *njaa* ‘hunger’ is *ndaa* in the Mvita dialect). Similarly, [t̪] is realized as [d] after nasals (e.g., *ndani* ‘inside’). Speakers of the dialect also omit redundant non-locative pronominal concords with the reference particle *-o-*. Polomé distinguishes two subdialects: Cijmvu (spoken by Jomvu) and Kingare (spoken in the Ngare area of Changamwe). For further details see Polomé (1967, p. 21–22). Mvita is also known as Kimvita and Mombasa (Paul, 2009). According to (Paul et al., 2016), Mvita shares 79% of its vocabulary with Mrima, 78% with Bajun and 86% with the Amu variant.

Cifundi Dialect spoken north of Vanga, on the Southern Kenya coast. Speakers of the dialect tend to affricate intervocalic fricatives, and weaken unaspirated /t/ to [r] and unaspirated /p/ to voiced bilabial fricative [β] in intervocalic position. Additionally, they palatalize /k/ and /g/ before front vowels and use the present negative of *jambwa* followed by the mere verbal stem (Polomé (1967, p. 23)). Cifundi is also known as Fundi (Hammarström et al., 2015).

Vumba Speech variant spoken on Wasin Island and Jimbo, as well as the opposite mainland in southeast Kenya, Vumba primarily differs from other variants in its verbal system. The present negative tense is characterized by the assimilation of the final vowel to the penultimate (*hamvuti* ‘you do not pull’ realizes as *k’amvuru*). The dialect also displays an ‘immediate perfect’ aspect with the tense-marker [a] (e.g., ‘I have understood [what you just told me]’) and its speakers do not use tense-marker in the positive past tense. Beyond its morphosyntactic peculiarities, Vumba also displays a post-radical vowel harmony in verbs (e.g., *nimepotea njia* ‘I have lost my way’ is realized as *nyaga njia*).

Mtanata The dialect is spoken on the Mrima Coast in Tanganyika, between Pangani and Tanga. It is characterized by the occurrence of a special pronominal subject-prefix (Polomé (1967, p. 23)). The post-radical vowel harmony in verbs described in the section of the Vumba dialect is also present in Mtanata. Mtanata is also known as Mrima and Mtang’ata (Paul, 2009).

Pemba Spoken on all parts of Pemba Island except its southern tip. Pemba is characterized by the absence of palatalization in the prefix *ki-* before nouns with an initial vowel. The vowel harmony in described in the section of the Vumba dialect is also present in Pemba. Pemba is also known as Phemba, Hadimu and Tambatu (Paul, 2009).

Tumbatu Spoken on Tumbatu Island north of Zanzibar, the northern tip of Zanzibar Island, and on the southern tip of Pemba Island. The vowel harmony described in the section of the Vumba dialect is also present in Tumbatu.

Hadimu Also known as Makunduchi, spoken in the southern part of Zanzibar Island. The post-radical vowel harmony described above is present in this dialect too. In addition, the dialect also displays the lenition of /p/ to [β], its future tense-marker, instead of [ta], appears as [t^ha], and it utilizes a unique paradigm of *kuwa* ‘to be’ Polomé (1967, p. 24).

Unguja The dialect, spoken in the central part of Zanzibar, is widely accepted as the basis of ‘standard’ Swahili. Unguja is used in education, administration and also in literary context. It is also known as Kiunguja and Zanzibar (Paul, 2009).

Mgao Variety spoken to the south of Kilwa, on the so-called Mgao coast (Tolmacheva, 1996, p. 22). According to Stigand and Taylor (2013), in Mgao (or Kimgao) [l] and [r] displays a variability (e.g., *ta/j/rili* ‘rich man’ are both acceptable). Another characteristic feature of the language is the tendency to resolve certain vowel clusters by inserting a *y* ‘letter’ (sic) (e.g., the standard form *aendae* ‘he who goes’ is realized as *aendaye*). Besides, certain grammatical borrowings also manifest, probably as a result of the contact with the surrounding island languages (Stigand and Taylor, 2013, p. 23).

Ngazija Spoken on the island of Grand Comoro.

Nzwani Spoken on the island of Anjouan. Polomé (1967, p. 25) notes that the latter two variants, i.e., the Comoro dialects are peripheral. Anjouan and Ngazija share certain features with the northern dialects (e.g., post-radical vowel harmony, sets of pronominal subject-prefixes).

Inland variants

Kisettla The contact vernacular, originally used by European settlers with their domestic and farm staff. It is characterized by oversimplification of morphological and syntactic patterns.

Kishamba Up-country Swahili, typically used in rural areas (*shamba* means ‘country’).

Kihindi Trade language used by Indian shopkeepers.

Kivita Army Swahili (has been extinct in the 1960s, current status not known).

Kingwana Dialect spoken by the descendants of the African helpers of the Arab merchants. It is mainly spoken in Ujiji, Tabora and Mwanza.

13.2 Main typological and syntactic features

Swahili belongs to the Bantoid branch of the Atlantic-Congo language family (Hammarström et al., 2015). Since most of Swahili’s speakers are L2 users, the language and its dialects encompasses many loan features from the mother-tongues used by its speaker population (Polomé, 1967). To overview the core properties, the typology below will focus on the ‘standard’, Unguja variant.

13.2.1 Linguistic typology

Phonological level Swahili maintains contrast between voiced and unvoiced stops and fricatives. It displays a moderately large consonant inventory (encompassing for instance somewhat unusual implosives), as well as a relatively high consonant to vowel ratio (22 consonants, but only 5 vowels Paul et al. 2016). Standard Swahili is not a tone language and displays no vowel harmony (however there is some dialectal variation, see Section 13.1.4). Stress usually falls on the penultimate syllable (Dryer and Haspelmath, 2013; UCLA Language Materials Project, 2006).

Morphological level Swahili is a weakly prefixing agglutinative language utilizing both full and partial reduplication (Dryer and Haspelmath, 2013; UCLA Language Materials Project, 2006).

Morphosyntactic level As compared to European languages, Swahili employs a rich gender-system. Besides female, male and neutral grammatical genders, the language also employs classes with certain semantic characteristics (e.g., human beings, animals, plants, artifacts, long objects, abstract concepts, etc.). Swahili verbs operate with a system of affixes that marks grammatical relations (e.g., subject, object, tense, aspect, mood) ([UCLA Language Materials Project, 2006](#)).

Syntactic level The typical word order in Swahilis is *svo* ([Dryer and Haspelmath, 2013](#)). Below a simplistic overview of the four basic sentence patterns are described. For further details, the reader is advised to consult with the comprehensive descriptive grammar of [Thompson and Schleicher \(2001\)](#).

13.2.2 Predication

The unmarked Swahili word order is *svo*. The majority of the informations needed to create simple sentences are carried by verbal prefixes ([Benjamin et al., 1998](#), p. 19). Verbal prefixes together with the verbal stem form so-called verbal complexes. The initial prefixes of the verbal complex usually carry subject agreement information, which is followed by tense and object agreement markers, and finally by the root. Following the root, suffixes and final vowels might be found ([Deen, 2002](#), p. 21).

- (77) *Mtoto anataka baisikeli.*
child 3SG.PRES.WANT bike

‘The child wants a bike.’

13.2.3 Possession

Below, possession is illustrated with a sample sentence in which the possessor is a pronoun. In addition, Swahili also employs seven possessive stems which indicate the possessor of a given object. For a detailed overview see [Benjamin et al. \(1998, p. 41\)](#).

- (78) *Nakaa na dada yangu.*
dwell.1SG with sister my
‘I live with my sister.’

13.2.4 Imperative

Imperatives in Swahili prototypically do not take an overt subject unless (as in other languages) the subject is focused or contrastive ([Deen, 2002](#), p. 31).

- (79) *Nunua mkate!*
buy.IND CLASS.bread
‘Buy bread!’

13.2.5 Interrogative

The general order of sentence constituents in interrogatives remains unchanged. To disambiguate homonymous expressions, speakers of Swahili utilize rising intonation in questions (Benjamin et al., 1998, p. 43).

- (80) *Aliona nini?*

3SG.PAST.see.IND what

‘What did he see?’

The same also holds for yes-no questions (c.f., *Watoto wanakula*. ‘The children are eating.’ and *Watoto wanakula?* ‘Are the children eating?’) (Wilson, 2006, p. 32).

13.3 Writing system, transcription

The earliest known Swahili texts are letters dating from 1711, written in the Arabic script. During the 19th century Swahili was an administrative language utilized by the European colonial powers in East Africa. Under their influence the Latin alphabet (see Figure 13.3) became more widespread. Contemporary Swahili is written in Latin script (Ager, 2015).

A a	B b	Ch ch	D d	E e	F f	G g	H h
a	be	che	de	e	ef	ge	he
[a]	[b̥]	[f/fʰ]	[d̥]	[ɛ]	[f̥]	[g̥]	[h̥]
I i	J j	K k	L l	M m	N n	O o	P p
i	je	ka	le	em	en	o	pe
[i]	[f~dʒ]	[k/kʰ]	[l]	[m]	[n]	[ɔ̥]	[p/pʰ]
R r	S s	T t	U u	V v	W w	Y y	Z z
re	se	te	u	ve	we	ye	ze
[r]	[s]	[t/tʰ]	[u]	[v]	[w]	[j]	[z]
Other letter combinations							
dh	gh	kh	mb	mv	nd	ng	ng'
[ð]	[ɣ]	[χ]	[m̥b̥]	[m̥v̥]	[n̥d̥]	[n̥g̥]	[ŋ̥]
nj	ny	nz	sh	th			
[n̥j~n̥dʒ]	[n̥]	[n̥z]	[ʃ̥]	[θ̥]			

Figure 13.3: The Latin alphabet of Swahili (Ager, 2015).

(Since the Latin script of Swahili contains no special characters, it is situated in the Basic Latin Unicode block.)

13.4 Previous research on the language

The most important grammatical and intonational properties of the language are detailed in (Ashton, 1944). Loogman (1965) provides a thorough overview of the (grammatical and) syntactic features of Swahili. The monograph entitled *Swahili language handbook* by Polomé (1967) caters with information about the various aspects of the language, coupled with information about its cultural background,

as well as some of its dialectal properties. The reader interested in a more pedagogic description of Swahili should consult with [Thompson and Schleicher \(2001\)](#).

For further resources, refer to the [OLAC entry on Swahili](#).

13.5 Data and sources

13.5.1 Basic vocabulary

The *An Crúbadán project* compiled a set of [Swahili resources written with Latin script](#) (consisting of character trigrams, word bigrams and word frequency tables) compiled from 1,043 documents with a total of 5,351,952 words. In addition, the team has also complied a set of [Swahili resources written with Arabic script](#). This set was compiled from 2 documents with a total of 20,208 words.

Publication	Entries	Pages
Kirkeby 2000	60,000	1069
Massamba 1996	50,000	882
Rechenbach and Gesuga 1967	21,600	641
Perrott and Russell 2003	14,000	266
Safari and Akida 1991	6,000	271
Mwalonya 2012	5,311	231
Awde 2002	5,000	194
Madan 2015	N/A	564
Snoxall and Mshindo 2006	N/A	321
Krapf and Binns 1925	N/A	301
Perrott 1981	N/A	184

Table 13.3: Major paper-based Swahili–English and English–Swahili dictionaries ordered by the number of their entries.

13.5.2 Dictionaries

Paper-based Swahili dictionaries

Table 13.3 provides an overview of the major paper-based dictionaries, indicating the number of entries, as well as their length in pages.

Online Swahili dictionaries

Table 13.4 provides an overview of the available Swahili–English as well as English–Swahili online dictionaries. Dictionaries that are mere wrappers of other engines are not included in the list.

Online Swahili news portals

The most important Swahili news outlets are reviewed in Table 13.5. To give an impression on the quality of the individual sites, post lengths and posting frequencies were measured. Mean post lengths are indicated in characters (not counting whitespaces) and are calculated from the random sample of 10 post bodies (titles are not included in the figures, artifacts, such as code chunks, were removed). Since most sites do not sport a list of their material ordered by release date, post frequencies for

Dictionary	Developer	Year	Size	Quality		Scrapeability		
				IPA	POS	LIST	QUERY	POST
Tuki English-Swahili Dictionary	University of Dar Es Salaam	2000	>50,000	×	×		N/A	
Online Swahili - English Dictionary	TshwaneDJe	2004-2016	>16,000	×			×	
Swahili-English Dictionary	Nino Vessella	2005-2012	12,851			×	N/A	
English Swahili Dictionary online	Glosbe	2011-2016	9,062	×		×	N/A	
Swahili-English xFried Dictionary	Morris Fried	2004	1,500	×	×		N/A	
English to Swahili dictionary	Dicts.info	2003-2016	N/A	×		×	N/A	
bab.la - Swahili-English dictionary	Andreas Schroeter and Patrick Uecker	2016	N/A	×	×	×	N/A	
freedict.com	Parvis	1998-2016	N/A				×	

Table 13.4: Summary table of Swahili–English as well as English–Swahili online dictionaries. Signs in the Quality multicolumn indicate whether a page have IPA transcription for the entries and whether the part-of-speech of a given entry is listed. Signs in the Scrapeability multicolumn indicate whether a word listing is available with hyper links to the entries, and whether an URL-query is possible. In the absence of the latter two, in the post column, the author indicates whether it is possible to send POST requests to the site through an automated tool such as [cURL](#).

Portal	Ranking			Quality		
	Daily unique visitors	Rank in Tanzania	Mean post length (N=10)	Update frequency	Video content	
Mwananchi	5342	39	788.5±223.85 [459-1119]	N/A		×
HabariLeo	5134	130	2582.7±1275 [915-5051]	N/A		
Swahili Times	3550	31	N/A	N/A		
Bongo5	3098	21	935±492 [275-1861]	26.67±21.42 [0-60]		×
Mtanzania	2485	222	3828.2±2806.03 [557-9780]	N/A		
Mwanaspoti	1059	379	1373±595.31 [536-2348]	348.78±911.23 [2-2772]		
Raia Mwema	606	501	3953.2±1946.13 [2132-8573]	N/A		
Radio France Internationale	442	N/A	1372.9±404.82 [973-2415]	106.11±159.31 [14-474]		×
VOA	120	1041	1132.4±587.24 [481-2198]	N/A		×
BBC Swahili	N/A	N/A	917.9±379.74 [491-1850]	N/A		×
DW Kiswahili	N/A	N/A	2912.7±613.46 [2203-4345]	N/A		×
NHK World	N/A	N/A	832.3±735.35 [156-2618]	N/A		×
ParsToday	N/A	N/A	2130±1033.14 [1116-3749]	216.44±308.83 [0-966]		×
Redio Ya Um	N/A	N/A	818.4±341.43 [345-1405]	N/A		×

Table 13.5: Swahili news portals ordered by the number of their daily unique visitors. In addition, the table shows how individual sites are ranked in Tanzania (the country with the most Swahili speakers). As an indicator of ‘quality’, the mean character length of a random sample of posts and mean post lengths are indicated (see Section 13.5.2). The table also highlights if a site hosts video content.

individual portals were measured by noting down the release time of the first 10 posts appearing on the main site, starting from page top. The dates were then ordered from oldest to latest, and the difference of individual steps were calculated and converted to minutes. From this set of 9 difference values, mean and standard deviation values, as well as ranges were calculated. The measurements were carried out between 01:00 AM and 02:00 AM, East African Time (EAT).

13.5.3 Corpora

Monolingual

As of 2016-10-28, the [Swahili Wikipedia](#) consisted of 34,580 public articles. In addition, the ELRA Universal Catalogue contains several monolingual corpora: the [ELRA-ST49](#) corpus contains fixed vocabulary utterances obtained through telephone calls, the [BBC corpus](#) contains general radio broadcasting speech, while the [ELRA-S0375](#) and the [ELRA-S0376](#) corpora were built to provide multilingual speech and text database for language independent speech recognition development.

Bilingual

The various sacred texts are ideal candidates for bilingual corpus building: the Bible is made available in .html format by the [Wordproject®](#), the Quran is also available aligned with various languages in .xml format on the site of the [OPUS corpus project](#). The [New World Translation](#) and the [Book of Mormon](#) is also available in Swahili. Additionally, the Universal Declaration of Human Rights is also uploaded in .txt format on the web page of the [Unicode Consortium](#). Besides religious scripts, the OPUS project also provides aligned corpora compiled from [Ubuntu](#) and [GNOME](#) localization files.

13.6 Computational tools

13.6.1 Language identification

There are several tools that can identify Swahili texts: [Shuyo \(2010\)](#) is a Java implementation, while the Compact Language Detector 2 ([2013](#)) is a C++ variant. The langdetect 1.0.7 Python library also supports Swahili recognition ([Danilak, 2014](#)).

Similarly, the [TextCat](#) language guesser and the [langid.py LangID tool](#) are both compatible with Swahili ([2015](#)).

13.6.2 Tokenizer

The SALMA Swahili Language Manager is capable of tokenization ([Hurskainen et al., 1985](#)). MorphAdorner 2.0, a Java implementation, is also compatible with Swahili ([Burns, 2013](#)). Rami Al-Rfou's [polygot](#) also support tokenization ([2015](#)). (The above tools are also capable of language identification.)

13.6.3 Stemmer

Bavin Ondieki's project entitled *Machine Translation in Swahili*, provides a [Swahili stemmer](#). Tree-Tagger developed by [Schmid \(1995\)](#) also supports Swahili affix removal.

13.6.4 Spell checker

Besides the open-source [HunSpell](#) extension, [Aspell](#) also supports Swahili. In addition, [Apache OpenOffice](#) also offers Swahili spell-checking.

13.6.5 Phrase level and higher tools

This section presents a collection of Swahili trained tools and, in the absence of them, articles dealing with higher-level computational tasks:

- **Swahili part-of-speech tagger:** [Daelemans et al. 2010](#); [Schmid 1995](#)
- **Swahili named entity recognizer:** [Shah et al. 2010](#)
- **Swahili chunker:** not available
- **Swahili parser:** [Schmid 1995](#); [Littell et al. 2014](#)
- **Swahili question answering system:** not available
- **Swahili speech recognizer:** [Gelas et al. 2012](#)
- **Swahili machine translator:** [De Pauw 2008](#); [Google 2006](#); [Babylon Ltd. 2014](#)

13.6.6 End-user support

- **Mac OSX** (as of 2016-10-28): [NO OS-level support](#)
- **Microsoft Windows** (as of 2016-10-28): [Language pack available](#)
- **Ubuntu** (as of 2016-10-28): [No coherent translation \(99% untranslated\)](#)

Bibliography

- S. Ager. Swahili – omniglot, 2015. URL <http://www.omniglot.com/writing/swahili.htm>. [Online; accessed 27-May-2015].
- R. Al-Rfou. polyglot, 2015. URL <http://polyglot.readthedocs.org/en/latest/index.html#>.
- E. O. Ashton. *Swahili Grammar, including intonation*. Longmans, Green and Co, 1944.
- N. Awde. *Swahili-English, English -Swahili Dictionary*. Hippocrene Books, 2002.
- Babylon Ltd. Babylon, 2014. URL <http://translation.babylon-software.com/english/to-hindi/>.
- M. Benjamin, C. Mironko, and A. Geoghegan. *Swahili Phrasebook*. Lonely Planet, 1998.
- P. R. Burns. Morphadornor v2.0, 2013. URL <https://devadornor.northwestern.edu/maserver/>.
- W. Daelemans, J. Zavrel, A. Van den Bosch, and K. Van der Sloot. Mbt: Memory-based tagger, 2010. URL <https://languagemachines.github.io/mbt/>.
- M. Danilak. langdetect 1.0.7, 2014. URL <https://pypi.python.org/pypi/langdetect?>
- G. De Pauw. Bootstrapping machine translation for the language pair English–Kiswahili. *Strengthening the Role of ICT in Development*, page 30, 2008.
- J. U. Deen. *The Acquisition of Nairobi Swahili: The Morphosyntax of Inflectional Prefixes and Subjects*. PhD thesis, UCLA, 2002.
- M. S. Dryer and M. Haspelmath, editors. *Language Swahili*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL http://wals.info/languoid/lect/wals_code_swa.
- East African Community. Community article (EAC) no 137/1999, 1999. URL http://www.eac.int/treaty/index.php?view=article&catid=66%3Achapter-29&id=206%3Aarticle-137--official-language&format=pdf&option=com_content&Itemid=331.
- H. Gelas, L. Besacier, and F. Pellegrino. Developments of Swahili resources for an automatic speech recognition system. In *SLTU*, pages 94–101, 2012.
- M. Goodman. No title, 1965. Presentation at the VIIIth Annual Meeting of the African Studies Association.
- Google. Google translate, 2006. URL <http://translate.google.com/>.
- H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 2.7 - Swahili, 2015. URL <http://glottolog.org/resource/languoid/id/swah1253>.

- D. Harper. Online etymology dictionary - Swahili, 2016. URL http://www.etymonline.com/index.php?allowed_in_frame=0&search=swahili.
- A. Hayden and J. Riesa. Compact language detector 2, 2013. URL <https://code.google.com/p/cld2/wiki/CLD2FullVersion>.
- A. Hurskainen, S. S. Sewangi, and W. Ng'ang'a. Salma, 1985. URL <http://www.njas.helsinki.fi/salama/index.html>.
- Joshua Project. Language - Swahili, 2016. URL <https://goo.gl/BgnFLi>. [Online; accessed 10-October-2016].
- W. Kirkeby. *English Swahili dictionary*. Kakepela Pub. Co., 2000.
- J.L. Krapf and H.K. Binns. *Swahili-English Dictionary: Being Dr. Krapf's Original Swahili-English Dictionary Revised and Re-arranged*. Society for Promoting Christian Knowledge, 1925.
- P. Littell, K. Price, and L. Levin. Morphological parsing of Swahili using crowdsourced lexical resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3333–3339, 2014.
- Alfons Loogman. *Swahili Grammar and Syntax*. Duquesne University Press, 1965.
- M. Lui and T. Baldwin. langid.py, 2015. URL <https://github.com/saffsd/langid.py>.
- A.C. Madan. *English-Swahili Dictionary*. BiblioLife, 2015.
- D. P. B. Massamba. *English-Swahili Dictionary*. Institute of Kiswahili Research, University of Dar es Salaam, 1996.
- J. Mwalonya. *Mgombato: Digo-English-Swahili Dictionary : with a Concise Grammar of the Digo Language*. Köppe, 2012.
- L. M. Paul. Swahili - ethnologue, 2009. URL http://archive.ethnologue.com/16/show_language.asp?code=swh. [Online accessed 20-October-2016].
- L. M. Paul, G. F. Simons, and D. F. Charles. Ethnologue: Languages of the World, 2015. URL <http://www.ethnologue.com>.
- L. M. Paul, G. F. Simons, and D. F. Charles. Swahili - ethnologue, 2016. URL <http://www.ethnologue.com/language/swa>. [Online; accessed 20-October-2016].
- D.V. Perrott. *Concise Swahili and English dictionary: Shwahili-English, English-Swahili*. Hodder and Stoughton, 1981.
- D.V. Perrott and J. Russell. *Swahili Dictionary*. Hodder Headline, 2003.
- E. C. Polomé. *Swahili language handbook*. ERIC, 1967.
- L. Radulesku. About the Swahili language, 2008. URL <https://www.pdx.edu/wll/about-the-swahili-language>.
- C.W. Rechenbach and A.W. Gesuga. *Swahili-English Dictionary. [By] Charles W. Rechenbach, Assisted by Angelica Wanjinu Gesuga [and Others], Etc.* Catholic University of America Press, 1967.

- J.F. Safari and H. Akida. *English-Swahili pocket dictionary*. Lightning Source Incorporated, 1991.
- H. Schmid. Treetagger, 1995. URL <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.
- R. Shah, B. Lin, A. Gershman, and R. Frederking. Synergy: a named entity recognition system for resource-scarce languages such as Swahili using misc machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26, 2010.
- N. Shuyo. Language detection library for java, 2010. URL <https://github.com/shuyo/language-detection>.
- R.A. Snoxall and H.B. Mshindo. *A Concise English-Swahili Dictionary*. Oxford University Press, 2006.
- C. H. Stigand and W. E. Taylor. *Dialect in Swahili: A Grammar of Dialectic Changes in the Kiswahili Language*. Cambridge University Press, 2013. Originally published in 1915.
- K. D. Thompson and A. F. Schleicher. *Swahili Learners' Reference Grammar. African Language Learners' Reference Grammar Series*. ERIC, 2001.
- M. Tolmacheva. Essays in Swahili geographical thought. *Afrikanistische Arbeitspapiere*, 47:173–196, 1996.
- UCLA Language Materials Project. Swahili, 2006. URL <http://lmp.ucla.edu/Profile.aspx?LangID=17&menu=004>.
- P. M. Wilson. *Simplified Swahili*. Longman, 2006.
- World Bank. Population 2015, 2015. URL <http://databank.worldbank.org/data/download/POP.pdf>. [Online; accessed 10-October-2016].

Chapter 14

Tagalog (Noémi Vadász)

Contents

14.1 Demography and ethnography	204
14.2 Main typological and syntactic features	207
14.3 Writing system, transcription	211
14.4 Previous research on the language	212
14.5 Data and sources	213
14.6 Computational tools	216
Bibliography	218

Introduction

Tagalog (/tə'ga:lɒg/; Tagalog pronunciation [te'ga:log]) is an Austronesian language spoken as a first language by a quarter of the population of the Philippines and as a second language by the majority. Its standardized form, officially named Filipino, is the national language and one of the two official languages of the Philippines, the other is English.

Tagalog is related to other Philippine languages, such as the Bikol languages, Ilocano, the Visayan languages, Kapampangan and Pangasinan, and more distantly to other Austronesian languages, such as the Formosan languages, Indonesian and Malaysian, Hawaiian, Malagasy and Māori.

Tagalog is a Central Philippine language within the Austronesian language family. Being Malayo-Polynesian, it is related to other Austronesian languages, such as Malagasy, Javanese, Malay (Malaysian & Indonesian), Tetum (of Timor), and Yami (of Taiwan). It is closely related to the languages spoken in the Bicol Region and the Visayas islands, such as the Bikol group and the Visayan group, including Hiligaynon and Cebuano.

For further information regarding the linguistic relatives of Tagalog see the corresponding [Wikipedia](#) entry.

The ISO 639-3 code (i.e. the [International Organization for Standardization](#)) of Tagalog is **tgl**. The top-level domain for Tagalog websites is **.ph** (Philippines).

For a detailed classification of the language see the [classification of Tagalog](#).

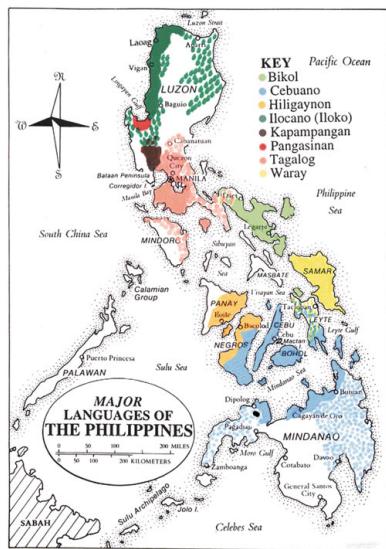


Figure 14.1: The Tagalog-speaking territories in the Philippines

14.1 Demography and ethnography

14.1.1 Name variants

The word *Tagalog* is derived from the endonym *taga-ilog* ('river dweller'), composed of *tagá-* ('native of' or 'from') and *ilog* ('river'). The English exonym is Tagalog.

14.1.2 Geographic spread

According to [Wikipedia](#), the Tagalog homeland, Katagalugan, covers roughly much of the central to southern parts of the island of Luzon—particularly in Aurora, Bataan, Batangas, Bulacan, Camarines Norte, Cavite, Laguna, Metro Manila, Nueva Ecija, Quezon, Rizal, and large parts of Zambales. Tagalog is also spoken natively by inhabitants living on the islands, Marinduque, Mindoro, and large areas of Palawan. It is spoken by approximately 64 million Filipinos, 96% of the household population; 22 million, or 28% of the total Philippine population, speak it as a native language.

Tagalog speakers are found in other parts of the Philippines as well as throughout the world, though its use is usually limited to communication between Filipino ethnic groups. In 2010, the US Census bureau reported (based on data collected in 2007) that in the United States it was the fourth most-spoken language at home with almost 1.5 million speakers, behind Spanish or Spanish Creole, French (including Patois, Cajun, Creole), and Chinese. Tagalog ranked as the third most spoken language in metropolitan statistical areas, behind Spanish and Chinese but ahead of French.

14.1.3 Speaker populations

According to [aboutworldlanguages.com](#), Tagalog is one of the major languages of the Republic of the Philippines. It functions as *lingua franca* and *de facto* national working language of the country. It is used as the basis for the development of Filipino, the national language of the Philippines, a country with 181 documented languages.

According to the Philippine Census of 2000, 21.5 million people claim Tagalog as their first/native language. In addition, it is estimated that 50 million Filipinos speak Tagalog as a second language.

English is the language of higher education. Many Filipinos who are fluent in English frequently switch between Tagalog and English for a variety of reasons. This mixed language is called Taglish. It

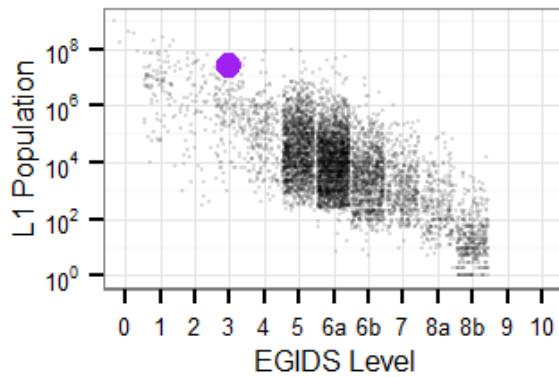


Figure 14.2: Tagalog in language cloud, see the explanation in section .

is more common among educated city dwellers than in rural areas. Frequent contact between Tagalog-speaking and Spanish-speaking people during the Spanish occupation of the Philippines has resulted in Philippine Creole Spanish known as Chabacano.

Since 1940, Filipino has been taught in schools throughout the Philippines. Tagalog is also the language of major literary works, of films, and of the media. According to [Ethnologue](#) Tagalog has been developed to the point that it is used and sustained by institutions beyond the home and community as well, its EGIDS-level is **3** (it is used in wider communication).

The numbers of the population using Tagalog is counted a little bit differently on [Ethnologue](#): 21.500.000 in Philippines (2000 census), and total users in all countries: 24.748.230.

14.1.4 Dialect situation

At present, no comprehensive dialectology has been done in the Tagalog-speaking regions, though there have been descriptions in the form of dictionaries and grammars of various Tagalog dialects. Ethnologue lists Lubang, Manila, Marinduque, Bataan, Batangas, Bulacan, Tanay-Paete (Rizal-Laguna), and Tayabas as dialects of Tagalog; however, traditionally four main dialects are differentiated of which the aforementioned are the parts. These are the following: Northern (exemplified by the Bulacan dialect), Central (including Manila), Southern (exemplified by Batangas), and Marinduque. The Northern and the Central dialects form the basis of the national language.

Some example of the dialectal differences are listed here:

- Many Tagalog dialects, particularly those in the south, preserve the glottal stop found after consonants and before vowels. This has been lost in Standard Tagalog. For example, standard Tagalog *ngayón* ('now', 'today'), *sinigáng* ('broth stew'), *gabi* ('night'), *matamís* ('sweet'), are pronounced and written *ngay-on*, *sinig-ang*, *gab-i*, and *matam-is* in other dialects.
- In Teresian-Morong Tagalog, [f] is usually preferred over [d]. For example, *bundók*, *dagat*, *dingdíng*, and *isdâ* become *bunrók*, *ragat*, *ringríng*, and *isrá*, e.g. *sandók sa dingdíng* becoming *sanrók sa ringríng*.
- In many southern dialects, the progressive aspect infix of *-um-* verbs is *na-*. For example, standard Tagalog *kumakain* ('eating') is *nákáin* in Quezon and Batangas Tagalog. This is the butt of some jokes by other Tagalog speakers, for should a Southern Tagalog ask *nákáin ka ba*

Manila Tagalog	Marinduqueño Tagalog
Susulat siná María at Esperanza kay Juan.	Másúlat da María at Esperanza kay Juan.
‘María and Esperanza will write to Juan.’	
Mag-aaral siya sa Maynilà.	Gaaral siya sa Maynilà.
‘[He/She] will study in Manila.’	
Maglutò ka na!	Paglutò!
‘You cook now!’	
Kainin mo iyán.	Kaina yaan.
‘Eat that.’	
Tinatawag tayo ni Tatay.	Inatawag nganì kitá ni Tatay.
‘Daddy is calling us.’	
Tútulungan ba kayó ni Hilario?	Atulungan ga kamo ni Hilario?
‘Is Hilario going to help you?’	

ng patíng? (‘Do you eat shark?’), he would be understood as saying ‘Has a shark eaten you?’ by speakers of the Manila Dialect.

- Some dialects have interjections which are considered a regional trademark. For instance, the interjection *ala e!* usually identifies someone from Batangas as does *hane?!* in Rizal and Quezon provinces.

Perhaps the most divergent Tagalog dialects are those spoken in Marinduque. Linguist Rosa Soberano identifies two dialects, western and eastern, with the former being closer to the Tagalog dialects spoken in the provinces of Batangas and Quezon.

One example is the verb conjugation paradigms: while some of the affixes are different, Marinduque also preserves the imperative affixes, also found in Visayan and Bikol languages. These affixes have mostly disappeared from most Tagalog varieties early in 20th century, and they have since merged with the infinitive.

The Tagalog language also boasts accentuations unique to some parts of Tagalog-speaking regions. For example, in some parts of Manila, a strong pronunciation of *i* exists and vowel-switching of *o* and *u* exists so words like *gisíng* (‘to wake’) is pronounced as *giseng* with a strong ‘e’ and the word *tagu-taguan* (‘hide-and-go-seek’) is pronounced as *tago-tagoan* with a mild ‘o’.

As it was mentioned, *Taglish* and *Englog* are names given to a mix of English and Tagalog. The amount of English vs. Tagalog varies from the occasional use of English loan words to outright code-switching, where the language changes in mid-sentence. Such code-switching is prevalent throughout the Philippines and in various languages of the Philippines other than Tagalog.

Code Mixing also entails the use of foreign words that are Filipinized by reforming them using Filipino rules, such as verb conjugations. Users typically use Filipino or English words, whichever comes to mind first or whichever is easier to use.

(81) *Magshoshopping kami sa mall. Sino ba ang magdadrive sa shopping center?*

‘We will go shopping at the mall. Who will drive to the shopping center?’

City-dwellers, the highly educated, and people born around and after World War II are more likely to do this. The practice is common in television, radio, and print media as well. Advertisements from

companies like Wells Fargo, Wal-Mart, Albertsons, McDonald's, and Western Union have contained Taglish.

14.2 Main typological and syntactic features

14.2.1 Linguistic typology

According to [seasite](#), Filipino (Tagalog) has been influenced, principally in vocabulary by the languages with which they have come into contact: Sanskrit, Arabic, Chinese, English, and Spanish.

Some of the grammatical features of the Philippine languages are the complex system of affixes, especially of the verbal affixes, which denote a relationship (subject agreement) between the verb and a particular noun phrase in the sentence often referred to by Philippine linguists as 'topic' or 'subject'. This relationship as actor, goal, or referent in the sentence is usually marked by an affix in the verb.

There are other prominent features of the language, such as the use of markers in a sentence, the reduplication of a syllable in a word, and the use of particles between words and phrases.

According to [Bennett \(2005\)](#) Tagalog words consist of roots and affixes. Roots are syntactically neutral. While a root word does refer to a concept which may be inherently a thing, an action or a state, its syntactic function is uncoded until a preposition or an affix is appended to it.

There are some case markers as well. *Ang*, *nang* (traditionally written *ng*) and *sa* are noun markers, namely the nominative, genitive and dative/locative, respectively. The nominative marker is usually described as the subject/topic marker. The genitive marker marks the possessor. Additionally, the topicalized agent or patient in (semantically) transitive sentences are marked by the genitive marker. Finally, the dative/locative marker marks location in its broadest sense.

Attribution is realized in Tagalog at clause level by parataxis, it is marked by zero and the attribute string may be said to be exocentric. At phrase level, it is realized hypotactically, in which case the attributive relation is marked by *na* and the attribute string may be said endocentric. This marker is generally known in Tagalog grammar as the 'linker'. In Tagalog, adjectives are 'linked' to their nouns, adverbs to their verbal words, and the relative string to its head noun by/with? *na*.

Word order in Tagalog has been characterized inaccurately as 'free' or 'completely free'. There are pragmatically motivated constraints in the word order. Whenever the agent is marked by the genitive *ng* it is usually comes immediately after the predicate to avoid ambiguity, particularly when there is another *ng*-phrase in the clause. To be more accurate, word order in Tagalog is 'relatively free'.

14.2.2 Predication

According to [Bennett \(2005\)](#) both at the clause and the phrase level *attribution* is the syntactic operation that "glues" Tagalog structure together. At the clause level, the attribution is realized by simple parataxis; at the phrase level is realized by hypotaxis marked by *na* and *ng*. Since the verbal predicate is syntactically nominal, such verb-based relations as subject and object cannot be said to exist in the predication system of Tagalog.

According to [Schachter and Otanes \(1982\)](#), unlike English predicates the Tagalog ones do not need to include a verb. The Tagalog predicates can be classified in three groups: **nominal**, **adjectival** and **verbal**; only the last of these includes a verb.

(82) nominal predicate-type

Arista ang babaे.

‘The woman is an actress.’

adjectival predicate-type

Maganda ang babaе.

‘The woman is beautiful.’

verbal predicate-type

Yumaman ang babaе.

‘The woman got rich.’

Tagalog sentences that include nominal or adjectival predicates may conveniently be grouped together into a single class of **equational sentences**. The Tagalog predicate is sentence-initial.

14.2.3 Possession

According to Schachter and Otanes (1982) the possessive phrases in Tagalog may occur as predicates of basic or derived sentences. These possessive predicates are of two main types which may be called **possessive *sa* phrases** and **possessive *may* phrases**. Possessive *sa* phrases occur as predicates of sentences expressing possession of some specific, already-identified object which is expressed by the topic of the sentence. If the nominal in the *sa* phrase has an animate referent, it designates the *owner* of the object, and the sentence is often equivalent to an English sentence of the type ‘The __ is/are __’s’ or ‘The __ belong(s) to __’: e.g.

(83) *Sa Nanay ang relos.*

‘The watch is Mother’s.’

‘The watch belongs to Mother.’

If the nominal in the *sa* phrase has an inanimate referent, it designates the *place* where the object properly belongs, and the sentence is equivalent to an English sentence of the type ‘The __ belong(s)/go(es) in/on __’: e.g.,

(84) *Sa kahon ang relos.*

‘The watch belong/goes in the box.’

Sa mesa ang relos.

‘The watch belong/goes on the table.’

A possessive *sa* phrase may be any of the following: (1) *sa* plus an unmarked noun; (2) *sa* plus the *sa* form of a personal pronoun; (3) the *sa* form of a personal pronoun without a preceding *sa* ((2) and (3) alternate freely); (4) *kay* (or its pluralized counterpart *kina*) plus a personal pronoun; (5) the *sa* form of a deictic pronoun.

Possessive *may* phrases occur as predicates of sentences expressing possession of some specific but not previously identified object which is expressed as the *topic* of the sentence. Sentences with

possessive *may* phrase predicates are often equivalent to English sentences of the type ‘_ has/have a/some _’: e.g.,

- (85) *May relos ang Nanay.*

‘Mother has a watch.’

- May pera ang Nanay.*

‘Mother has some money.’

The label **possessive *may* phrase** is used as a cover term for possessive phrases introduced by any one of the following: *may*, *mayroon*, *marami*, *wala*. *Wala* has a negative meaning. *May* and *mayroon* are identical in meaning, and are often in free alternation. A possessive phrase introduced by *marami* expresses possession of a large quantity or number of an object.

Next to the possessive sentences Tagalog has possessive modification constructions as well. There are four types of possessive modification constructions. Two of these belong to the class of the modification constructions. In these contructions, the components – the head (possessed) and the modifier (possessor) – are connected by the linker *na*, and the head and the modifier correspond respectively to the topic and the predicate of an underlying constituent sentence. In one of these contructions, the modifier is a possessive *may* phrase; in the other, the modifier is a possessive *sa* phrase. These constructions allow alternative word orders.

- (86) *batang may lapis / may lapis na bata*

‘children with a pencil’ (‘child that has a pencil’)

- lapis na sa bata / sa batang lapis*

‘the child’s pencil’ (‘pencil that belongs to the child’)

The third type of possessive modification constructions has the following structure:

Head (possessor)	Linker Modifier	(Thing possessed)
Noun	<i>na/-ng</i>	Adjective+ang+Noun

- (87) *batang bago ang lapis*

‘child with the new pencil’

This contruction does not allow alternative word order and its head and modifier are not directly referable to the topic and the predicate of an underlying constituent sentence.

The fourth type of possessive modification construction has the following structure:

Head (Thing possessed)	Modifier (Possessor)
Noun	<i>ng Phrase</i>

- (88) *lapis ng bata*

‘the/a child’s pencil’

This construction does not include the linker *na/-ng*, it is not directly referable to an underlying constituent sentence and the order of its components is not free.

14.2.4 Imperative

According to Schachter and Otanes (1982), Tagalog has five types of imperative constructions. The **basic imperative** is produced by eliminating the aspect marker from the predicate verb of a declarative sentence that includes a second-person-pronoun actor. When the aspect marker is eliminated, the basic form of the verb remains. In the Tagalog basic imperative, the second-person-pronoun actor is explicitly expressed.

(89) Narrational Imperative

Wawalisan/Winawalisan/Winalisan mo ang sahig. Walisan mo ang sahig.

‘You will sweep/are sweeping/swept the floor.’ ‘Sweep the floor.’

The second type is the **equational imperative** that is produced by eliminating the aspect marker from the nominalized topic verb of an equational sentence that includes a second-person-pronoun actor. In the equational imperative, the element (usually a noun or pronoun) that occurs in the predicate position is emphasized.

(90) Equational Imperative

Ito ang BAbasahin mo. Ito ang basahin mo.

‘This is what you’ll read.’ ‘Read this.’

Thirdly, the **immediate imperative** construction consists of an unaffixed verb base, plus – optionally – one or more enclitic particles (notably *na* ‘now’). The construction lacks any explicit expression of the actor.

(91) *Alis (na)!*

‘Leave!’

Bili (na)!

‘Buy (some).’

The fourth type of imperatives, i.e. the **abbreviated imperative** is derived from a basic imperative that includes a predicate secondary-actor-focus indirect-action verb formed with *pa-...in* and a first-person-pronoun topic (*ako* or *kami*).

(92) *Painumin mo/ninyo ako/kami ng tubig.*

‘Let me/us have some water.’

Finally, the **habitual imperative** is derived from a basic or equational imperative by replacing the basic form of the verb with the contemplated-aspect form.

(93) *Mag-aaral ka ng liksyon mo.*

‘Study your lessons (regularly).’

Ito ang gagawin mo.

‘Do this (regularly).’

Any of the five types of imperatives is changed from a command to a polite request by the inclusion in the sentence of the enclitic particle *nga*.

14.2.5 Interrogative

According to Schachter and Otanes (1982) in Tagalog several interrogative words function as interrogative substitutes for the class of structures they normally elicit. Thus, for instance, *kailan* has the grammatical function similar to those of the time expressions as *ngayon* ‘today’, *sa Lunes* ‘next Monday’, or *tuwing Sabado* ‘every Saturday’ have.

- (94) *Kailan ang miting?*

‘When is the meeting?’

- Kailan siya mangisda?*

‘When will he go fishing?’

Tagalog interrogative words generally occur at or near the beginning of the sentences. In this position they substitute predicates or pseudo-predicates, or adverbs occurring initially in emphatic inversion constructions.

Any Tagalog interrogative word may be accompanied, optionally, by the interrogative enclitic particle *ba*.

- (95) *Kailan (ba) siya mangisda?*

14.3 Writing system, transcription

According to Omnidot, Tagalog used to be written with the Baybayin alphabet, which was probably developed from the Kawi script of Java, Bali and Sumatra. The Kawi is in turn descended from the Pallava script, that is one of the southern Indian scripts derived from Brahmi. Today the Baybayin alphabet is used mainly for decorative purposes and the Latin alphabet is used to write to Tagalog.

The earliest known book in Tagalog is the *Doctrina Cristiana* (Christian Doctrine) Guthrie (1994) which was published in 1593. It was written in Spanish and Tagalog, with the Tagalog text in both Baybayin and the Latin alphabet.

The writing system of the Baybayin alphabet is syllabic, in which each consonant has an inherent vowel /a/. Other vowels are indicated either by separate letters, or by dots - a dot over a consonant changes the vowel to an /i/ or and /e/, while a dot under a consonant changes the vowel to /o/ or /u/. The inherent vowel is muted by adding a + sign beneath a consonant. This innovation was introduced by the Spanish. Direction of writing: left to right in horizontal lines.

Tagaloglang introduces the Tagalog alphabet, i.e. the A-ba-ka-da for children. Filipino children learn the alphabet through the ABAKADA, which is the basic Tagalog alphabet with 20 letters. Today, the official Filipino alphabet has 28 letters because of the inclusion of letters from other Western alphabets, but Filipino children still begin their education by learning the basic Tagalog alphabet: a-ba-ka-da. The webpage provides audio sources to ease to learn the Tagalog alphabet.

Cabuay (2012) is written by Baybayin artist and translator, Christian Cabuay who runs Baybayin.com. Baybayin (incorrectly known as Alibata) is a pre-Filipino writing system from the islands

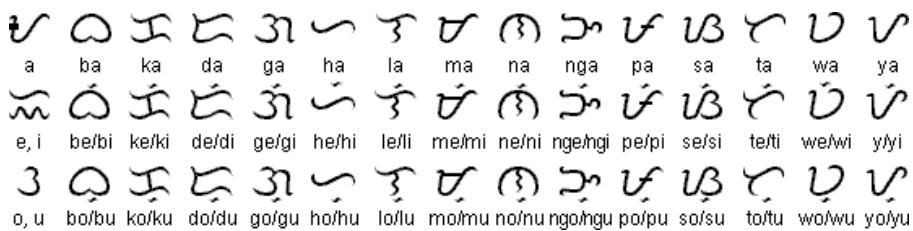


Figure 14.3: Baybayin alphabet

called as the “Philippines”. Baybayin comes from the word ‘baybay’, which literally means ‘spell’. Today, this ancient script is being resurrected thanks to young soul searching Filipinos.

14.4 Previous research on the language

Culwell-Kanarek (2005) introduces word order facts arguing against a theory that Tagalog has a process of *Specificity Shift* in which a specific argument moves to a position above the External Argument. It claims that the *ang*-angargument does not raise above the External Argument ([PDF](#)).

Asarina and Holt (2005) provides an analysis of all Tagalog modal verbs. It shows that there is a clear relationship between the syntax and the semantics of Tagalog modal verbs. They show that Tagalog modals assign a theta-role to the subject if, and only if, they syntactically mark the subject ([PDF](#)).

Sabbagh (2009) investigates the syntax of existential sentences in Tagalog, which pose a particular analytical challenge due to morpho-syntactic as well as semantic properties that are uncharacteristic or ordinary types of finite clauses in the language. Detailed arguments are provided, however, to show that existential sentences have a rather unexceptional syntax after all—they are projected from an unaccusative predicate (i.e. from the existential verb) which selects a DP as its sole internal argument. The unusual characteristics of existential sentences are argued to follow from a particular compositional semantic analysis of the existential predicate ([PDF](#)).

Nagaya (2007) spells out the way syntax and pragmatics interact with each other inside and outside the clause in Tagalog. Inside the clause, different constructions are employed to express different types of focus structure: a canonical construction for predicate focus and sentence focus, a cleft construction for argument narrow focus, and a fronting construction for adjunct narrow focus ([PDF](#)).

Kroeger (1991) presents an analysis of Tagalog within the framework of Lexical-Functional Grammar ([PDF](#)).

Kroeger (1993) tries to show that, at least on language-internal grounds, the identity (and existence) of the grammatical subject in Tagalog is far less problematic than many linguists currently assume ([PDF](#)).

Meladel Mistica (2009) describes research on parsing Tagalog text for predicate–argument structure (PAS). It first outlines the linguistic phenomenon and corpus annotation process, then details a series of PAS parsing experiments ([PDF](#)).

Dionisio (2012) is a formal analysis of the syntax and semantics of the Tagalog plural marker *mga*. It presents data that show that syntactically, *mga* is a predicate modifier that combines with a one-place predicate to form another one-place predicate ([PDF](#)).

14.4.1 Grammars

Adelaar and Himmelmann (2005) is designed to serve as a reference work and in-depth introduction to

these languages, providing a source of basic information for linguists and other professionals concerned with austronesian languages. It contains a detailed Tagalog grammar (the chapter is available in [PDF](#)).

[MacKinlay \(1905\)](#) was made for the War department of the US government (its [online](#) version is available). Since this historical book may have numerous typos and missing text there are several renewed and corrected editions.

[Ramos \(1971a\)](#) approaches the grammar of Tagalog through an examination of word formation, sentence construction, and sentence types. There is also a discussion of the phonology.

Language learning sites mentioned in the section below ([Learningtagalog.com](#), [Mylanguages](#), [Learn101](#), [Tagaloglang](#)) also contain grammar sections.

[Tagalog grammar](#) also provides grammar lessons for beginning Tagalog learners.

[Wikipedia](#) gives a detailed description of Tagalog grammar including sections of Verbs, Nouns, Pronouns, Modifiers, Word order and so on.

14.5 Data and sources

The World Atlas of Language Structures ([WALS](#)) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars). There are materials on [Tagalog](#).

The [Omniglot](#) is an encyclopedia of writing systems and languages. It can be used to learn about languages, to learn alphabets and other writing systems, and to learn phrases in many languages. There is also advice on how to learn languages in general. There are some information about [Tagalog](#).

The [OLAC](#), i.e. the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources. OLAC Archives contain over 100,000 records, covering resources in half of the world's living languages including [Tagalog](#).

[Swadesh lists](#) were originally devised by the linguist Morris Swadesh. In the 1940s to 1950s, Swadesh created/made? word lists of body parts, verbs, natural phenomena, in order to compute the relationships of languages, and in particular their age. A Swadesh list may also be useful to achieve knowledge of some universal terms in other languages. This is because, for basic communication, knowledge of vocabulary is more important than knowledge of grammar and syntax. Sometimes it is even possible to achieve (very) basic communication skills with no knowledge of the target language syntax whatsoever. [Tagalog Swadesh list](#) is provided as well.

Learning Tagalog The [Learningtagalog.com](#) created Learning Tagalog. It is a resource to learn Tagalog that includes literal (word-for-word) and natural (whole-sentence) translations, lots of authentic dialogues and a good grammar reference with plenty of examples. Their mission is to help people become fluent in Tagalog in the fastest, easiest and most enjoyable way possible, by providing high-quality Tagalog learning materials. Their philosophy is to keep all the materials as simple and practical as possible, while making sure everything is accurate. The site contains a detailed grammar organized in topics and subtopics.

The [Mylanguages](#) is designed to teach and help learn many languages for free through vocabularies, phrases, grammars and flashcards (i.e. vocabulary trainers). The developers offer to learn Tagalog very quickly and easily through their lessons.

The [Learn101](#) helps the user learn Tagalog, by going step by step. All the lessons contain audio data and are all offered for free. It teaches the alphabet, also reviews some simple grammatical rules, practices common phrases, and helps memorizing many important vocabulary lists.

The [Tagaloglang](#) provides useful words and phrases in Tagalog with English translations. It contains also sections of grammar, pronunciation and conversation and a Tagalog–English–Tagalog dictionary as well.

Finally, the [SEAsite project](#) is designed with both the beginning and intermediate students of Tagalog in mind. To make the study of the language both challenging and enjoyable, the developers have incorporated a number of learning activities to enhance the user's listening, speaking, reading, and writing skills in Tagalog.

14.5.1 Basic vocabulary

Next to other languages, [Dicts](#) provides the basic vocabulary of Tagalog sorted into main categories.

The language learning sites mentioned in the section above such as the ([Learningtagalog.com](#), the [Mylanguages](#), the [Learn101](#), and the [Tagaloglang](#)) also contain basic vocabulary sections.

14.5.2 Dictionaries

Paper editions

[Perdon \(2013\)](#), is a handy, convenient volume, that is the ideal dictionary for beginner students and travellers. Featuring the essential vocabulary appropriate to beginning and lower intermediate students, the Pocket Tagalog Dictionary covers all the words needed for the everyday situations encountered by travellers.

[Rubino and Llenado \(2002\)](#) is the best-selling two-way dictionary. It is designed for students of the Tagalog language and native Tagalog speakers in need of a bilingual dictionary. It includes a grammatical introduction to the language, a vocabulary appendix with numbers and menu terms, and over 20.000 total dictionary entries, with idiomatic expressions, slang, loan words and derivations.

Online dictionaries

The [Tagalog dictionary](#) is an online browsable dictionary. There are other dictionaries with a search-box, for instance the [tagalog-dictionary](#), the [lingvozone](#), and the [tagalogphrases](#). The [tagalogtranslator](#) is an online translator as well.

The [SEAlang Tagalog dictionary](#) is based on the Tagalog Dictionary by [Ramos \(1971b\)](#).

Scrape The [Tagalog dictionary](#) is a scrapable dictionary due to its relative big size and transparent structure. The words are collected into letters which contain pages. An entry always includes the Tagalog word form, its part-of-speech (or more part-of-speech categories) and its English translation. Sometimes, an expanded or varied word form or an example is included as well, and if there are more translations they are listed.

Searchbox dictionaries with an easily parametrizable search interface like the [tagalog-dictionary](#), and the [lingvozone](#) are scrapable with the help of a world-list which contains the lexical forms of Tagalog words. If there is no more completed list is available, the [Tagalog Swadesh list](#) can be used for this purpose.

14.5.3 Corpora

The [SEAkang Library](#) Tagalog Text Corpus contains more than two million words taken from examples extracted from the Ramos Tagalog–English Dictionary [Ramos \(1971b\)](#); texts from the Tagalog Literary Text collection, prepared by the Philippine Languages Online Corpora project ([Shirley and Rachel Edita \(2011\)](#), [PDF](#)) and various Internet sources, including material located by Kevin Scannell as part of his work on corpus building for minority langauges, [An Crúbadán](#).

The [2008 NIST Speaker Recognition Evaluation Test Set](#) contains 942 hours of multilingual telephone speech and English interview speech along with transcripts and other materials used as test data in the [2008 NIST Speaker Recognition Evaluation \(SRE\)](#). Participants were native English and bilingual English speakers. The telephone speech in this corpus is predominantly English, but also includes the above languages. All interview segments are in English. Telephone speech represents approximately 368 hours of the data, whereas microphone speech represents the other 574 hours.

The [HC corpora](#) is a collection of corpora for various languages freely available to download. The corpora have been collected from numerous different webpages with the aim of building a varied and comprehensive corpus of current use of the respective language. It has been built from many different types of sources, such as newspapers, magazines, (personal and professional) blogs, and Twitter updates. The [statistics](#) show that the total size of the Tagalog corpus is 13.792.000 words. The corpora are collected by a web crawler that checks for language, so as to mainly gets texts consisting of the desired language. Each entry is tagged with it's date of publication. Where user comments are included they will be tagged with the date of the main entry.

Wikipedia The [Wikipedia](#) offers free copies of all available content to interested users. The [Wikimedia](#) is a global movement whose mission is to bring free educational content to the world. The content of Wikimedia is dumped and available as well. The [Tagalog Wikipedia](#) contains 65.562 articles as of 21/10/2016.

These dumps are available in Tagalog in 21/10/2016: [Tagalog Wikipedia](#) and [Tagalog Wikibooks](#) and [Tagalog Wiktionary](#).

Bible The [Jehovah's witnesses](#) and the [Multilingual bible](#) provides Tagalog Bible translations.

Bilingual SEAlang provides a [Tagalog Bitex Corpus](#) as well, which shows words, phrases, and sentences in translation. Insofar as possible, translated texts are aligned sentence-by-sentence. It is based on material taken from The Teresita Ramos Tagalog Dictionary (1971) [Ramos \(1971b\)](#), which conatins well over a thousand examples.

The United Nations Human Rights Office of the High Commissioner [OHCHR](#) worldwide collection of materials on the Universal Declaration of Human Rights (UDHR), including various resources developed by governmental and non-governmental organizations both on the occasion of the Declaration's 50th Anniversary (1998) and prior to/after the Anniversary year. The collection is unique in the world and comprises more than 400 items. Since UDHR is the most translated document – and it is available online in [Tagalog](#) – it can be used as a multilingual corpus.

14.5.4 News portals

The [Wikipedia](#) lists all the main Philipine newspapers. In this list no broadsheets can be found in Tagalog, but a lot of tabloids (Inquirer Libre, Abante, Abante Tonite, Pinoy Weekly...) and some

regional or community newspapers (e.g. *Agila ng Bayan*) are accessible in Tagalog. Some of them, for example the [Abante](#), the <http://balita.net.ph/category/balita-main/nacional/Balita>, the [Pinoy Weekly](#) and the [Philipino Star Ngayon](#) has an online version as well.

14.5.5 Contact person

The [SEAsite](#) is the Center for Southeast Asian Studies at the Northern Illinois University. The purpose of this project is to provide language instructions and other cultural, political, and social information about Southeast Asia. In particular, interactive language instructional materials for – among other languages – the Philippines can be found (see [SEAsite Tagalog](#)). Although the SEAsite server is went down in December 2013, the materials are available. To contact the authors of SEAsite, please send email to henry@cs.niu.edu.

14.6 Computational tools

[Roxas et al. \(2009\)](#) presents the diverse research activities on Philippine languages from all over the country, with focus on the Center for Language Technologies of the College of Computer Studies at the De La Salle University in Manila, where the majority of the work are conducted. The paper is available online in [PDF](#) format.

14.6.1 Language identification

The [CLD2](#), i.e. Compact Language Detector 2, the [TextCat](#), the [Lingua::LanguageGuesser](#) and the [langid.py](#) supports Tagalog.

14.6.2 Tokenizer

No separate tokenizer tool for Tagalog has been found so far. vagy vmi ilyesmi? The [Languagetool](#) has a tokenizer function for Tagalog. The [OmegaT](#) computer aided translation (CAT) tool has a tokenizer for Tagalog.

14.6.3 Stemmer

The [TagSA](#), i.e. a Tagalog Stemming Algorithm, was developed for all forms of Tagalog words. It can be used specifically for morphological analysis to derive root words. In addition, it can also be applied to information retrieval (IR) to conflate different word forms to a common canonical form.

14.6.4 Spell checker

Besides the Open office extension there are some online spell checker tools like the [spellchecker.net](#) and the [stars](#).

The [SpellChef](#) is a spell checker for Filipino that uses a hybrid approach detecting and correcting misspelled words in a document. Its approach is composed of dictionary-lookup, n- gram analysis, Soundex and character distance measurements. It is a plug-in to OpenOffice Writer.

The [Languagetool](#) provides a proof reading device for Tagalog.

14.6.5 Phrase level and higher tools

POS-tagger Miguel and Roxas (2007) conducted empirical tests on part-of-speech taggers for Tagalog without stemming, and implemented rule-based and probabilistic tagging (see the [PDF](#)). It was developed for all forms of Tagalog words. It can be used specifically for morphological analysis to derive root words. In addition, it can also be applied to information retrieval (IR) to conflate different word forms to a common canonical form. It uses the principle of iterative affix removal and is context sensitive. The system implementation was tested and evaluated based on the counting of actual understemming and overstemming errors using a total of 6,382 words variants derived from three sources (duplicates included). The resulting understemming error of less than 15% and overstemming error of less than 0.005% indicates a good performance of TagSA.

Roxas et al. (2006) has shown that the use of a more comprehensive grammar for ambiguity resolution resulted in an improvement to the over-all performance of the tagger. Thus, a study on the development of the Tagalog grammar is recommended to further improve the accuracy of the tagger (see [DOC](#)).

Cheng and See (2006) is a template-based n-gram Part-of-speech (POS) tagger for Tagalog. It is designed to utilize few lexical resources (see [PDF](#)).

Manguilimotan and Matsumoto (2009) investigates factors contributing to the performance of the POS Tagger for Tagalog language (see [PDF](#)).

Dominique et al. (2011) paper explores the use of Support Vector Machines and bi-grams in the Part-of-speech tagging of Filipino words through the creation of SVPOST. Experiments were conducted in order to determine the effect of SVM parameters, such as its kernel or gamma, in the accuracy of the tagger (see [PDF](#)).

Named entity recognition Eboña et al. (2013) is a study intended mainly for the development of a named entity recognition system specifically for handling texts written in the Filipino language (see [PDF](#))

In Castillo et al. (2013), a system for a named entity recognizer for Filipino texts using support vector machine was developed, and its performance was evaluated and compared to an existing named entity recognizer intended for the same language, but uses a rule-based approach (see [PDF](#)).

Speech recognition In Ang et al. (2011) the development of a closed captioning system for Filipino TV news programs is discussed. The researchers tested the system for offline captioning and evaluated the performance of the system based on word error rate (WER) (see [PDF](#)).

14.6.6 End-user support

According to [support.microsoft.com](#), Windows does not have a language pack for Tagalog. For Mac OS X no Tagalog language pack is available either ([support.apple.com](#)).

For Linux, with the help of open-source command-line package-management utility for computers running the Linux operating system [Yum languages](#), including Tagalog.

The [AlanWood](#) provides Philippine unicode fonts for Windows computers. The [101languages](#) provides an on-line Tagalog keyboard. [TagalogKB-Mac](#) is a Tagalog (Baybayin) keyboard layout for Macs.

Bibliography

- K.A. Adelaar and N. Himmelmann. *The Austronesian Languages of Asia and Madagascar*. Routledge language family series. Routledge, 2005.
- F. Ang, M. C. Burgos, and M. De Lara. Automatic Speech Recognition for Closed-Captioning of Filipino News Broadcasts. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2011 7th International Conference on*, pages 328–333, 2011.
- Alya Asarina and Anna Holt. Syntax and semantics of tagalog modals. In *Proceedings of AFLA XII*, number 12 in UCLA Working Papers in Linguistics, 2005.
- N. Bennett. *Subject, Voice and Ergativity*. Taylor & Francis, 2005. ISBN 9781135751890. URL <https://books.google.hu/books?id=KeaQAgAAQBAJ>.
- C. Cabuay. *An Introduction to Baybayin*. LULU Press, 2012. ISBN 9781105422287. URL <https://books.google.hu/books?id=VVfGAQAAQBAJ>.
- Jonalyn M. Castillo, Marck Augustus L. Mateo, Antonio D. C. Paras, Ria A. Sagum, and Vina Danica F. Santos. Named-Entity Recognition Using Support Vector Machine for Filipino Text Documents. *International Journal of Future Computer and Communication*, 2(5):530–532, 2013.
- Charibeth Cheng and S. See. TPOST: A Template-Based, n-gram Part-of-Speech Tagger for Tagalog. *Journal of Research in Science*, 3(1), 2006.
- Nathan Culwell-Kanarek. Word order and the syntax of *ang* in tagalog. In *LSO Working Papers in Linguistics 5: Proceedings of WIGL 2005*, pages 40–50, 2005.
- Michelle Dionisio. The syntax and semantics of the tagalog plural marker *Mga*, a thesis, 2012. Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of The Ohio State University.
- Camille Dominique, E. Reyes, Kevin Rainier, S. Suba, Abigail Razon, and Prospero C. Naval. SV-POST: A Part-of-Speech Tagger for Tagalog using Support Vector Machines, 2011. Conference: 11th Philippine Computing Science Congress, At Ateneo de Naga University, Philippines.
- Karen Mae L. Eboña, Orlando S. Llorca Jr., Genrev P. Perez, Jhustine M. Roldan, Iluminda Vivien R. Domingo, and Ria A. Sagum. Named-Entity Recognizer (NER) for Filipino Novel Excerpts using Maximum Entropy Approach. *Journal of Industrial and Intelligent Information*, 1(1):63–67, 2013.
- S.C. Guthrie. *Christian Doctrine*. Westminster/J. Knox Press, 1994.
- Paul R. Kroeger. *Phrase Structure and Grammatical Relations in Tagalog*. PhD thesis, Department of Linguistics Stanford University, August 1991.

- Paul R. Kroeger. Another look at subjecthood in tagalog. *Philippine Journal of Linguistics*, 2(24): 1–16, 1993.
- William Egbert Wheeler MacKinlay. *A Handbook and Grammar of the Tagalog Language*. U.S. Government Printing Office, Washington, 1905.
- Erlyn Manguilimotan and Yuji Matsumoto. Factors Affecting Part-of-Speech Tagging for Tagalog. In Olivia Kwong, editor, *PACLIC*, pages 763–770. City University of Hong Kong Press, 2009.
- Timothy Baldwin Meladel Mistica. Recognising the predicate–argument structure of tagalog. In *Proceedings of NAACL HLT 2009*, pages 257–260, 2009.
- Dalos Miguel and Rachel Edita O. Roxas. Comparative Evaluation of Tagalog Part-of-Speech Taggers. In *4th National Natural Language Processing Research Symposium Proceedings*, 2007.
- Naonori Nagaya. Information structure and constituent order in tagalog. *Language and Linguistics*, 1(8):343–372, 2007.
- R. Perdon. *Pocket Tagalog Dictionary: Tagalog-English English-Tagalog*. Periplus Dictionaries. Tuttle Publishing, 2013. ISBN 9781462909834. URL https://books.google.hu/books?id=XI_QAgAAQBAJ.
- Teresita V. Ramos. *Tagalog structures*. University of Hawaii Press, Honolulu, 1971a.
- T.V. Ramos. *Ramos: Tagalog Dictionary*. PALI language texts. University of Hawaii Press, 1971b.
- Rachel Edita Roxas, Charibeth Cheng, and Nathalie Rose Lim. Philippine Language Resources: Trends and Directions. In *Proceedings of the 7th Workshop on Asian Language Resources*, ALR7, pages 131–138, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-56-5. URL <http://dl.acm.org/citation.cfm?id=1690299.1690318>.
- Rachel Edita O. Roxas, Bryan Anthony S. Hong, Chris Ian Lim, and Peter Tan. An Integrated Approach to Tagalog Part Of Speech Tagging, 2006. 8th Science and Technology Congress De La Salle University-Manila, oral presentation.
- C.R.G. Rubino and M.G.T. Llenado. *Tagalog-English, English-Tagalog Dictionary*. Hippocrene Standard Dictionary. Hippocrene Books, 2002. ISBN 9780781809603. URL <https://books.google.hu/books?id=ZtHrvikdSJkC>.
- Joseph Sabbagh. Existential sentences in tagalog. *Natural Language and Linguistic Theory*, 27(4), 2009.
- P. Schachter and F.T. Otanes. *Tagalog Reference Grammar*. California Library Reprint Series. University of California Press, 1982.
- Dita Shirley and O. Roxas Rachel Edita. Philippine Languages Online Corpora: Status, issues, and prospects. In *Proceedings of the 9th Workshop on Asian Language Resources, Chiang Mai, Thailand, November 12 and 13*, pages 59–62, 2011.

Chapter 15

Tamil (Nikolett Mus)

Contents

15.1 Demography and ethnography	221
15.2 Main typological and syntactic features	224
15.3 Writing system, transcription	226
15.4 Previous research on the language	226
15.5 Data and sources	227
15.6 Computational tools	229
Bibliography	231

The section describes certain characteristics of the Tamil language (Dravidian; Indian). The general outline of the present section is as follows: the ethnolinguistic situation (see 15.1), main typological features (see 15.2), writing system (see 15.3), the results of previous research concerning the grammatics of the language (see 15.4), the linguistic data and sources (such as dictionaries, corpora, news portals, etc. 15.5), and the computational tools (see 15.6) will be discussed in detail.

15.1 Demography and ethnography

This section deals with the (genetic) classification and characterization of the Tamil language.

15.1.1 Name variants

The **Tamil** language is also called as Damulian, Tamal, Tamalsan, Tambul, Tamili. It belongs to the Southern branch of the Dravidian language family. The ISO 639-3 code of the Tamil language is **tam**.

15.1.2 Geographic spread

The language has official status in the Indian state of Tamil Nadu and the Indian Union Territory of Puducherry. It was the first Indian language, which was declared a classical language by the Government of India in 2004. Figure (15.1) illustrates the Tamil speaking areas of India.

Furthermore, Tamil is also an official and national language of Sri Lanka, and one of the official languages of Singapore. Figure (15.2) shows those areas of Sri Lanka, where the Tamil language is spoken. Besides, Tamil is one of the languages of (medium) education in Malaysia (along with English, Malay and Mandarin).

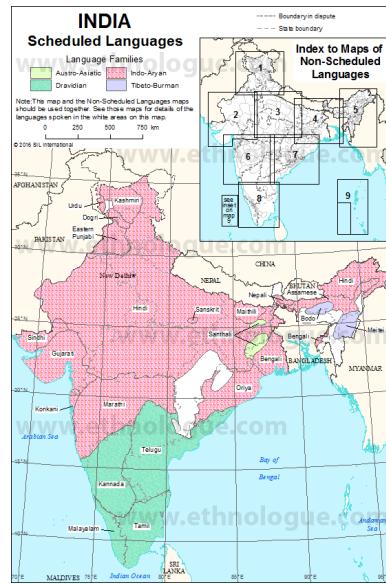


Figure 15.1: The Tamil speaking areas of India (source:[Ethnologue](http://www.ethnologue.com))

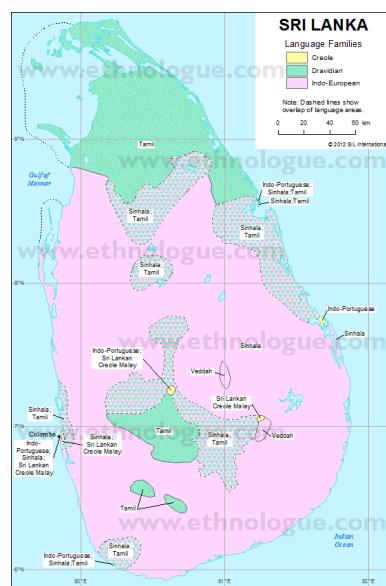
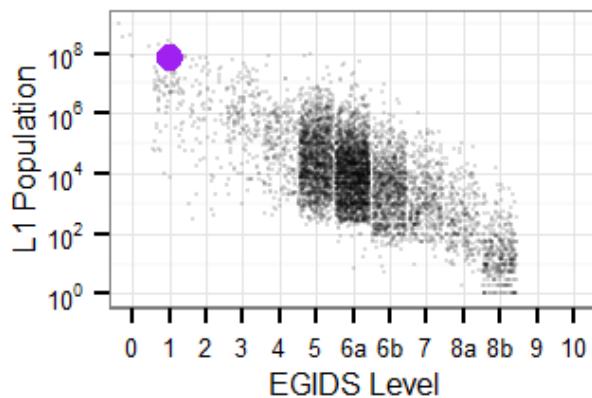


Figure 15.2: The Tamil speaking areas of Sri Lanka (source:[Ethnologue](http://www.ethnologue.com))

Table 15.1: The Persian dialects

Countries	Number of speakers
India	72,000,000
Sri Lanka	3,900,000
Malaysia	1,400,000
Réunion	120,000
Singapore	100,000
Mauritius	32,000
South Africa	26,000

Figure 15.3: The EGIDS level for Tamil (source: [Ethnologue](#))

Finally, the Tamil language is spoken by minority groups across Europe (e.g. in England, Germany, Netherlands and France, a.o.) and in South Africa, as well as, in Canada.

15.1.3 Speaker populations

Tamil is spoken by cc. **78 million** people mainly in India, and in Sri Lanka. The number of speakers is illustrated in Table 15.1.

According to the corresponding [Ethnologue entry of Tamil](#), the EGIDS (Expanded Graded Intergenerational Disruption Scale) level of Tamil is **2**. It means, that the language is a *provincial* language used in education, work, government and mass media. The EGIDS level for Tamil is illustrated in (15.3).

15.1.4 Dialect situation

The Tamil language has further dialects. The classification is, however, not unified. The commonly accepted categorization provided by the [Dialects of Tamil language](#) can be listed in the following:

- the dialects of Kovai, Madurai, Chettinadu, Chennai, Tanjavur, Tirunelveli, Jaffna etc.
- the dialects of the group Brahmin, Pillai, Chettiar, Goundar, Dalit etc. (this is a regional dialect)
- the dialects of immigrants to Tamilnadu, Telugu Naikars, Naidus, Chettiars, Palgat Brahmins, Kannada immigrants etc.
- the dialects of Diaspora Tamils (e.g. in Singapore, Malaysia, Europe, Canada, etc.)

According to Zvelebil (1998), the Southern dialect around Madurai is the literary standard, while the Eastern dialect is the colloquial standard.

15.2 Main typological and syntactic features

15.2.1 Linguistic typology

As previously mentioned, the Tamil language belongs to the Dravidian language family. The Dravidian languages in general are agglutinative languages in which the grammatical relations are expressed by suffixes. In addition, the languages exhibit the SOV basic pattern. For a detailed linguistic description of Tamil see the corresponding [WALS](#) entries.

Phonological level There are twelve vowels (5 shorts and 7 longs) and 18 consonants in Tamil (cf. [A Brief Introduction to Tamil Language and Culture](#)).

Morphological level As mentioned, the grammatical relations are expressed by suffixes.

Morphosyntactic level The Tamil nouns and pronouns can be classified based on animacy. The nouns are inflected for cases and are marked by gender markers. The Tamil verbs functioning as predicates are specified for tense, aspect, mood and agreement. Furthermore, the verbs can take voice and causative suffixes. The inflectional tense category covers the past, present and future tenses. The other tenses require auxiliaries. There are three moods in Tamil, which are the indicative, the optative and the imperative moods. The affective and effective voices belong to the voice paradigm.

Syntactic level The Tamil language is a head final language. This feature characterizes the internal structure of the phrases, as well as the structure of the clause. The basic word order of the declarative clauses is SOV. The phrases appear in the following orders: AdjN, DemN, NumN, GenN, VAux.

15.2.2 Predication

The predicate verbs agree with their subject in person, number, and in 3rd person in gender too. The nominals functioning as predicates do not involve any copular verb so in gender. The structures with predicate nominals do not involve the use of any copular verb.

செய்கிறேன்.

Cey-kir=e-n.
do-TNS-1SG
'I am doing.'

அது பெரிய வீடு.

atu periya v=itu.
that big house
'That is a big house.'

15.2.3 Possession

In adnominal possession, the possessor precedes its head, i.e. the possessed item. The possessor can be pronominal or lexical.

உன் பேராசிரியர் புத்தகம்
 un p=er=aciriyar puttakam
 2SG.GEN professor book
 'your/the professor's book'

15.2.4 Imperative

The 2nd person imperative in Tamil is expressed by the verbal root without any suffices. In addition, there is a polite form of the 2nd person imperative, in which the plural marker is added to the verbal stem.

பார்!
 p=ar!
 'See!'

பாருங்கள்!
 p=aru-.nka!
 see-PL
 'Please see!'

In contrast, for negating the imperatives there is a prohibitive structure available, that differs from the combination of negative and imperative constructions.

15.2.5 Interrogative

The polar or Yes-No questions are expressed by an interrogative suffix (*-a*) that either appears clause finally, or it is attached to the word questioned. In wh-(or content) questions, the interrogative phrases remain *in-situ*.

உங்கள் பெயர் கண்ணா?
 u.nka₁peyar Ka₂an=a?
 2SG name Kannan-Q
 'Is your name Kannan?'

கண்ணன் உங்கள் பெயர்-ா?
 Ka₂an u.nka₁peyar=a
 Kannan 2SG name-Q
 'Is Kannan your name?'

உங்கள் பெயர் என்ன?
 u.nka₁peyar enna?
 2SG name what
 'What is your name?'

15.3 Writing system, transcription

The earliest known inscriptions in Tamil date back to 2,200 BC. Tamil was originally written with a version of the Brahmi script, i.e. *Tamil Brahmi*. The modern Tamil language is written with the so-called Tamil Elluttu ('Tamil letter'). This script is a syllabic one, i.e. not an alphabetic, which is derived from the Brahmi. The syllables are written from the left to the right in horizontal lines. This script was created during the 7th century based on the *Grantha* script. In the 19th and 20th centuries, the script was simplified (cf. [omniglot](#), see figure 15.4).

a	ா	i	ி	u	ு	e	ே	ai	o	ஒ	au
அ	ஆ	இ	ி	உ	ஊ	எ	ஏ	ஐ	ஒ	ஒன்	ஓன்
k	க	கா	கி	க்கீ	கு	கெ	கே	கை	கொ	கோ	கெள்
ங	ங	ஙா	ஙி	ங்ஙீ	ஙு	ஙே	ஙே	ஙை	ஙொ	ஙோ	ஙென்
c	ச	சா	சி	ச்சீ	சு	செ	சே	சை	சொ	சோ	செள்
ஞ	ஞ	ஞா	ஞி	ஞ்ஞீ	ஞு	ஞூ	ஞே	ஞை	ஞொ	ஞோ	ஞென்
t	ட	டா	டி	ட்டீ	டு	டெ	டே	டை	டொ	டோ	டெள்
ந	ந	நா	நி	ந்நீ	நு	நூ	நே	நை	நொ	நோ	நென்
p	ப	பா	பி	ப்பீ	பு	பெ	பே	பை	பொ	போ	பெள்
m	ம	மா	மி	ம்மீ	மு	மெ	மே	மை	மொ	மோ	மெள்
y	ய	யா	யி	ய்யீ	யு	யெ	யே	யை	யொ	யோ	யெள்
r	ர	ரா	ரி	ர்ரீ	ரு	ரெ	ரே	ரை	ரொ	ரோ	ரெள்
l	ல	லா	லி	ல்லீ	லு	லூ	லே	லை	லொ	லோ	லெள்
v	வ	வா	வி	வ்வீ	வு	வூ	வே	வை	வொ	வோ	வெள்
!	ழ	ழா	ழி	ழ்ஷீ	ழு	ழே	ழே	ழை	ழொ	ழோ	ழெள்
!	ள	ளா	ளி	ள்ளீ	ளு	ளூ	ளே	ளை	ளொ	ளோ	ளெள்
r	ற	றா	றி	ற்றீ	று	றூ	றே	றை	றொ	றோ	றெள்
n	ன	னா	னி	ன்னீ	னு	னூ	னே	னை	னொ	னோ	னென்

Figure 15.4: Tamil script

The contemporary Tamil script is an abugida script, which is a segmental writing system, i.e. consonant–vowel sequences are written as a unit. Tamil characters charts can be found on [Omniglot](#).

15.4 Previous research on the language

There are grammatical descriptions of the Tamil language (e.g. [Asher, 1982, 1985; Lehman, 1993; Annamalai and Steever, 1998; Steever, 2001](#)).

Furthermore, the Max Planck Institute for Psycholinguistics examines the Tamil language acquisition within the frame of the [L1 Acquisition Bhuvana Narasimhan](#) longitudinal project.

Research and control bodies There are research centres dealing with the Tamil language, literature, and culture in India, such as the [Pondicherry Institute of Linguistics and Culture](#); the colleagues of the [Central Institute of Indian Languages](#); the [Department of Art and Culture](#) of the Government Puducherry; as well as, outside of India, e.g. at the [Institute for South Asia Studies](#) at the University of Berkeley.

15.5 Data and sources

In this section, sources of Tamil will be presented.

15.5.1 Basic vocabulary

As mentioned in Section 15.4, word (frequency) lists, character trigrams, word bigrams and word frequency lists based on a corpus containing 14,689,508 words are provided by the [An Crubadan project](#).

Additionally, the [300 Languages](#) subproject of The Rosetta Project presents archives of parallel texts and audio recordings of Tamil.

The [Tamilcube](#) provides freely downloadable Tamil sources.

15.5.2 Dictionaries

Traditional (paper) dictionaries

Freely accessible PDF dictionaries can be found at [Tamilcube](#). These are the followings: the Tamil to English dictionary; the Tamil dictionary of Tamil terms for Administration; the Names of birds in Tamil; the Tamil to Tamil dictionary; and the Sanskrit to Tamil dictionary.

The Tamil–English/English–Tamil Dictionary and Phrasebook: Romanized (Hippocrene Dictionary and Phrasebooks) (Tamil Edition) by Clement J. Victor includes over 6,000 dictionary entries, a listing of essential phrases, and a helpful section on pronunciation. The entries are romanized.

The English–Tamil, Tamil–English Dictionary (Tamil Edition) by Jayalalitha Swamy provides the user with English translations, however, the English data constitute the 64% of the dictionary (from the total of 829 pages).

English-Tamil Dictionary (Readwell's) (Tamil) by S. Krishnamurti contains entries written in the Tamil script and English translations.

Online dictionaries

The following online dictionaries are available for Tamil:

- the English–Tamil online dictionary of [Tamilcube](#) contains entries written in the Tamil script, and English translations with contexts. This source can be scraped.
- the scrapeable [English–Tamil–German Dictionary](#) includes words in the Tamil script, and English and/or German translations.
- the English–Tamil Dictionary of [Shabdkosh](#) provides words in the Tamil scripts and their transliterated forms, their English meanings, POS tags, and pronunciation (the entries can be listened).

15.5.3 Corpora

Monolingual corpora

The [Tamil web corpus](#) was prepared by the Corpus factory method. It contains over 26 million words gathered from the Internet in the year 2015.

The Stanford NLP Group provides [monolingual written corpus data](#) for the Tamil language. The corpus is downloadable.

There is an [eBook](#) collection available that contains downloadable PDF-formatted files and audio books.

As of 2016/09/30 the Tamil [Wikipedia](#) contains 261,833 articles. Besides, there is a corpus available which is based on automatically collected data from Wikipedia (see the [W2C-Web to Corpus-Corpora Project](#)).

The following texts and corpora of Tamil exist:

- the [English-Tamil Parallel Corpus \(EnTam v2.0\)](#)
- the [W2C - Web to Corpus - Corpora](#)
- the [HamleDT 2.0.](#)
- the [L1 Acquisition Bhuvana Narasimhan](#)
- the [Tamil](#) translation of the *Bible*
- the full translation of *The book of Mormon*
- the [Tamil](#) translation of *Quran*
- the [translation](#) of *The Universal Declaration of Human Rights*

Bilingual corpora

A [Tamil Dependency Treebank](#) that contains 600 sentences is available. Within the frame of the [Universal Dependencies \(UD\)](#) project, a cross-linguistically consistent treebank annotation for Tamil is developed. The project provides the Tamil data with tokenization, POS tagging, syntactic annotation. The data were manually annotated.

15.5.4 News portals

Radio stations that broadcast in the Tamil language are found primarily in India, Sri Lanka, Malaysia, Singapore, United Kingdom, United States, South Africa, Canada as well as other parts of the world containing a significant Tamil diaspora population.

- Africa: [SA Tamil Radio](#) and [Lotus FM](#)
- London: [BBC Tamil World Service](#)
- Canada:[A9RADIO](#)
- China: [Cheena Tamil Radio](#) which broadcasts in Sri Lanka too
- Australia: [Australian Tamil Broadcasting Corporation](#)
- Malaysia: [Minnal FM](#)
- Singapore: [Oli 96.8FM](#)

There are many types of Tamil TV channels, such as general entertainment channels, movie channels, action movie channels, music channels, kids' channels, etc. Here, the news channels will be focused.

- Lotus News
- Puthiya Thalaimurai
- Raj News

Finally, there are news portals in Tamil, e.g. [thoothuonline](#), [Samachar](#), [ValaiTamil](#) available.

15.6 Computational tools

15.6.1 Language identification

The [CLD2](#) supports the Tamil language. Furthermore, there is also a [TextCat](#) support of Tamil. In addition, a language identifier is provided by the Basis technology ([Rosette Language Identifier](#)). Besides, the [saffsd/langid.py](#) is compatible with Tamil. Finally, there is a language identifier of Tamil developed by [Translated Labs \(T-Labs\)](#).

15.6.2 Tokenizer

The [Indic NLP Library](#) provides Python-based libraries for common text processing and Natural Language Processing in Indian languages. The project developed the following tools for Tamil: text normalizer, tokenizer, word segmenter, script conversion (romanization and indicization, as well as, transliteration) and translator. The [Tamil Dependency Treebank](#) provides, *inter alia*, tokenizer for Tamil.

15.6.3 Stemmer

For the summary of the experiences in developing Tamil stemmer(s) see e.g. [Ramachandran and Krishnamurthi \(2012\)](#) and [M.Thangarasu & Manavalan \(2013\)](#). There is an [open source stemmer](#) available via github. The [Tamil Dependency Treebank](#) provides morphological analyzer, that can be used as stemmer. A Tamil POS tagger is provided by the [Universal Dependencies](#) project. Additionally, see [Dhanalakshmi et. al 2009](#); a.o.) for the documentation of developing Tamil POS tagger and chunker.

15.6.4 Spell checker

There is an open source [spell checker](#) of HunSpell. Additionally, a [Free Online Spell Checker](#) is also available for Tamil. Furthermore, there is a [spell checker](#) developed by Mozilla Firefox.

15.6.5 Phrase level and higher tools

Tamil morphological analyzer A Morphological analyzer developed by [Polyglot](#) is available for Tamil. Additionally, there are further morphological analyzers (see [Vijay Sundar Ram et al., 2010a](#)) and word form generators (cf. [Vijay Sundar Ram et al., 2010b](#)) available.

Tamil part-of-speech tagger [Dhanalakshmi et al. \(2009\)](#) discuss the method of developing Tamil POS tagger and chunker ([PDF](#)). Additionally, the [langtool.org](#) provides POS tagger and sentence tokenizer for Tamil.

Tamil chunker There is also a chunker for the Tamil language (see e.g. [Sobha and Vijay Sundar Ram, 2006, 2010](#)).

Tamil named entity recognizer [Srinivasagan et al. \(2014\)](#) introduces a method of developing a Tamil NER by using hybrid approach. There is a domain focused NER for Tamil (see [Vijayakrishna and Devi \(2008\); PDF](#)). Additionally, [Abinaya et al. \(2015\)](#) present a new approach for Named Entity Recognition (NER) in Tamil language using Random Kitchen Sink algorithm.

Tamil sentence parser A compilation of existing dependency treebanks (or dependency conversions of other treebanks) transformed to the same annotation style is provided by the [HamleDT 2.0 Project](#). This unification is beneficial both to comparative corpus linguistics and to machine learning of syntactic parsing.

Tamil speech recognizer [Dharun and Karnan \(2012\)](#) reports the results of the development of a voice and speech recognition system for Tamil. Furthermore, there is [Triphone based Automatic Speech Recognition engine](#) for the Tamil Language. For further speech recognition tools see e.g. [Vimala & Radha 2012](#)).

Tamil machine translator The [Google translator](#) supports Tamil (for a detailed description and the supported languages see the [Wikipedia entry of Google Translate](#)). Additionally, tools of [machine translation by stars21](#) for Tamil are also available. Finally, the translator developed by [T-labs](#) have Tamil–English pair.

Tamil question answering machine [Ravi and Artstein \(2016\)](#) introduce methods and consequences of translating a Question-Answering System from English into Tamil (see [PDF](#)).

15.6.6 End-user support

The following OS supports are available for Tamil:

- the Tamil language is supported by [Mac OS X](#)
- a [Microsoft Windows](#) language pack is available in Tamil
- there is a language pack of [Linux](#) for Tamil

In addition, there is a Tamil language pack for [Firefox](#).

The Unicode range for Tamil is U+0B80–U+0BFF. There is a unified [Unicode code chart](#) for Tamil available. In addition, there is an online [Tamil Unicode Converter](#).

There are some script converters available via internet (see e.g. [Tamil Typing](#)).

The [Tamil keyboard](#) can be downloaded.

[Online Tamil Tools](#), such as script converters, dictionary, etc. are there available online.

There is [OCR support](#) for Printed Tamil texts.

Bibliography

- N Abinaya, M Anand Kumar, and KP Soman. Randomized kernel approach for named entity recognition in tamil. *Indian Journal of Science and Technology*, 8(24), 2015.
- E. Annamalai and S. B. Steever. Modern tamil. In S. B. Steever, editor, *The Dravidian Languages*, pages 100–128. Routledge, London, 1998.
- R. E. Asher. *Tamil*. North-Holland, Amsterdam, 1982.
- R. E. Asher. *Tamil*. Croom Helm, London, 1985.
- V Dhanalakshmi, M Anand Kumar, S Rajendran, and KP Soman. Pos tagger and chunker for tamil language. In *Proceedings of Tamil Internet Conference*, 2009.
- VS Dharun and M Karnan. Voice and speech recognition for tamil words and numerals. *International Journal of Modern Engineering Research (IJMER)* Vol, 2:3406–3414, 2012.
- T. Lehman. *A grammar of Modern Tamil*. Institute of Language and Culture, Pondicherry, 1993.
- V. A. Ramachandran and I. Krishnamurthi. Noun phrase chunker for tamil. In *ACIIDS'12 Proceedings of the 4th Asian conference on Intelligent Information and Database Systems*. vol. 3., pages 197–205, 2012.
- Satheesh Ravi and Ron Artstein. Language portability for dialogue systems: Translating a question-answering system from english into tamil. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 111, 2016.
- L. D. Sobha and R. Vijay Sundar Ram. Noun phrase chunker for tamil. In *Proceedings of Symposium on Modeling and Shallow Parsing of Indian Languages*, pages 194–198, 2006.
- L. D. Sobha and R. Vijay Sundar Ram. Noun phrase chunker using finite state automata for an agglutinative language. In *Proceedings of the Tamil Internet - 2010 at Coimbatore*, pages 218–224, 2010.
- KG Srinivasagan, S Suganthi, and N Jeyashenbagavalli. An automated system for tamil named entity recognition using hybrid approach. In *Intelligent Computing Applications (ICICA), 2014 International Conference on*, pages 435–439. IEEE, 2014.
- S. B. Steever. Tamil. In J. Garry and C. Rubino, editors, *Facts About the World's Languages. An Encyclopedia of the World's Languages: Past and Present*, pages 716–718. HW Wilson, New York and Dublin, 2001.
- R. Vijay Sundar Ram, S. Menaka, and L. D. Sobha. Tamil morphological analyser. In M. Parakh, editor, *Morphological Analysers and Generators*, pages 1–18. LDC-IL, Mysore, 2010a.

- R. Vijay Sundar Ram, S. Menaka, and L. D. Sobha. Morphological generator for tamil. In M. Parakh, editor, *Morphological Analysers and Generators*, pages 82–96. LDC-IL, Mysore, 2010b.
- R. Vijayakrishna and Sobha Lalitha Devi. Domain focused named entity recognizer for tamil using conditional random fields. In *IJCNLP*, pages 59–66, 2008.
- K. Zvelebil. *Tamulica et Dravidica: A Selection of Papers on Tamil and Dravidian Linguistics*. Karolinum–Charles University Press, Prague, 1998.

Chapter 16

Thai (Ekaterina Georgieva)

Contents

16.1 Demography and ethnography	233
16.2 Main typological and syntactic features	237
16.3 Writing system, transcription	239
16.4 Previous research on the language	240
16.5 Data and sources	240
16.6 Computational tools	244
Bibliography	247

This chapter deals with the Thai (Tai; Thailand) language. More specifically, it focuses on the ethno-linguistic situation, the main typological features, and the writing system of the Thai language. It also gives an overview of the previous research on the Thai grammar and the available linguistic data and sources, such as dictionaries and corpora. Finally, it summarizes the available computational tools for the Thai language.

16.1 Demography and ethnography

16.1.1 Name variants

Thai (ภาษาไทย) is the national and official language of Thailand and the native language of the Thai people and the vast majority of Thai Chinese. The [Ethnologue](#) lists the following alternate names: Bangkok Thai, Central Thai, Siamese, Standard Thai, Thai Klang, Thaiklang.

The vast majority of the inhabitants of Thailand (about 35 % of the population) are Thai. Their ancestors first entered the central part of the Southeast Asian mainland about 1000 CE, bringing with them cultural characteristics shaped by contact with the Chinese. In their new home, they were influenced by Khmer and Mon peoples. The Tai became dominant in the 13th century. They combined the linguistic, cultural, and sociopolitical heritage of their Tai ancestors with the Buddhism of the Mon and the statecraft of the Indianized Khmer, resulting in a distinctive Thai culture (source: [Encyclopædia Britannica – Thailand](#)).

As for its linguistic classification, Thai belongs to the Tai group of the Tai–Kadai language family. According to the [Ethnologue](#), Thai belongs to the Southwestern subgroup of the Tai group with 31 other languages spoken in Thailand, China, Vietnam, Laos, Myanmar and India (on the linguistic

diversity of Thailand see the maps below). Below, some of the Tai languages are listed (source: Wikipedia):

- Isan (Northeastern Thai), the language of the Isan region of Thailand, a collective term for the various Lao dialects spoken in Thailand that show some Siamese Thai influences. It is spoken by about 20 million people. Thais from both inside and outside the Isan region often simply call this variant “Lao” when speaking informally.
- Northern Thai (Phasa Nuea, Lanna, Kam Mueang or Thai Yuan), spoken by about 6 million (1983) in the formerly independent kingdom of Lanna (Chiang Mai). Shares strong similarities with Lao to the point that in the past the Siamese Thais referred to it as Lao.
- Southern Thai (Thai Tai, Pak Tai or Dambro), spoken by about 4,5 million (2006).
- Phu Thai, spoken by about half a million around Nakhon Phanom Province, and 300,000 more in Laos and Vietnam (2006).
- Phuan, spoken by 200,000 in central Thailand and Isan, and 100,000 more in northern Laos (2006).
- Shan (Thai Luang, Tai Long, Thai Yai), spoken by about 100,000 in north-west Thailand along the border with the Shan States of Burma, and by 3,2 million in Burma (2006).
- Lü (Tai Lue, Dai), spoken by about 80,000 (2001) in northern Thailand, and 600,000 more in China, Burma, and Laos (1981–2000).
- Nyaw language, spoken by 50,000 in Nakhon Phanom Province, Sakhon Nakhon Province and Udon Thani Province of Northeast Thailand (1990).
- Song, spoken by about 30,000 in central and northern Thailand (2000).

The ISO 639-3 of Thai is **tha**.

16.1.2 Geographic spread

Thai is the official language of Thailand. In addition to Central Thai, Thailand is home to other related Tai languages. The following maps illustrate the geographic distribution of Thai in Thailand, and more generally, the linguistic diversity of Thailand, see Figure 16.1 and Figure 16.2. Additionally, as the [Ethnologue](#) reports, there is a small group of Thai speakers in Cambodia as well, see Figure 16.3.

16.1.3 Speaker populations

According to the [Ethnologue](#), Thai has 20,200,000 speakers in Thailand (2000). Furthermore, there are 40,000,000 Thai L2 users in Thailand. Additionally, the [Ethnologue](#) mentions that Khorat Thai (a dialect of Thai, spoken in the Nakhon Ratchasima Province of Thailand) has 400,000 speakers. In total, there are 60,489,750 Thai users in all countries (as L1: 20,489,750; as L2: 40,000,000).

According to the [Ethnologue](#), the EGIDS level for the Thai language is 1 (de facto national language), cf. Figure 16.4. This means that the language is used in education, work, mass media, and government at the national level.

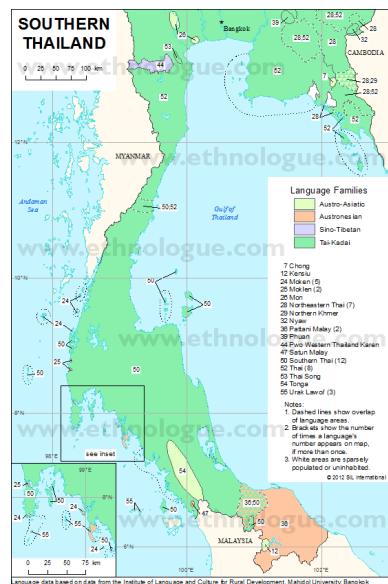


Figure 16.1: The linguistic diversity of Southern Thailand (source: [Ethnologue](http://www.ethnologue.com))

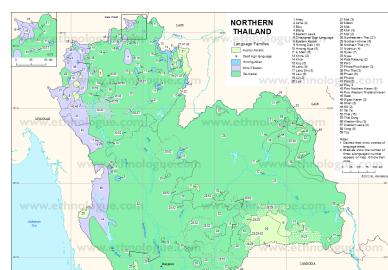


Figure 16.2: The linguistic diversity of Northern Thailand (source: [Ethnologue](http://www.ethnologue.com))



Figure 16.3: The linguistic diversity of Cambodia (source: [Ethnologue](http://www.ethnologue.com))

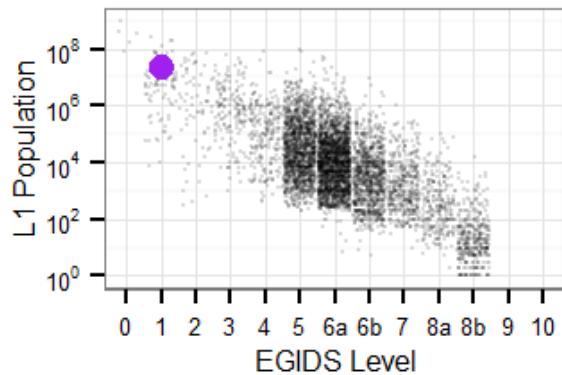


Figure 16.4: Thai – EGIDS Level (source: [Ethnologue](#))

16.1.4 Dialect situation

According to the [Ethnologue](#), Thai has two **dialects**, namely Central (centered in Bangkok) and Khorat dialect (spoken in the Nakhon Ratchasima Province of Thailand). [Smyth \(2014, pp. 1\)](#) claims that distinct regional dialects of Thai are spoken in the northern, northeastern and southern parts of Thailand, however, the language of the Central Region is regarded as the standard and is used both in schools and for official purposes throughout the country.

It is noteworthy that Thai uses several distinct **registers**, i.e. forms for different social contexts, see below (source: [Wikipedia](#)):

- Street or common Thai (ภาษาพูด, phasa phut, spoken Thai): informal, without polite terms of address, as used between close relatives and friends.
- Elegant or formal Thai (ภาษาเขียน, phasa khian, written Thai): official and written version, includes respectful terms of address; used in simplified form in newspapers.
- Rhetorical Thai: used for public speaking.
- Religious Thai: (heavily influenced by Sanskrit and Pāli) used when discussing Buddhism or addressing monks.
- Royal Thai (ราชาศพท์, racha sap): (influenced by Khmer) used when addressing members of the royal family or describing their activities.

As mentioned in the subsections above, several other languages are spoken in Thailand. [Rappa and Wee \(2006, pp. 106\)](#) claim that “despite Thailand’s linguistic diversity, some languages are treated as being variants of the Thai language while others are treated as being clearly foreign”. Moreover, [Rappa and Wee \(2006, pp. 114-115\)](#) mention that “a number of regional varieties are often glossed as being Thai, regardless of their actual linguistic properties”, for instance Northern Thai (for more details regarding the language policy in Thailand see [Rappa and Wee 2006, Ch 5](#) and the references therein).

16.2 Main typological and syntactic features

This section gives an overview of the main typological features of the Thai language (further information can be found in the [WALS](#)'s entry of Thai). Additionally, several constructions, such as predication, possession, imperative and interrogative clauses, will be discussed.

16.2.1 Linguistic typology

With respect to **phonology**, Thai has an average consonant inventory and a large vowel inventory ([Maddieson, 2013b,c](#)). Thai is tonal. This means that the meaning of each syllable is determined by the pitch at which it is pronounced. Thai has a complex tone system ([Maddieson, 2013a](#)) since it uses five tones, namely mid, low, high, rising and falling ([Smyth, 2014](#), pp. 9-10). Much of the original Thai lexicon is monosyllabic; a high percentage of polysyllabic words are foreign borrowings, particularly from the classical Indian languages, Sanskrit and Pali ([Smyth, 2014](#), pp. 1).

As far as its **morphology** is concerned, Thai can be considered to be an analytic language. This means that Thai has no inflections: nouns have a single form, with no morphological distinction between singular and plural. Similarly, past, present and future time can be conveyed by a single verb form. Like many other South-East Asian languages, Thai has a complex pronoun system, which reflects gender, age, social status, the formality of the situation and the degree of intimacy between speakers ([Smyth, 2014](#), Ch 4).

With regard to **syntax**, Thai follows the SVO word order. Furthermore, Thai shows SV and VO word order, and it uses prepositions (for additional information on other syntactic features consult the [WALS](#) entry of the Thai language).

16.2.2 Predication

Word order in indicative clauses generally follows the SVO pattern, cf. the following example (the sentence was taken from [Smyth \(2014](#), pp. 116)):

- (96) พ่อ ชีอ รถ
father bought car

‘The father bought a car.’

Additionally, in spoken Thai, it is common for the subject noun to be followed immediately by a pronoun coreferential with it ([Smyth, 2014](#), pp. 116).

- (97) พ่อ เขายีอ รถ
father he bought car

‘The father bought a car.’

It should be emphasized that either subject or object, or even both, may be omitted when they are understood from the context ([Smyth, 2014](#), pp. 117). Topicalisation is very common in Thai; in this case, the topicalised constituent stands in the beginning of the sentence. In spoken Thai, topics can be marked by a particle that stands at the end of the topicalized constituent or by a preposition introducing it ([Smyth, 2014](#), pp. 117-118).

Additionally, Thai also makes use of non-verbal predicates. In Thai, non-verbal predicates always require a copula ([Stassen, 2013](#)), as shown in the next example (taken from [Smyth 2014](#), pp. 56).

- (98) เขายเป็นนักเรียน
he COP student
'He is a student.'

In this example, the copula is เป็น which is always followed by a noun or a noun phrase, however, it cannot be followed by an adjective. Adjectives in Thai function like stative verbs, i.e. verbs which describe a state rather than an action (Smyth, 2014, pp. 59). Hence, the word เล็ก can be used either as a static verb, i.e. 'to be small', or as an adjective modifying a noun phrase (meaning 'a small house'). Adjectives typically do not occur with the copula used with nominal predicates, the (only) exceptions are some idiomatic expressions (Smyth, 2014, pp. 59).

- (99) บ้านเล็ก
house small
'The house is small.' / 'a small house'

The interested reader may wish to consult Wongwattana (2015) for a detailed description of the Thai copular clauses. Additionally, Phillips and Thiengburanathum (2007) present a detailed description of the Thai verb classes, including the adjectival state verbs.

16.2.3 Possession

In Thai, possession is expressed in the following way. As for the word order, the *possessee* (or *possessed, possessum*) precedes the *possessor* (Dryer, 2013a). Possession is marked by the preposition ของ 'of'. However, this preposition is optional and is often omitted, cf. the example below (source: Smyth 2014, pp. 38):

- (100) บ้าน (ของ) ฉัน
house (of) I
'my house'

16.2.4 Imperative

In Thai, a simple verb form or a verb phrase can be used to express commands (without any morphological marking) as illustrated in the following example:

- (101) ดู
look
'Look!'

As Smyth (2014, pp. 123) notes, these type of commands can sound abrupt; hence, normally some particles are used in imperative sentences, such as the mild command particles or the more insistent particle or can be further softened by the use of polite particles. The interested reader may wish to consult Smyth (2014, pp. 124) for other ways of expressing commands in the Thai language.

Prohibitions are expressed in a different way. According to van der Auwera et al. (2013), in Thai, second singular prohibitives use the second singular imperative verb form, but the negative strategy

is different from the sentential negative found in declaratives. According to Smyth (2014, pp. 145-146), prohibitives are expressed by the particle อย่า which stands in the beginning of the sentence. Additionally, some other particles can be used in order to emphasize the prohibition.

16.2.5 Interrogative

In this subsection, interrogative phrases are discussed. Thai uses both *content* and *polar questions*.

Content questions are introduced by an interrogative phrase. According to Dryer (2013b), Thai interrogative phrases do not occupy sentence initial position, as shown in the example below. However, Smyth (2014, pp. 159) argues that in Thai, the position of some question words varies according to their grammatical function in the sentence, while others have a fixed position. In the example below, the interrogative phrase ใคร ‘who’ is in the end of the sentence (source: Smyth 2014, pp. 160). The interested reader may wish to consult Smyth (2014, pp. 159-169) for more information on content interrogative clauses in Thai.

(102) บ้าน (ของ) ใคร
house (of) who

‘Whose house?’

Polar questions are formed with a question participle (Dryer, 2013c), cf. the particle ไหม in the end of the next example. According to Smyth (2014, pp 153), this is a an information-seeking question particle used in neutral questions, i.e. this particle does not anticipate either a positive or negative response. However, other participles can be also used to form questions, see the discussion in Smyth (2014, pp. 153-159).

(103) อาหาร ญี่ปุ่น แพง ไหม
food Japanese expensive Q

‘Is Japanese food is expensive?’

16.3 Writing system, transcription

Thai is written in a unique script. This script has evolved from a script which originated in South India and was introduced into mainland South-East Asia during the fourth or fifth century AD (Smyth, 2014, pp. 11). The neighbouring Lao and Cambodian scripts resemble Thai. The first recorded example of Thai writing is widely believed to be a stone inscription found by the future King Mongkut (Rama IV, 1851–68) at Sukhothai in 1833, and dated 1283 AD.

It should be emphasized that the Thai writing system is alphabetic (Smyth, 2014). However, in the Thai writing system each consonant may invoke an inherent vowel sound; hence, this script is not a true alphabet, but an *abugida* (source: Wikipedia). In this script, the implicit vowel is ‘a’ or ‘o’. Consonants are written horizontally from left to right, with vowels arranged above, below, to the left, or to the right of the corresponding consonant, or in a combination of positions. Moreover, Thai is written across the page from left to right with no spaces between words; when spaces are used, they serve as punctuation markers, instead of commas or full stops (Smyth, 2014, pp. 11). An example of the Thai script is provided below in Figure 16.5. Additionally, Thai uses a different numeral system, exemplified below in Figure 16.6.

เราทุกคนเกิดมาอย่างอิสระ เราทุกคนมี
 ความคิดและความเข้าใจเป็นของเรางเอง เรา
 ทุกคนควรได้รับการปฏิบัติในทางเดียวกัน.

Figure 16.5: Thai script – sample text (source: [Omniglot](#))

๐	๑	๒	๓	๔	๕	๖	๗	๘	๙	๑๐
ศูนย์	หนึ่ง	สอง	สาม	สี่	ห้า	หก	เจ็ด	แปด	เก้า	สิบ
sōon	nèung	sōng	sām	sèe	hâa	hòk	jèt	bpàet	gâo	sìp
0	1	2	3	4	5	6	7	8	9	10

Figure 16.6: Thai numerals (source: [Omniglot](#))

It should be emphasized that there is no universally applied method for **transcribing** Thai into the Latin alphabet ([Smyth, 2014](#), pp. 2). Official standards are the [Royal Thai General System of Transcription \(RTGS\)](#), published by the Royal Institute of Thailand, and the almost identical [ISO 11940-2](#), defined by the International Organization for Standardization.

Additionally, some linguistic works use a system which is based on the phonemic transcription devised by the American scholar, Mary Haas, in the early 1940s and slightly modified in J. Marvin Brown's AUA Thai course materials ([Smyth, 2014](#), pp. 2). Furthermore, [SAMPA alphabet for Thai](#) is also available ([SAMPA](#) is a machine-readable phonetic alphabet).

16.4 Previous research on the language

This section deals with the linguistic description of Thai (for more information see the [WALS](#) entry of the Thai language and the [OLAC](#) resources about Thai as well). There are several grammars and grammatical descriptions of the Thai language, such as ([Noss, 1964](#)), ([Anthony et al., 1968](#)), ([Warotamasikkhadit, 1972](#)), ([Panthumetha, 1982](#)), ([Higbie and Thinsan, 2002](#)), ([Iwasaki and Horie, 2005](#)), ([Patpong, 2006](#)), and ([Smyth, 2014](#)).

Thailand's [National Electronics and Computer Technology Center \(NECTEC\)](#) aims at undertaking, supporting, and promoting the development of electronic, computing, telecommunication and information technologies through research and development activities. They have also developed several NLP tools for the Thai language (see in the sections below). Additionally, the [Department of Linguistics at the Chulalongkorn University](#) has several research projects connected to the Thai language. The [Thai Studies Center at the Chulalongkorn University](#) offers M.A. and Ph.D Programs in Thai Studies, conducted in English, and does research on several topics connected to the Thai society and culture. One of the most important scientific events in the field of Thai Studies is the International Conference on Thai Studies (ICTS) that has been organized every three year starting from 1981 (the next edition of this conference will take place in 2017 at Chiang Mai University, Thailand).

16.5 Data and sources

This section shows the available sources for the Thai language, such as vocabularies, paper-based and online dictionaires and corpora. In addition, it provides an overview of the news portals available on the internet.

16.5.1 Basic vocabulary

The *An Crúbadán* project provides [Thai resources](#) (consisting of character trigrams, word bigrams and word frequency tables) compiled from 1160 documents, with a total of 3,398,756 words. Furthermore, there is a [Lego](#) Thai vocabulary list and a Thai [Swadesh list](#).

Additionally, there are quite a few learners' sources, including vocabulary lists, available for Thai, see for example [Say Hello in the Thai Language](#), [English speakers' online resource for the Thai language](#), and [Free Android software to learn Thai language](#). [Mylanguages](#) aims at teaching basic level knowledge of several languages and cultures, including Thai. This website categorizes words by part-of-speech (nouns, verbs, adjectives, etc.) and also provides information about the alphabet and transliteration.

16.5.2 Dictionaries

This section gives an overview of the Thai dictionaries (both paper-based and online ones).

Traditional (paper-based) dictionaries

Below, some paper-based dictionaries for Thai are listed:

- [Becker and Pirazzi \(2009\)](#) (28,000 entries and 36,000 definitions, 982 pages);
- [Haas \(1964\)](#) (20,000 entries, includes notes on pronunciation, 638 pages);
- [Allison \(1997\)](#) (10,000 entries, 627 pages);
- [Domnern and Sathienpong \(1995\)](#) (727 pages);
- [McFarland \(1944\)](#) (1019 pages);
- [Thongchai \(1993\)](#) (662 pages);
- [Robertson et al. \(2004\)](#) (5,000 words and expressions, with pronunciation guide and grammar notes, 146 pages).

Online dictionaries

Below, some online dictionaries for Thai are discussed. In addition to the monolingual Official Standard Thai—Thai Dictionary, [The Royal Institute Dictionary](#) (which contains about 38,000 entries, however, it is not scrapeable), there are some bilingual dictionaries as well:

- [Cambridge English–Thai Dictionary](#) contains more than 40,000 entries (example sentences for each word are also provided). The dictionary translates only from English to Thai. (scrapeable)
- [NECTEC Thai–English Dictionary](#) (not scrapeable);
- [English–Thai Bilingual Dictionary](#) (not scrapeable);
- [LongdoDict](#);
- [Thai–English Dictionary](#);
- [LEXiTRON-based Thai—English dictionary](#) (provides transliteration as well);

- Lexilogos, an English–Thai and Thai–English dictionary;
- VOLUBILIS, (Romanized Thai–Thai–English–French) : free databases (ods/xlsx) and dictionaries (PDF) – Thai transcription system (it contains about 100,000 entries);
- Daoulagad Thai, a mobile OCR Thai—English dictionary;
- MyMemory;
- Forvo Pronunciation Dictionary (As of 01/10/20016, it contains 6,883 pronounced words by 1,207 speakers);
- Thai Wiktionary. (As of 11/01/2017, it contains 133,065 entries.)

16.5.3 Corpora

Monolingual

The [Thai National Corpus \(TNC\)](#) has been developed by the [Department of Linguistics at the Chulalongkorn University](#) (the corpus' homepage is only in Thai, however, [Aroonmanakun et al. 2009; Aroonmanakun 2007](#) provide some information about the corpus). According to [Aroonmanakun et al. \(2009\)](#), the corpus is targeted at 80 million words. The aim of this corpus is to be comparable to the British National Corpus in terms of its domains and medium proportions. The TNC contains only written texts coming from different genres, such as fiction, newspaper, academic, non-academic, law, and misc. According to [Aroonmanakun et al. \(2009\)](#), the linguistic annotation provided in the corpus contains marking word boundaries and transcription.

The second corpus is the [HSE Thai Corpus](#), developed by the [HSE School of Linguistics](#) in Moscow. This is a corpus of modern texts written in Thai language. The texts, containing in whole 50 million tokens, were collected from various Thai websites (mostly news websites). Each token was assigned its English translation as well as a part-of-speech tag.

The monolingual [SEAlang Library Thai Text Corpus](#) is part of the [SEAlang Library](#). This project provides language reference materials for Southeast Asia, including bilingual and monolingual dictionaries, monolingual text corpora, aligned bitext corpora, and a variety of tools for manipulating, searching and displaying complex scripts. Its Thai corpus contains 50 million characters, and consists of Thai texts published on the Internet, sampled for research and educational purposes. It is possible to make context searches and collocation searches; it has options for merged view, raw context or restrict collocates.

Furthermore, [NECTEC](#) has developed the [Benchmark for Enhancing the Standard of Thai language processing \(BEST Corpus\)](#) (the webpage of the corpus is only in Thai). According to [Poltree et al. \(2011\)](#), the BEST corpus is a free word corpus with segmented words. It has been created as a benchmark tool to use for Thai segmentation program. BEST 2009 corpus has three different segmented documents. There are articles, encyclopedia, news and novels. The corpus contains about 32,700 non-duplicated common words, 34,100 non-duplicated proper nouns and specific words.

NECTEC has also worked on the [Orchid Thai POS tagged Corpus](#) (for more information see [Charoenporn et al. 1997](#) and [Sornlertlamvanich et al. 1999](#)). As of 01/10/2016, the webpage of the corpus does not work.

[Aw et al. \(2014\)](#) discuss the TaLaPi corpus. TaLaPi is a fully annotated Thai corpus (word segmentation, part-of-speech and named entity). The corpus contains 2,720 articles (1,043,471 words)

from the entertainment and lifestyle domain and 5,489 articles (3,181,487 words) in the news domain, with a total of 35 POS tags and 10 named entity categories. As of 26/10/2016, the corpus does not have a homepage.

Ruangrajitpakorn et al. (2009) present a Categorial Grammar Treebank for Thai (implementation of a method for deriving (minimally labeled) dependency trees from the [Thai Categorial Grammar Bank](#) is available). There is one more treebank, [NAiST Treebank](#), however, as of 11/01/2017, its homepage is inaccessible).

Additionally, Thai is also represented in the [300 Languages Project](#), which is part of the Rosetta Project that aims at collecting materials in every variety of the 300 most widely-spoken languages and macrolanguages in the world.

As of 05/01/2017, [Thai Wikipedia](#) has 114,031 content pages.

Bilingual

There are some good candidates for building parallel text corpora:

- the [Thai](#) translation of the *Bible* is available online on the website of Yehovah's witnesses as well as on the homepage of the [Wordproject](#);
- the [Thai](#) translation of the *Book of Mormon*;
- the *Quran* is also translated into [Thai](#) (also available in .xml format on the homepage of the [OPUS corpus project](#));
- the [Thai](#) translation of the *Universal Declaration of Human Rights*;
- Thai is also represented in the [OpenSubtitles2016](#) (10,238 files, 17.3M tokens, 8.3M sentences);
- [Ubuntu](#) and [GNOME](#) localization files are also available aligned with different languages, including Thai, in .xml format on the homepage of the OPUS corpus project. In addition, a parallel corpus of [KDE4](#) localization files consisting of 75,535 items / 60,75M tokens in 92 languages is also available for Thai (0.2M tokens).

Thu et al. (2016) introduce an ongoing project, Asian Language Treebank (ALT). By 2018, the corpus should have covered the following languages: Indonesian, Japanese, Khmer, Laos, Malay, Myanmar, Philippine, Thai and Vietnamese. In 2014, the project commenced development for Japanese and Myanmar. The domain is news and 1888 articles were randomly selected from English Wikinews. The corpus is targeted at 20,000 sentences, and includes the following steps: word segmentation, word alignment to parallel English translation, part-of-speech tagging and constituency parse trees. These tasks are performed with the ALT Tool, a web-based application. The application was developed with PHP, Javascript, HTML and CSS and is served using the Apache web server (Apache Software Foundation, 1997).

Speech corpora

NECTEC has also worked on speech corpora. Within the Speech-to-Speech Project, the [TSync](#) and [LOTUS](#) corpora have been built. (Within the same project, several speech recognition and machine translation tools have been developed, too, see [16.6.](#))

16.5.4 News portals

Information about Thai newspapers is available at [w3newspapers](#). Among the great number of Thai online newspapers, only a couple of them will be mentioned:

- [Thai Rath](#) (includes video as well);
- [Daily News](#) (includes video as well);
- [Kom Chad Luek](#);
- [Google News](#) in Thailand.

Information about radio stations and TV channels in Thailand is available at [Thailand Live Radio Stations](#) and [TV channels from Thailand](#), respectively (with links).

16.6 Computational tools

This section summarizes the available computational tools developed for the Thai language. (Additionally, [Kawtrakul and Praneetpolgrang 2014](#) present an overview of the research conducted in Thailand in the field of Artificial Intelligence.)

16.6.1 Language identification

Some language identification tools available for the Thai language are the following: [CLD2](#), [TextCat](#), [Polyglot3000](#), [LabsTranslated](#), [saffsd/langid](#), and [Rosette Project](#).

16.6.2 Tokenizer

Since Thai does not have explicit word boundary (see Section 16.3), word segmentation is basic and essential step for processing the Thai language. Different approaches have been proposed for word segmentation in the case of the Thai language, as discussed by [Kawtrakul and Thumkanon \(1997\)](#), [Meknavin et al. \(1997\)](#), [Aroonmanakun \(2002\)](#), [Limcharoen et al. \(2009\)](#) and [Poltree et al. \(2011\)](#). [Phaholphinyo et al. \(2011\)](#) discuss a Thai Word Segmentation Verification Tool. Additionally, it has been proposed that the problem of the Thai romanization and word segmentation should be handled simultaneously ([Aroonmanakun and Rivepiboon, 2004](#)). Furthermore, ([Kruengkrai et al., 2006](#)) suggest that the ambiguities of both word segmentation and POS tagging should/can be solved simultaneously.

The [WordCut](#), [Swath](#), [ThaiWordSeg](#), and [OpenNLP project's Thai sentence detector](#) and [Thai tokenizer](#) are word segmentation tools available for Thai.

16.6.3 Stemmer

This section is not applicable to isolating languages like Thai.

16.6.4 Spell checker

There is an [Online Spell Checker for Thai](#). Additionally, [HunSpell](#) provides a spell checker as well.

16.6.5 Phrase level and higher tools

This section presents a collection of Thai trained tools as well as scientific articles dealing with such tools:

- **part-of-speech tagger.** Within the [OpenNLP](#) project, a [Thai POS-tagger](#) has been developed. Additionally, [RDRPOSTagger](#) provides pre-trained Universal POS tagging models for 40 languages, including Thai. [Kawtrakul and Thumkanon \(1997\)](#) mention that Thai words can belong to more than one word class which posits problems for POS tagging. Additionally, [Murata et al. \(2002\)](#) aim at constructing more accurate taggers by developing new tagging methods: the decision list, maximum entropy and support vector machine methods.
- **named entity recognizer.** [Polyglot](#) supports Named Entity Extraction in Thai.
- **chunker:** As of 11/01/2017, no chunker is available for Thai.
- **morphological analyzer.** [Polyglot](#) provides a tool for morphological analysis for Thai. (See also the tools for tokenization discussed above)
- **sentence parser.** [Tongchim et al. \(2008\)](#) discuss a pilot dependency parser for Thai. A LALR parser was used in the Thai Categorial Grammar ([Ruangrajtpakorn et al., 2009](#)). [Satayamas and Kawtrakul \(2004\)](#) discuss treebank building methods for Thai and report their results with a free implementation of PCFG parser. [Wacharamanotham et al.](#) present AnnoThai, a web-based annotation system which can be used to annotate sentences in the Thai Treebank. (It should be noted that Thai does not mark sentence boundaries, see Section 16.3 and also the studies/tools on tokenization discussed above.)
- **question answering system.** There are several studies dealing with this topic, see [Jitkrittum et al. \(2009\)](#), [Suktarachan et al. \(2009\)](#), [Kongthon et al. \(2011\)](#), [Wanichayapong et al. \(2014\)](#), [Chantrapornchai et al. \(2014\)](#) [SAIKAEW et al. \(2016\)](#), [Pechsiri \(2016\)](#), and [Pechsiri and Piriaykul \(2016\)](#).
- **speech recognizer.** Google Cloud Platform provides [Speech API](#) for Thai. Furthermore, [PUTTiPan Salika](#) is a commercial text-to-speech software, developed by PPA Innovation. This software reads Thai text as well as mixed language (Thai and English) in e-mail, MS-Word document, web pages with human-like natural voice.
- **machine translation.** [Systran Language Server 8](#) supports English–Thai and Thai–English translation. Additionally, [Googletranslator](#), [Bing Translator](#), [Yandex Translator](#), and [Babylon](#) also support Thai. Furthermore, Google Cloud Platform provides [Translation API](#) for Thai.

As far as speech corpora, speech recognition tools and machine translation are concerned, [NECTEC](#) has conducted research in this field, too. Their Speech-to-Speech Project included the following strands of activity: speech synthesis, speech recognition, speech corpora and machine translation (all of the tools developed within this project are listed at: [Thai speech synthesis, recognition, corpora](#), however the website does not contain detailed description of the tools in English). The interested reader may wish to consult [Wutiwiwatchai et al. \(2007\)](#) who provide information about the relevant tools. Below, I will summarize some of them:

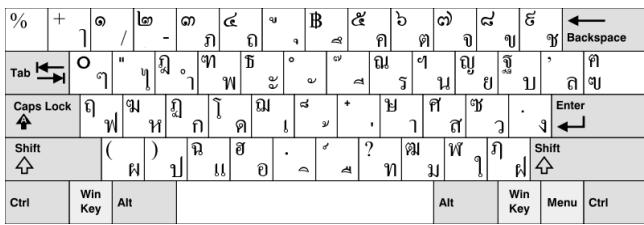


Figure 16.7: Thai Kedmanee Keyboard Layout

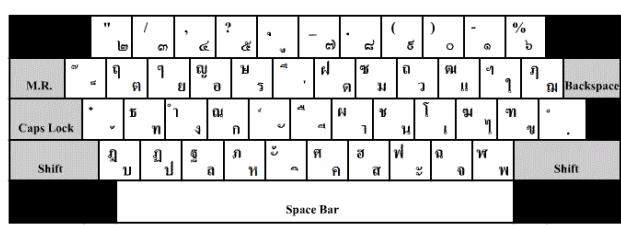


Figure 16.8: Thai Pattajoti Keyboard Layout

- [Vaja](#) is a Thai text-to-speech software developed since 1997. Its most recent version, Vaja 6.0, uses a statistical model called Hidden Markov and a prosody prediction module that make the synthesized speech sounds more natural. Furthermore, Vaja 6.0 is able to synthesize all Thai words, even if they are not included in a dictionary since it has a text analysis module.
- [iSpeech](#) is a Thai automatic speech recognition (ASR) project. It has three versions: iSpeech-W (isolated word recognition); iSpeech-R (continuous speech recognition in restricted grammar); and iSpeech-N (continuous speech recognition in free grammar).
- [Parsit](#) English-to-Thai machine translation system (as of 18/12/2016, its website is under construction).

Additionally, [Potisuk \(2007\)](#) presents a preliminary work on prosody modelling aspects of a text-to-speech system for Thai.

16.6.6 End-user support

Some useful links, including character support, character properties, string manipulators, string collocation, input/output method and word breaking, can be found on [LibThai](#).

Operation systems. A [Windows](#) language pack is available for Thai. Moreover, [MAC OS X](#) and Ubuntu also support Thai as well.

All modern **browsers** support the Thai language: Google Chrome, [Mozilla Firefox](#), [Opera](#), [Internet Explorer](#) and [Safari](#).

Typing Thai text poses some difficulties, since Thai does not use Latin script, moreover, it does not break the text into words (see [16.3](#)). The following Microsoft Windows **typefaces** support Thai: Microsoft San Serif, Segoe UI and Tahoma. Additionally, there are some fonts designed specially for Thai, such as *Garuda*, *Kinnari*, *Loma*, and *Norasi*. In [LaTeX](#), it is possible to type Thai with the *CJKutf8* package (if PDFLaTeX is used as a default compiler). Alternatively, if XeLaTeX is used as a deafult compiler, the *polyglossia* and *fontspec* packages should be used.

Thai has two main **keyboard layouts**: *Thai Kedmanee* and *Thai Pattajoti*, see Figures [16.7](#) and [16.8](#). For additional information on the Thai input methods the interested reader may wish to consult [NECTEC](#)'s webpage.

Additionally, there are several online Thai keyboards, for instance [Branah](#), [Gate2Home](#), [Puttipan](#), and [Lexilogos](#).

Thai is supported by [Unicode](#). There is a [Thai Romanization](#) tool available as well.

[Microsof Office 2013 proofing tools](#) are available for the Thai language.

Additionally, the optical character recognition system [ABBYY FineReader Engine 10](#) recognizes Thai (with dictionary support).

Bibliography

Gordon H Allison. *Jumbo English-Thai Dictionary*. Odeon Store, 5th edition, 1997.

Edward Mason Anthony, Deborah P French, and Udom Warotamasikkhadit. *Foundations of Thai*. University of Michigan Press, 1968.

Wirote Aroonmanakun. Collocation and Thai word segmentation. In *Proceedings of the 5th SNLP & 5th Oriental COCOSDA Workshop*, pages 68–75. Citeseer, 2002. URL <http://pioneer.chula.ac.th/~awirote/ling/SNLP2002-0051c.pdf>.

Wirote Aroonmanakun. Creating the Thai National Corpus. *Manusaya. Special Issue*, 13:4–17, 2007. URL https://www.researchgate.net/publication/27808932_Creating_the_Thai_National_Corpus.

Wirote Aroonmanakun and Wanchai Rivepiboon. A unified model of Thai romanization and word segmentation. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation, Tokyo*, pages 205–214, 2004. URL https://www.researchgate.net/publication/33010877_A_Unified_Model_of_Thai_Romanization_and_Word_Segmentation.

Wirote Aroonmanakun, Kachen Tansiri, and Pairit Nittayanuparp. Thai National Corpus: a progress report. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 153–158. Association for Computational Linguistics, 2009. URL https://www.researchgate.net/profile/Wirote_Aroonmanakun/publication/228971209_Thai_National_Corpus_a_progress_report/links/0deec5272788fce6f0000000.pdf.

AiTi Aw, Sharifah Aljunied Mahani, Nattadaporn Lertcheva, and Sasiwimon Kalunsima. TaLaPi – a Thai linguistically annotated corpus for language processing. In *LREC*, pages 125–132, 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/59_Paper.pdf.

Benjawan Poomsan Becker and Chris Pirazzi. *New Thai-English, English-Thai Compact Dictionary for English Speakers with Tones and Classifiers*. Paiboon Publishing, 2009.

Chantana Chantrapornchai, Archara Sangchan, Atitaya Kantankul, Chidchanok Choksuchat, and Suchitra Adulkasem. Ontology-based question answering system development for case study of Thai cats. *International Journal of Multimedia and Ubiquitous Engineering*, 9(3):187–202, 2014. URL http://www.sersc.org/journals/IJMUE/vol9_no3_2014/18.pdf.

Thatsanee Charoenporn, Virach Sornlertlamvanich, and Hitoshi Isahara. Building a large Thai text corpus-part-of-speech tagged corpus: Orchid. In *Proc. Natural Language Processing Pacific Rim Symposium 1997*, pages 509–512. Citeseer, 1997. URL https://www.researchgate.net/publication/2630580_Building_a_Thai_part-of-speech_tagged_corpus_ORCHID.

- Garden Domnern and Wannnapok Sathienpong. *Thai-English dictionary*. Amarin, 1995.
- Matthew S. Dryer. Order of genitive and noun. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013a. URL <http://wals.info/chapter/86>.
- Matthew S. Dryer. Position of interrogative phrases in content questions. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL <http://wals.info/chapter/93>.
- Matthew S. Dryer. Polar questions. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013c. URL <http://wals.info/chapter/116>.
- Mary Rosamond Haas. *Thai-English student's dictionary*. Stanford University Press, 1964.
- James Higbie and Sneá Thinsan. *Thai reference grammar: The structure of spoken Thai*. Orchid Press, 2002.
- Shoichi Iwasaki and Inkapiromu Puriyā Horie. *A reference grammar of Thai*. Cambridge University Press, 2005.
- Wittawat Jitkrittum, Choochart Haruechaiyasak, and Thanaruk Theeramunkong. QAST: question answering system for Thai Wikipedia. In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions*, pages 11–14. Association for Computational Linguistics, 2009. URL <http://www.aclweb.org/anthology/W09-2703>.
- Asanee Kawtrakul and Prasong Praneetpolgrang. A history of AI research and development in Thailand: three periods, three directions. *AI Magazine*, 35(2):83–92, 2014. URL <http://aaai.org/ojs/index.php/aimagazine/article/viewFile/2522/2430>.
- Asanee Kawtrakul and Chalatip Thumkanon. A statistical approach to Thai morphological analyzer. In *Proceedings of the 5th Workshop on Very Large Corpora*, 1997. URL <http://www.aclweb.org/anthology/W97-0126>.
- Alisa Kongthon, Sarawoot Kongyoung, Choochart Haruechaiyasak, and Pornpimon Palingoon. A semantic based question answering system for Thailand tourism information. In *Proceedings of the KRAQ11 Workshop: Knowledge and Reasoning for Answering Questions*, pages 38–42, 2011. URL <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=7F4354C43D2C67A9C095E23067A500EF?doi=10.1.1.222.1349&rep=rep1&type=pdf>.
- Canasai Kruengkrai, Virach Sornlertlamvanich, and Hitoshi Isahara. A conditional random field framework for Thai morphological analysis. In *Proceedings of LREC*, pages 2419–2424, 2006. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/137_pdf.pdf.
- Piya Limcharoen, Cholwich Nattee, and Thanaruk Theeramunkong. Thai word segmentation based-on GLR parsing technique and word n-gram model. In *Eighth International Symposium on Natural Lanugage Processing*, 2009. URL http://thailang.nectec.or.th/interbest/downloads/InterBEST_1.pdf.

- Ian Maddieson. Tone. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013a. URL <http://wals.info/chapter/13>.
- Ian Maddieson. Consonant inventories. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL <http://wals.info/chapter/1>.
- Ian Maddieson. Vowel quality inventories. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013c. URL <http://wals.info/chapter/2>.
- George Bradley McFarland. *Thai-English Dictionary*. Stanford University Press, 1944.
- Surapant Meknavin, Paisarn Charoenpornsawat, and Boonserm Kijssirikul. Feature-based Thai word segmentation. In *Proceedings of NLPRS97*, 1997. URL <http://www.cs.cmu.edu/~paisarn/papers/nlprs97.pdf>.
- Masaki Murata, Qing Ma, and Hitoshi Isahara. Comparison of three machine-learning methods for thai part-of-speech tagging. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(2):145–158, 2002. URL <https://pdfs.semanticscholar.org/b0bf/21b150defa8bd309a51e591760c275a5d21b.pdf>.
- Richard B Noss. *THAI, REFERENCE GRAMMAR*. ERIC, 1964.
- N Panthumetha. *Thai grammar*. Bangkok: Chulalongkorn University Press, 1982.
- Pattama Patpong. *A systemic functional interpretation of Thai grammar: An exploration of Thai narrative discourse*. PhD thesis, Macquarie University Sydney, Australia, 2006.
- C Pechsiri and R Piriyakul. Developing a why–how question answering system on community web boards with a causality graph including procedural knowledge. *Information Processing in Agriculture*, 3(1):36–53, 2016. URL <http://www.sciencedirect.com/science/article/pii/S2214317316000032>.
- Chaveevan Pechsiri. Explanation based why question answering system. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 104–109, 2016. URL http://www.iaeng.org/publication/IMECS2016/IMECS2016_pp104-109.pdf.
- Supon Phaholphinyo, Kanyanut Klaithin, Sitthaa Kriengket, and Krit Kosawat. Thai word segmentation verification tool. In *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011)*, page 16. Citeseer, 2011. URL <http://www.aclweb.org/anthology/W11-3003>.
- Audra Phillips and Prang Thiengburanathum. Verb classes in thai. *Language and Linguistics*, 8(1):167–191, 2007. URL http://www.ling.sinica.edu.tw/Files/LL/Documents/Journals/8.1/j2007_1_06_0385.pdf.
- Seksan Poltree, KR Saikaew, and Khon Kaen University. Thai word segmentation web service. In *Proceedings of the Joint International Symposium on Natural Language Processing and Agricultural Ontology Service, King Mongkut’s Institute of Technology Ladkrabang, Bangkok, Thailand*, pages 115–9, 2011. URL <https://gear.kku.ac.th/~krunapon/research/pub/wordsegment-snlp-2012.pdf>.

- Siripong Potisuk. Prosodic annotation in a Thai text-to-speech system. In *Proceeding of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 405–414, 2007. URL <http://www.aclweb.org/anthology/Y07-1042>.
- Antonio L Rappa and Lionel Wee. *Language policy and modernity in Southeast Asia*. Springer, 2006.
- Richard G Robertson, Michael Golding, and Benjawan Jai-Ua. *Robertson's practical English-Thai dictionary*. Tuttle Publishing, 2004.
- Taneth Ruangrajitpakorn, Kanokorn Trakultawee, and Thepchai Supnithi. A syntactic resource for thai: CG treebank. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 96–101. Association for Computational Linguistics, 2009. URL http://lexitron.nectec.or.th/KM_HL5001/file_HL5001/Paper/Inter%20Conference/krrn_53922.pdf.
- Kanda Runapongsa SAIKAEW, Seksan POLTREE, Kornchawal CHAIPAH, and Choochart HARUE-CAIYASAK. Improving answer retrieval from web forums with topic model and ontology. *Walailak Journal of Science and Technology (WJST)*, 13(6):451–463, 2016. URL <http://www.thaiscience.info/journals/Article/WJST/10982723.pdf>.
- Vee Satayamas and Asanee Kawtrakul. Wide-coverage grammar extraction from Thai treebank. In *Proceedings of Papillon 2004 Workshops on Multilingual Lexical Databases*, 2004. URL http://www.papillon-dictionary.org/static/info_media/42808974.pdf.
- David Smyth. *Thai: An essential grammar*. Routledge, 2014. URL http://www.uta.edu/faculty/cmfitz/swnal/projects/CoLang/courses/Ped_Grammar/thai_grammar.pdf.
- Virach Sornlertlamvanich, Naoto Takahashi, and Hitoshi Isahara. Building a Thai part-of-speech tagged corpus (ORCHID). *Journal of the Acoustical Society of Japan (E)*, 20(3):189–198, 1999. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.3496&rep=rep1&type=pdf>.
- Leon Stassen. Zero copula for predicate nominals. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL <http://wals.info/chapter/120>.
- Mukda Suktarachan, Patthrawan Rattanamanee, and Asanee Kawtrakul. The development of a question-answering services system for the farmer through SMS: query analysis. In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions*, pages 3–10. Association for Computational Linguistics, 2009. URL <http://www.anthology.aclweb.org/W/W09-W09-27.pdf#page=13>.
- Iamwaramet Thongchai. *A New Thai Dictionary with Bilingual Explanation*. Ruam San, Bangkok, 1993.
- Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Introducing the Asian Language Treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1574–1578, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/435_Paper.pdf.

- Shisanu Tongchim, Randolph Altmeyer, Virach Sornlertlamvanich, and Hitoshi Isahara. A dependency parser for Thai. In *LREC*, pages 136–139, 2008. URL <https://pdfs.semanticscholar.org/c410/32cb0577f6bb6a139f4caadb868dda952f44.pdf>.
- Johan van der Auwera, Ludo Lejeune, and Valentin Goussev. The prohibitive. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL <http://wals.info/chapter/71>.
- Chatchavan Wacharamanotham, Mukda Suktarachan, and Asanee Kawtrakul. The development of web-based annotation system for thai treebank. Manuscript. URL <http://citeseervx.ist.psu.edu/viewdoc/download?doi=10.1.1.592.4021&rep=rep1&type=pdf>.
- Napong Wanichayapong, Wasan Pattara-Atikom, and Ratchata Peachavanish. Road traffic question answering system using ontology. In *Joint International Semantic Technology Conference*, pages 422–427. Springer, 2014. URL http://link.springer.com/chapter/10.1007/978-3-319-15615-6_32.
- Udom Warotamasikkhadit. *Thai syntax: An outline*, volume 68. Mouton De Gruyter, 1972.
- Unchalee S. Wongwattana. Complexities of Thai copular constructions. *Journal of the Southeast Asian Linguistics Society*, 9:97–120, 2015. URL <http://pacling.anu.edu.au/series/SEALS-PDFs/wongwattana2015complexities.pdf>.
- Chai Wutiwiwatchai, Thepchai Supnithi, and Krit Kosawat. Speech-to-speech translation activities in Thailand. In *Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, page 7, 2007. URL <http://www.aclweb.org/anthology/I08-8002>.

Chapter 17

Turkish (Noémi Vadász)

Contents

17.1 Demography and ethnography	254
17.2 Main typological and syntactic features	256
17.3 Writing system, transcription	261
17.4 Previous research on the language	262
17.5 Data and sources	263
17.6 Computational tools	266
Bibliography	269

Introduction

Turkish (/Türkçe/) also referred to as Istanbul Turkish, is the most widely spoken of the Turkic languages, with around 10–15 million native speakers in Southeast Europe (mostly in East and West Thrace) and 60–65 million native speakers in Western Asia (mostly in Anatolia). Outside of Turkey, significant smaller groups of speakers exist in Germany, Bulgaria, Macedonia, Northern Cyprus (only recognized by Turkey), Greece, the Caucasus, and other parts of Europe and Central Asia.

To the west, the influence of Ottoman Turkish – the variety of the Turkish language that was used as the administrative and literary language of the Ottoman Empire – spread as the Ottoman Empire expanded. In 1928, as one of Atatürk’s Reforms in the early years of the Republic of Turkey, the Ottoman Turkish alphabet was replaced with a Latin alphabet.

The distinctive characteristics of Turkish are vowel harmony and extensive agglutination. The basic word order of Turkish is subject–object–verb. Turkish has no noun classes or grammatical gender. Turkish has a strong T–V distinction and usage of honorifics. Turkish uses second-person pronouns that distinguish varying levels of politeness, social distance, age, courtesy or familiarity toward the addressee. The plural second-person pronoun and verb forms are used referring to a single person out of respect.

Turkish is a member of the Oghuz group of languages, a subgroup of the Turkic language family. There is a high degree of mutual intelligibility between Turkish and the other Oghuz Turkic languages, including Azerbaijani, Turkmen, Qashqai, Gagauz, and Balkan Gagauz Turkish. The Turkic family comprises some 30 living languages spoken across Eastern Europe, Central Asia, and Siberia. Some linguists believe the Turkic languages to be a part of a larger Altaic language family. About 40% of all speakers of Turkic languages are native Turkish speakers. The characteristic features of Turkish,



Figure 17.1: Map of the main subgroups of Turkish dialects across Southeast Europe and the Middle East

such as vowel harmony, agglutination, and lack of grammatical gender, are universal within the Turkic family.

The source of this section is [Wikipedia](#).

The ISO 639-3 ([International Organization for Standardization](#)) code for Turkish is **tur**. The top-level domain for Turkish websites is **.tr**.

17.1 Demography and ethnography

Turkish is the official language of Turkey and is one of the official languages of Cyprus. It also has official (but not primary) status in the Prizren District of Kosovo and three municipalities of the Republic of Macedonia, based on the concentration of Turkish-speaking local population.

In Turkey, the regulatory body for Turkish is the Turkish Language Association (Türk Dil Kurumu or TDK), which was founded in 1932 under the name Türk Dili Tetkik Cemiyeti ("Society for Research on the Turkish Language"). The Turkish Language Association was influenced by the ideology of linguistic purism: indeed one of its primary tasks was the replacement of loanwords and foreign grammatical constructions with equivalents of Turkish origin. These changes, together with the adoption of the new Turkish alphabet in 1928, shaped the modern Turkish language spoken today. TDK became an independent body in 1951, with the lifting of the requirement that it should be presided over by the Minister of Education. This status continued until August 1983, when it was again made into a governmental body in the constitution of 1982, following the military coup d'état of 1980.

The source of this section is [Wikipedia](#).

17.1.1 Name variants

The endonym of the language is *Türkçe*. The English exonym is *Turkish*.

17.1.2 Geographic spread

Turkish is natively spoken by the Turkish people in Turkey and by the Turkish diaspora in some 30 other countries. Turkish language is mutually intelligible with Azeri language and other Turkic languages. In particular, Turkish-speaking minorities exist in countries that formerly (in whole or part) belonged to the Ottoman Empire, such as Bulgaria, Cyprus, Greece (primarily in Western Thrace), the Republic of Macedonia, Romania, and Serbia. More than two million Turkish speakers live in

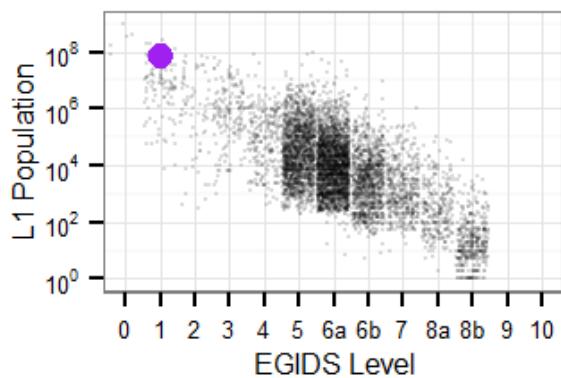


Figure 17.2: Turkish in language cloud, see the explanation in [section](#).

Germany; and there are significant Turkish-speaking communities in the United States, France, The Netherlands, Austria, Belgium, Switzerland, and the United Kingdom. Due to the cultural assimilation of Turkish immigrants in host countries, not all ethnic Turkish immigrants speak the language with native fluency.

In 2005, 93% of the population of Turkey were native speakers of Turkish, about 67 million at the time, with Kurdish making up most of the remainder. However, most linguistic minorities in Turkey are bilingual, speaking Turkish with native-like fluency.

The source of this section is [Wikipedia](#).

17.1.3 Speaker populations

According to [Ethnologue](#) the Turkish speaker population is 66.500.000 in Turkey (European Commission 2006). L2 users: 350.000 in Turkey (European Commission 2006). Total users in all countries: 71.785.850 (as L1: 71.435.850; as L2: 350.000).

The EGIDS-level of Turkish is **1** (National), which means that the language has been developed to the point that it is used and sustained by institutions beyond the home and community.

17.1.4 Dialect situation

Modern standard Turkish is based on the dialect of Istanbul. Dialectal variation persists, in spite of the levelling influence of the standard used in mass media and the Turkish education system since the 1930s. Academically, researchers from Turkey often refer to Turkish dialects as *ağız* or *şive*, leading to an ambiguity with the linguistic concept of accent, which is also covered with these words. Projects investigating Turkish dialects are being carried out by several universities, as well as a dedicated work group of the Turkish Language Association. Work is currently in progress for the compilation and publication of their research as a comprehensive dialect atlas of the Turkish language.

Rumelice is spoken by immigrants from Rumelia, and includes the distinct dialects of Deliorman, Dinler, and Adakale, which are influenced by the theorized Balkan language area. Kıbrıs Türkçesi is the name for Cypriot Turkish and is spoken by the Turkish Cypriots. Edirne is the dialect of Edirne. Ege is spoken in the Aegean region, with its usage extending to Antalya. The nomadic Yörük tribes of the Mediterranean Region of Turkey also have their own dialect of Turkish. This group is not to be confused with the Yuruk nomads of Macedonia, Greece, and European Turkey who speak Balkan Gagauz Turkish.



Figure 17.3: Dark blue: Countries where Turkish is an official language

Light blue: Countries where it is recognized as a minority language

Güneydoğu is spoken in the southeast, to the east of Mersin. Doğu, a dialect in Eastern Anatolia, has a dialect continuum. The Meskhetian Turks who live in Kazakhstan, Azerbaijan and Russia as well as in several Central Asian countries, also speak an Eastern Anatolian dialect of Turkish, originating in the areas of Kars, Ardahan, and Artvin and sharing similarities with Azeri Turkish, the language of Azerbaijan.

The Central Anatolia region speaks Orta Anadolu. Karadeniz, spoken in the Eastern Black Sea Region and represented primarily by the Trabzon dialect, exhibits substratum influence from Greek in phonology and syntax; it is also known as Laz dialect (not to be confused with the Laz language). Kastamonu is spoken in Kastamonu and its surrounding areas. Karamanlıca is spoken in Greece, where it is also named Karamanlidika. It is the literary standard for Karamanlides.

The source of this section is [Wikipedia](#).

17.2 Main typological and syntactic features

17.2.1 Linguistic typology

Turkish is an agglutinative language and frequently uses affixes, and specifically suffixes, or endings. One word can have many affixes and these can also be used to create new words, such as creating a verb from a noun, or a noun from a verbal root. Most affixes indicate the grammatical function of the word. The only native prefixes are alliterative intensifying syllables used with adjectives or adverbs: for example *sımsıcak* ('boiling hot' ← *sicak*) and *masmavi* ('bright blue' ← *mavi*).

There is no definite article in Turkish, but definiteness of the object is implied when the accusative ending is used. Turkish nouns decline by taking case-endings, as in Latin. There are six noun cases in Turkish, with all the endings following vowel harmony. The plural marker *-ler* immediately follows the noun before any case or other affixes (e.g. *köylerin* 'of the villages').

The accusative case marker is used only for definite objects; compare (*bir*) *ağaç gördük* 'we saw a tree' with *ağaç gördük* 'we saw the tree'. The plural marker *-ler* is generally not used when a class or category is meant: *ağaç gördük* can equally well mean 'we saw trees [as we walked through the forest]'

– as opposed to *ağaçları gördük* ‘we saw the trees [in question]’.

Additionally, nouns can take suffixes that assign person: for example *-imiz*, ‘our’. With the addition of the copula (for example *-im*, ‘I am’) complete sentences can be formed. The interrogative particle *mi* immediately follows the word being questioned: *köye mi?* ‘[going] to the village?’, *ağaç mı?* ‘[is it a] tree?’.

The Turkish personal pronouns in the nominative case are *ben* (1s), *sen* (2s), *o* (3s), *biz* (1pl), *siz* (2pl, or formal/polite 2s), and *onlar* (3pl). They are declined regularly with some exceptions: *benim* (1s gen.); *bizim* (1pl gen.); *bana* (1s dat.); *sana* (2s dat.); and the oblique forms of *o* use the root *on*. All other pronouns (reflexive *kendi* and so on) are declined regularly.

Two nouns, or groups of nouns, may be joined in either of two ways:

- definite (possessive) compound (*belirtili tamlama*)
- indefinite (qualifying) compound (*belirtisiz tamlama*)

There is a third way of linking the nouns where both nouns take no suffixes (*taksız tamlama*).

Turkish adjectives are not declined.

Turkish verbs indicate person. They can be made negative, potential (‘can’), or impotential (‘cannot’). Furthermore, Turkish verbs show tense (present, past, future, and aorist), mood (conditional, imperative, inferential, necessitative, and optative), and aspect. Negation is expressed by the infix *-me* immediately following the stem. Almost all Turkish verbs are conjugated in the same way, most notable exception being the irregular and defective verb *i-*, the Turkish copula (corresponding to English ‘to be’), which can be used in compound forms (the shortened form is called an enclitic).

There are 9 simple and 20 compound tenses in Turkish. 9 simple tenses are simple past (*di'li geçmiş*), inferential past (*miş'li geçmiş*), present continuous, simple present (*aorist*), future, wish, demand, necessitative (‘must’) and order. There are three groups of compound forms. Story (*hikaye*) is the witnessed past of the above forms (except command), rumor (*rivayet*) is the unwitnessed past of the above forms (except inferential past and command), conditional wish (*koşul*) is the conditional form of the first five basic tenses.

There are also so-called combined verbs, which are created by adding certain verbs (like *bil* or *ver*) to the stem of the verb. *Bil* is the sufficiency mood. It is the equivalent of English auxiliary verbs ‘able to’, ‘can’ or ‘may’. *Ver* is the swiftness, *kal* is the perpetuity and *yaz* is the approach (‘almost’) moods. Thus while *gittin* is ‘you went’, *gidebildin* is ‘you could go’ and *gidiverdin* is ‘you went swiftly’. The tenses of the combined verbs are the same as the other verbs.

Turkish verbs have attributive forms, including present (with the ending *-en*), future (*ecek*), indirect/inferential past (*-miş*), and aorist (*-er* or *-ir*). These forms can function as either adjectives or nouns: *oynamayan çocuklar* ‘children who do not play’, *oynamayanlar* ‘those who do not play’; *okur yazar* ‘reader-writer = literate’, *okur yazarlar* ‘literates’.

The most important function of attributive verbs is to form modifying phrases equivalent to the relative clauses found in most European languages. The attributive forms used in these constructions are the future (*-ecek*) and an older form (*-dik*), which covers both present and past meanings.

Word order in simple Turkish sentences is generally subject–object–verb, as in Korean and Latin, but unlike English. In more complex sentences, the basic rule is that the qualifier precedes the qualified: this principle includes, as an important special case, the participial modifiers discussed above. The definite precedes the indefinite: thus *çocuğa hikâyeyi anlattı* ‘she told the child a story’, but *hikâyeyi bir çocuğa anlattı* ‘she told the story to a child’.

It is possible to alter the word order to stress the importance of a certain word or phrase. The main rule is that the word before the verb has the stress without exception. For example, if one wants to say “Hakan went to school” with a stress on the word “school” (*okul*, the indirect object) it would be *Hakan okula gitti*. If the stress is to be placed on *Hakan* (the subject), it would be *Okula Hakan gitti* which means ‘it’s Hakan who went to school’.

The source of this section is [Wikipedia](#).

17.2.2 Predication

According to [Göksel and Kerslake \(2005\)](#) the predicate of a simple sentence, or of the main clause in a complex sentence, is described as finite. According to the type of predicate they have, sentences in Turkish are divided into two main groups, verbal sentences and nominal sentences. Verbal sentences’ predicates are finite verbs.

- (104) *Bu gün evde kal-a-ma-m.*
 - - - stay-PSB-NEG.AOR-1SG

‘I can’t stay at home today.’

Nominal sentences’ predicate either does not contain an overt verb at all or whose verb is one of the forms of the copula (*ol-* ‘be’, ‘become’, ‘exist’ or *-(y)-* ‘be’).

- (105) *Necla öğretmen.*

‘Necla is a teacher.’

Turkish nominal sentences are of two kinds: linking and existential. **Linking sentences** correspond to the pattern *x is y*, and contain the following:

1. a subject (if overtly expressed)
2. a subject complement as (part of) the predicate
3. a copular marker (suffixed to, or immediately following, the subject complement). In present-tense sentences which are not aspectually or modally marked the copula has no overt expression. Person/number marking of the predicate is attached to the copular marker, if there is one, otherwise to the subject complement
4. (optionally) one or more adverbials.

- (106) *Ben osrada öğretmen-di-m.*
 SUBJ. ADV. SUBJ.COMPL.-COPULAR.MARKER-PERS.MARKER
 I at.that.time teacher-P.COP-1SG

‘I was a teacher at the time.’

The function of the subject complement is to provide some kind of description of the subject, such as identification, characterization, location, state of belonging, etc. The subject complement may be an adjectival, a noun phrase or a postpositional phrase. If a noun phrase, it may be marked for any grammatical case except the accusative case:

(107) *Biraz yorgundum.* (*Adjectival*)

‘I was rather tired.’

Siz çok iyi bir doktorsunuz. (*Non-case-marked noun phrase*)

‘You’re a very good doctor’

Sōzüm sanaydi. (*Dative-marked noun phrase*)

‘My words were for you.’

Herkes ona karşıymış. (*Postpositional phrase*)

‘Apparently, everyone is/was against him/her.’

Existential sentences merely assert the existence or presence of a subject. The statement is usually made in relation to either (i) a location in time or space, or (ii) a possessor. There are thus two kinds of existential sentence: locative and possessive.

Locative existential sentences are of the type there is an *x* (*in y*) or *x has y*. The basic constituents of a locative existential sentence (listed in the order in which they occur) are:

1. at least one adverbial of place or time
2. the subject (shown in bold below)
3. one of the two expressions *var* ‘present/existent’ and *yok* ‘absent/nonexistent’. (These can only be used in predicative function.)
4. copular marker (as in linking sentences, not present in the case of presenttense sentences which are not marked for aspect or modality).
5. 1st or 2nd person marking if required. (*Yok* but not *var* may also take a 3rd person plural marker.)

(108) *Buzdolabın-da iki şişe bira var.*
fridge-LOC two bottle beer existent

‘There are two bottles of beer in the fridge.’

O gün ben yok-tu-m.
that day I non-existent-P.COP-1SG

‘I wasn’t [there] on that day.’

Possessive existential sentences express the concept *x has y*. Their basic constituents are:

1. a genitive-possessive construction or a possessive-marked noun phrase, which is the subject
2. (optionally) one or more adverbials
3. *var* ‘present/existent’ or *yok* ‘absent/non-existent’
4. a copular marker (not overtly expressed in the case of present-tense sentences which are not marked for aspect or modality).

The adverbials, if there are any, are generally placed before the possessive-marked noun phrase. They may either precede or follow the genitive-marked possessor constituent (if there is one).

- (109) *Ayten-in İstanbul'da iki arkadaş-var.* / *İstanbul'da Ayten-in iki arkadaş-var.*

‘Ayten has two friends in Istanbul.’

O gün paramız yoktu.

‘We had no money that day.’

Possessive existential sentences are used for expressing possession or relations between people and things which are either of a permanent nature or are considered as such. The possessed constituent in these constructions cannot be definite or specific. Possessive existential sentences are mainly used for expressing the following: (i) familial and other personal relations, (ii) part-whole relations, (iii) authorship or (vi) ownership. Locative existential sentences, on the other hand, denote contiguity between persons or things at a particular time.

17.2.3 Possession

According to Göksel and Kerslake (2005) in Turkish the possessive suffixes correspond to the six grammatical persons. A noun phrase marked with a possessive suffix (except where this is a 3rd person suffix functioning as a compound marker or pronominalizer) is understood as denoting a person or thing that is possessed. The possessive suffix indicates only whether the possessor is 1st, 2nd or 3rd person, singular or plural.

- (110) *Arkadaş-lar-iniz ne kadar kalacaklar?*

‘How long are your friends going to stay?’

Ahmet oda-sın-ıarryordu. Numara-sıakl-in-da kalmamıştı.

‘Ahmet was looking for his room. Its number was no longer in his head.’

For the description of possessive existential sentences see Section 17.2.2.

17.2.4 Imperative

According to Göknel (2013) direct orders are given to a second person by using a verb root, a verb stem or a verb frame without using any suffixes.

- (111) *Bura-/y/a gel.*

‘Come here’

Kuş-lar-a bak.

‘Look at the birds.’

Süt-ün-ü iç.

‘Drink your milk.’

Pencere-den bak.

‘Look out of the window.’

Bir fincan kahve buyur!

‘Have a cup of coffee!’

Eğlen-me-en-e bak!

‘Have a nice time!/Enjoy yourself!’

Orders are given to the second person as a rule. However, an order may also be given to the third person indirectly. A speaker gives orders to the second person to be transferred to a third person. The last syllable of an imperative sentence is primarily stressed and dropped sharply. The orders that are given with the verb *ol* and *et* ‘be’ are widely used in both English and Turkish. In such sentences the primarily stressed syllables are the last syllables of the adjectives and adverbials.

17.2.5 Interrogative

According to Göknel (2013) There are two kinds of interrogative words in Turkish: Simple interrogative words like *kim?* (who?), *ne?* (what?), *nasıl?* (how?), *nıçın?* (why?), and the simple interrogative words that are followed by some inflectional morphemes such as *kim-sin?* (who?), *kim-im?* (who?), *kim-iz?* (who?), *kim-i?* (whom?), *kim-e?* (to whom?), *kim-den?* (from whom?), *kim-le?* (with whom?), *kim-de?* (?), *kim-in?* (whose?), *ne/y/-le?* (how?), (with what instrument?), *ne-den?* (why?), *nere-/y/e?* (where?), *nere-de?* (where?), *nere-den?* (from where?).

The interrogative sentences having the question words above are pronounced with a rising intonation both at the end of the interrogative sentences, and after the people or things that the question words are inquiring.

In order to make up Turkish sentences containing one of the interrogative words above, one can put one of these words in a positive or negative sentence without changing its order. In other words, one can use such interrogative words in Turkish positive or negative sentences without changing their positive or negative sentence structures.

The words that change positive and negative sentences into *yes-no* interrogative sentences differ from one tense to another.

17.3 Writing system, transcription

Turkish is written using a Latin alphabet introduced in 1928 by Atatürk to replace the Ottoman Turkish alphabet, a version of Perso-Arabic alphabet. The Ottoman alphabet marked only three different vowels – long ā, ū and ī – and included several redundant consonants, such as variants of z (which were distinguished in Arabic but not in Turkish). The omission of short vowels in the Arabic script was claimed to make it particularly unsuitable for Turkish, which has eight vowels.

The reform of the script was an important step in the cultural reforms of the period. The task of preparing the new alphabet and selecting the necessary modifications for sounds specific to Turkish was entrusted to a Language Commission composed of prominent linguists, academics, and writers. The introduction of the new Turkish alphabet was supported by public education centers opened throughout the country, cooperation with publishing companies, and encouragement by Atatürk himself, who toured the country teaching the new letters to the public. As a result, there was a dramatic increase in literacy from its original Third World levels.

The Latin alphabet was applied to the Turkish language for educational purposes even before the 20th-century reform. Instances include a 1635 Latin-Albanian dictionary by Frang Bardhi, who also

incorporated several sayings in the Turkish language, as an appendix to his work (e.g. *alma agatsdan irak duschamas* – “An apple does not fall far from its tree”).

Turkish now has an alphabet suited to the sounds of the language: the spelling is largely phonemic, with one letter corresponding to each phoneme. Most of the letters are used approximately as in English, the main exceptions being ⟨c⟩, which denotes [dʒ] (⟨j⟩ being used for the [ʒ] found in Persian and European loans); and the undotted ⟨i⟩, representing [u]. As in German, ⟨ö⟩ and ⟨ü⟩ represent [ø] and [y]. The letter ⟨ğ⟩, in principle, denotes [y] but has the property of lengthening the preceding vowel and assimilating any subsequent vowel. The letters ⟨ş⟩ and ⟨ç⟩ represent [ʃ] and [tʃ], respectively. A circumflex is written over back vowels following ⟨k⟩, ⟨g⟩, or ⟨l⟩ when these consonants represent [c], [ɟ], and [l] – almost exclusively in Arabic and Persian loans. An apostrophe is used to separate proper nouns from inflectional suffixes: e.g. *İstanbul'da* “in Istanbul” (but not from derivational suffixes, e.g. *İstanbullu* “from/of Istanbul”).

The Turkish alphabet consists of 29 letters (q, x, w omitted and ç, ş, ğ, ı, ö, ü added); the complete list is:

a, b, c, ç, d, e, f, g, ğ, h, ı, i, j, k, l, m, n, o, ö, p, r, s, ş, t, u, ü, v, y, and z (Note that capital of i is İ and lowercase I is ı.)

The source of this section is [Wikipedia](#).

17.4 Previous research on the language

The Turkish Language Institution [TDK](#) is the official regulatory body of the Turkish language, founded on July 12, 1932 by the initiative of Atatürk and headquartered in Ankara, Turkey. Besides acting as the official authority on the language (without any enforcement power), contributes to linguistic research on Turkish and other Turkic languages, and is responsible for publishing the official dictionary of the language, *Güncel Türkçe Sözlük*. The institution heads academic linguistic research in Turkey into the Turkish language and its sister Turkic languages of Central Asia. It publishes [Türkçe Sözlük](#), the official Turkish dictionary, and [Yazım Kılavuzu](#), the Turkish writing guide, in addition to many other specialized dictionaries, linguistics books and several periodicals.

The institution, in addition to maintaining *Güncel Türkçe Sözlük* and *Yazım Kılavuzu* has published more than 850 linguistics related books, mainly consisting of studies on Turkic languages, specialized dictionaries, philological books, and works of literature.

The 2005 edition of *Güncel Türkçe Sözlük*, the official dictionary of the Turkish language published by Turkish Language Association, contains 104.481 entries, of which about 86% are Turkish and 14% are of foreign origin.

In the series of [Typological Studies in Language, Studies in Turkish Linguistics](#) ([Slobin and Zimmer \(1986\)](#)) deals with the morphological, syntactic, semantic and discourse-based, synchronic and diachronic aspects of the Turkish language. Although an interest in morphosyntactic issues pervades the entire collection, the contributions can be grouped in terms of relative attention to syntax, semantics and discourse, and acquisition.

[Ay \(2009\)](#) contains 48 papers presented at the Fourteenth International Conference on Turkish Linguistics, held by Ankara University in August 6-8, 2008. The contributions to this conference cover a wide range of topics in theoretical, descriptive and applied linguistics relating to Turkish and Turkic languages in discussing a great variety of issues related to phonology and phonetics, morphology, syntax and semantics, pragmatics and discourse, language acquisition, language contact, and applied linguistics, as they have been grouped in this volume. Although the main focus of the volume is on

Turkish linguistic issues, there are also a number of articles in different modern linguistic frameworks dealing with Turkic languages and Turkish dialects. The book will be appealing to anyone interested in current issues in theoretical linguistics as well as those who are working on Turcology, linguistic typology, contact linguistics, and applied linguistics.

17.4.1 Grammars

[Wikipedia](#) has an overview of Turkish grammar. In the section of References a lot of Grammars, Dictionaries and other resources of the Turkish language are listed.

Here I list the most important grammars written in English. A classic, still used to teach Turkish grammar in many universities is [Underhill \(1976\)](#). It is available on [scribd](#). The most recent comprehensive grammar in English is [Göksel and Kerslake \(2005\)](#). An other standard work on the language throughout the English-speaking world is [Lewis \(2000\)](#). [Swift \(1997\)](#) is written for the linguist of the linguistically-oriented intermediate student of modern turkish. [Göknel \(2013\)](#) is a traditional grammar books for academic usage. It is available in [PDF](#).

17.5 Data and sources

The World Atlas of Language Structures ([WALS](#)) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars). There are materials about [Turkish](#).

[Omniglot](#) is an encyclopedia of writing systems and languages. It can be used to learn about languages, to learn alphabets and other writing systems, and to learn phrases in many languages. There is also advice on how to learn languages. There are some information about [Turkish](#).

[OLAC](#), the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources. OLAC Archives contain over 100,000 records, covering resources in half of the world's living languages including [Turkish](#).

17.5.1 Basic vocabulary

[Swadesh lists](#) were originally devised by the linguist Morris Swadesh. In the 1940s to 1950s, Swadesh developed word lists of body parts, verbs, natural phenomena, in order to compute the relationships of languages, and in particular their age. A Swadesh list may also be useful to achieve knowledge of some universal terms in other languages. This is because, for basic communication, knowledge of vocabulary is more important than knowledge of grammar and syntax. Sometimes it is even possible to achieve (very) basic communication skills with no knowledge of the target language syntax whatsoever. [Turkish Swadesh list](#) is provided as well.

[Turkeytravelplanner](#) lists the 100 most important words in Turkish. [turkishlanguage.co.uk](#) provides a short introduction to Turkish pronunciation. It lists the most important words alphabetically and sorted into categories as well.

Learning Turkish There are some webpages for Turkish learners containing an overview of the language, useful phrases, grammar and pronunciation like [turkisbasics](#), [turkishclass](#) and [ielanguages.101languages](#) lists Turkish radio stations as well, because listening to radio is a great way to develop an ear for a foreign language.

17.5.2 Dictionaries

Paper editions

Three classic Turkish-English-Turkish dictionaries: [Kornrumpf \(1989\)](#); [Redhouse \(1991\)](#); [Serap et al. \(2001\)](#).

Online dictionaries

There are a lot of online Turkish dictionaries as [en.bab.la](#), [cambridge](#), [lexilogos](#), [babylon turkishdictionary.net](#), [tureng](#), and [ectaco](#).

Scraping Searchbox dictionaries with an easily parametrizable search interface like [en.bab.la](#), [cambridge](#), [lexilogos](#), [babylon turkishdictionary.net](#), [tureng](#), and [ectaco](#) are scrapable with the help of a world-list which contains the lexical forms of Turkish words. If no more complete list is available, [Turkish Swadesh list](#) can be used for this purpose.

17.5.3 Corpora

[TS Corpus](#) is a Free&Independent Project that aims building Turkish corpora, developing Natural Language Processing tools and compiling linguistic datasets. It contains 551.411.894 tokens in 7 subcorpora (TS Corpus v2, TS Wikipedia Corpus, TweetS Corpus, Abstract Corpus, TS Gezi Corpus, Constitution Corpus, Idioms&Proverbs Corpus).

[Turkish National Corpus](#) (TNC) with a size of 50 million words, is a balanced and a representative corpus of contemporary Turkish. It consists of samples of textual data across a wide variety of genres covering a period of 20 years (1990-2009). Written component consists of texts produced in different domains on various topics. Transcriptions from spoken data constitute 2% of TNC's database, which involves spontaneous, every day conversations and speeches collected in particular communicative settings.

[METU](#) (Middle East Technical University) Turkish Corpus is a collection of 2 million words of post-1990 written Turkish samples. A subset of the corpus is used in METU-Sabancı Turkish Treebank. METU Turkish Corpus is XCES tagged at the typographical level. The distribution of the corpus also includes a workbench and related publications. The words of METU Turkish Corpus were taken from 10 different genres. At most 2 samples from one source is used; each sample is 2000 words or the sample ends when the next sentence ends.

The [Spoken Turkish Corpus](#) presents a total of 18 selection of audio recordings and archives from Radyo ODTÜ. The recordings in std_demo have been analyzed only for basic transcription and certain discursive features.

Wikipedia [Wikipedia](#) offers free copies of all available content to interested users. [Wikimedia](#) is a global movement whose mission is to bring free educational content to the world. The content of Wikimedia is dumped and available as well. [Turkish Wikipedia](#) contains 283.839 articles as of 21/10/2016.

These dumps are available in Turkish in 21/10/2016:

- [Turkish Wikipedia](#)
- [Turkish Wikibooks](#)

- [Turkish Wiktionary](#)
- [Turkish Wikisource](#)
- [Turkish Wikimedia](#)
- [Turkish Wikiquote](#)
- [Turkish Wikinews](#)

Bible The text of the Bible in Turkish is available from a lot of sources like [wordproject](#), [wordplanet](#), [worldbibles](#) and [sacred-text](#).

Bilingual The main goal of [The English-Swedish-Turkish Corpus](#) is to promote research and teaching in the Turkish language. More specifically, the aim is to build a language resource for Turkish, Swedish and English allowing contrastive studies between the involved languages. The language resource consists of linguistically analyzed parallel texts that are linked to each other in the three languages. The corpus consists of original texts and their translations from Turkish to Swedish and English, and from Swedish and English to Turkish. The corpus is organized as a parallel corpus, where the texts, paragraphs, sentences and words are linked to each other. The corpus is built semi-automatically by using a basic language resource kit (BLARK) for the particular languages. The texts are linguistically analyzed with morphological features and part-of-speech as well as with dependency structures. The parallel corpus is intended to be used in research, teaching and applications such as machine translation.

[Turkish-English Parallel Corpus](#) contains numerous parallel corpora. (1) SETimes is parallel corpus of news articles in the Balkan languages, originally extracted from [South East European Times](#), contains 207K Turkish-English parallel sentences. (2) BU parallel corpus contains 688K parallel sentences and 5 million words and built by using literature sources. (3) WebCorpus contains 131K Turkish-English parallel sentences collected from web. (4) The parallel texts collected from [yeminlisozluk.com](#) which is an online translation memory interface includes Turkish and English translation examples. The Corpus contains 175K 1-1 sentence alignments for English and Turkish Languages. (5) OpenSubtitles'2011 is a collection of documents from <http://www.opensubtitles.org/>. The total number of Turkish-English parallel sentences: 17M. (6) Academik Corpus is generated from bilingual abstracts of academical thesis which are collected from the database of The Council of Higher Education. It contains 1.4K parallel sentence pairs.

United Nations Human Rights Office of the High Commissioner [OHCHR](#) worldwide collection of materials on the Universal Declaration of Human Rights (UDHR), including various resources developed by governmental and non-governmental organizations both on the occasion of the Declaration's 50th Anniversary (1998) and prior to/after the Anniversary year. The collection is unique in the world and comprises more than 400 items. Since UDHR is the most translated document – and it is available in [Turkish](#) – it can be used as a multilingual corpus.

17.5.4 News portals

[101languages](#) lists the top Turkish language newspapers for example [Hürriyet](#), [Milliyet](#), [Radikal](#), [Vatan](#) and [Posta](#). [Kibris](#) is a Turkish language newspaper from Cyprus.

17.5.5 Contact person

[Turkish Studies](#) at the Department of Near Eastern Studies, University of Michigan provides graduate and undergraduate programs in arabic and middle eastern studies. This website is maintained by Mehmet Süreyya Er, to whom comments and suggestions may be addressed at sureyya@umich.edu.

Founded and incorporated in the District of Columbia in 1982, the Institute of Turkish Studies ([ITS](#)) is the only non-profit, private educational foundation in the United States exclusively dedicated to the support and development of Turkish Studies in American higher education. The fundamental purposes and objectives of the Institute are: (1) to support individual scholars, especially younger members of the academic profession in the United States, for advanced research on Turkish history and culture as well as contemporary political, social, and economic developments in Turkey; (2) to assist American universities to develop their library resources, programs of study, scholarly conferences, and outreach activities in the field of Turkish Studies; (3) to support the publication of books and journals that contribute to American scholarship on Turkey and broaden the understanding and knowledge of Turkish history, society, politics, and economics in the United States; (4) to promote better understanding of Turkish politics, economy, and society through conferences and lecture series. To contact ITS please send email to itsdirector@turkishstudies.org.

17.6 Computational tools

17.6.1 Language identification

There are several language identification tools which supports Turkish such as [elastic](#). [CLD2](#) Compact Language Detector 2, [TextCat](#), [Lingua::LanguageGuesser](#) and [langid.py](#) supports Turkish as well.

17.6.2 Tokenizer

[Engerek](#) is an Apache v2.0 licensed Python library that brings together commonly used natural language processing techniques, with a special focus on Turkish text. Our aim is to provide a clean and consistent interface for existing open source Turkish processing libraries. It has tokenizer and stemmer features and it can deascify Turkish text that is written only by ASCII characters.

[Ivory](#) supports tokenization on a lot of languages including Turkish. Is has a stemming feature as well.

17.6.3 Stemmer

The tokenizer tools mentioned above have stemmer features as well. Newt to them there are some individual stemmer tools as [resha stemmer](#), which is a fast and “less aggressive” stemmer for Turkish written in Java. It uses a stem dictionary using a statistical language model based on morpheme n-grams. So it returns the most possible stem for a word without considering the neighbor words. It contains more than 1.1 million word-stem pairs.

[Snowball](#) is a small string processing language designed for creating stemming algorithms for use in Information Retrieval. This site describes Snowball, and presents several useful stemmers which have been implemented using it. [Snowball stemmer](#) is an external contribution of Snowball. Find the details of the Turkish stemmer in [Gülşen and Eşref \(2004\)](#).

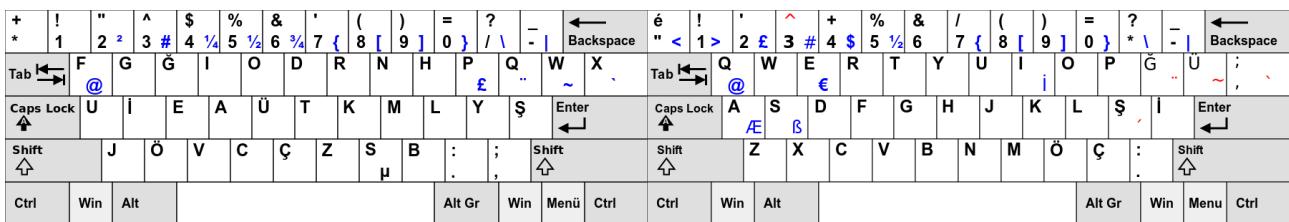


Figure 17.4: Turkish-F Keyboard Layout

Figure 17.5: Turkish-Q Keyboard Layout

17.6.4 Spell checker

Some free on-line spell checker sites are available in Turkish like [SpellChecker.net](#) and [SpellStar](#).

Microsoft provides [spell checker tools for Turkish](#) in its products such as Word, Excel PowerPoint, Access and so on. LibreOffice has a [Turkish Spellcheck Dictionary extension](#) as well.

17.6.5 Phrase level and higher tools

[ITU](#) provides the Turkish NLP Tools and APIs developed by our [Natural Language Processing group](#) at [Istanbul Technical University](#). [Gülşen \(2014\)](#) summarizes the main phases of the tool ([PDF](#)). The platform (available operates as a SaaS (Software as a Service) and provides the researchers and the students the state of the art NLP tools in many layers: preprocessing, morphology, syntax and entity recognition. The users may communicate with the platform via three channels: 1. via a user friendly web interface, 2. by file uploads and 3. by using the provided Web APIs within their own codes for constructing higher level applications. In the paper [Gülşen \(2014\)](#) presents their new web service which provides both a whole Turkish NLP pipeline and its atomic NLP components for stand-alone usage, namely Tokenizer, Deasciifier, Vowelizer, Spelling Corrector, Normalizer, isTurkish, Morphological Analyzer, Morphological Disambiguator, Named Entity Recognizer and Dependency Parser.

[LingPipe](#) tool kit for processing text using computational linguistics. LingPipe is used to do tasks like finding the names of people, organizations or locations in news, automatically classification of Twitter search results into categories, suggesting correct spellings of queries. Next to other languages it supports Turkish.

17.6.6 End-user support

Although modern Turkish is written in the Roman alphabet, it is encoded as Unicode or **ISO-8859-9** and requires special font and keyboard support separate from languages like Spanish and French. Modern versions of many fonts such as Times New Roman, Arial, Verdana, Tahoman Times CE (Mac OS X) or Palatino (Mac OS X) are Unicode fonts and contain the letters needed for this language.

Keyboard layout Turkish has two different keyboard layouts: Turkish-F and Turkish-Q, both of which are quite commonly used. Turkish-F keyboard layout was designed in 1955 by İhsan Yener. During its design, the Turkish Language Academy (TDK) investigated letter frequencies in Turkish and used this statistical basis to design the Turkish-F keyboard. It provides a balanced distribution of typing effort between the hands - 49% for the left hand and 51% for the right.

Windows If you are using a recent version of Microsoft Word (2003+), you can use the following ALT key plus a numeric code can be used to type a Latin character (accented letter or punctuation symbol) in any Windows application.

Cns	name	ALT codes
Ç	capital C cedille	ALT+0199
ç	lower c cedille	ALT+0231
Ğ	capital G breve	ALT+0286
ğ	lower g breve	ALT+0287
Ş	capital S cedille	ALT+0350
ş	lower s cedille	ALT+0351

Table 17.1: Word ALT codes for Turkish characters

I	name	ALT codes
ı	dotted capital I	ALT+0199
ç	dotless lower i	ALT+0231
Ö		ALT+0286
ö		ALT+0287
Ü		ALT+0350
ü		ALT+0351

Table 17.2: Turkish Vowels

Users with older versions of Windows or not using may need to use the Character Map utility.

[Code Page 1254](#) Windows Latin 5 (Turkish) contains all characters used in Turkish.

Macintosh The new CE fonts for OS X (Times CE) as well as new versions of Palatino, Times New Roman, Arial and others now contain Turkish letters. It is recommended that you transition to these fonts whenever possible. Apple now has several Turkish keyboards, includig a QWERTY keyboard and QWERTY PC keyboard, but they only work for Unicode Aware applications. If you are working with a Unicode aware application you can activate the U.S. Extended and use the following option codes along with the older accent codes.

Browser and Font Setup All modern browsers support this script like [Firefox](#), [Internet Explorer](#), [Safari](#), [Opera](#).

Cns	name	ALT codes
Ç	capital C cedille	Shift+Option+C
ç	lower c cedille	Option+C
Ğ	capital G breve	Option+V,Shift+G
ğ	lower g breve	Option+V,g
Ş	capital S cedille	Option+C,Shift+S
ş	lower s cedille	Options+C,S

Table 17.3: Turkish Consonants

Bibliography

- Sıla Ay. *Essays on Turkish Linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, August 6-8, 2008*. Turcologica Series. Harrassowitz, 2009. ISBN 9783447060592. URL <https://books.google.hu/books?id=N9kggNEZGdoC>.
- Yüksel Göknel. *Turkish Grammar Updated Academic Edition*. Ege Basım, 2013.
- A. Göksel and C. Kerslake. *Turkish: A Comprehensive Grammar*. Comprehensive Grammars. Routledge, 2005.
- Eryiğit Gülsen. ITU Turkish NLP Web Service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), Gothenburg, Sweden, April 2014*, 2014.
- Eryiğit Gülsen and AdalıEşref. An Affix Stripping Morphological Analyzer for Turkish. In *Proceedings of the IAESTED International Conference ARTIFICIAL INTELLIGENCE AND APPLICATIONS, February 16-18, 2004, Innsbruck, Austria*, 2004.
- H. J. Kornrumpf. *Langenscheidt's Universal Turkish Dictionary: Turkish-English/English-Turkish*. Langenscheidt Publishers, 1989.
- Geoffrey Lewis. *Turkish Grammar, Second Edition*. Oxford University Press, 2000.
- James W. Sir Redhouse. *Redhouse yeni Türkçe-İngilizce sözlük = New Redhouse Turkish-English Dictionary*. Redhouse Yayınevi, İstanbul, 12 edition, 1991.
- Bezmez Serap, Blakney Richard, Brown C. H., and Aydın Nilüfer. *Redhouse büyük elsözlüğü : İngilizce -Türkçe, Türkçe -İngilizce = The Larger Redhouse Portable Dictionary : English-Turkish, Turkish-English*. Redhouse Yayınevi, İstanbul, 2001.
- Dan I. Slobin and Karl Zimmer, editors. *Studies in Turkish Linguistics*. Number 8 in Typological Studies in Language. John Benjamins Publishing Company, 1986.
- L.B. Swift. *A Reference Grammar of Modern Turkish*. Uralic and Altaic Series. Taylor & Francis Group, 1997.
- R. Underhill. *Turkish Grammar*. MIT Press, 1976.

Chapter 18

Uzbek (Nikolett Mus)

Contents

18.1 Demography and ethnography	271
18.2 Main typological and syntactic features	274
18.3 Writing system, transcription	277
18.4 Previous research on the language	277
18.5 Data and sources	278
18.6 Computational tools	280
Bibliography	282

Introduction

The present chapter describes certain features of Uzbek, a member of the Turkic language family. Following an overview of the ethnolinguistic situation of the language and its speaking community (see 18.1), a typological overview will be provided (see 18.2). After surveying the writing system (and linguistic transcriptions; see 18.3), the results of previous researches on the Uzbek language will be introduced (see 18.4). The final part of the chapter details the various offline and online linguistic data repositories (such as dictionaries, corpora and news portals; see 18.5), while the last section reviews the available computational tools for Uzbek (see 18.6).

18.1 Demography and ethnography

18.1.1 Name variants

The Uzbek language, also referred to with the endonyms as O’zbek tili or O’zbekcha, belongs to the Eastern Turkic (or Karluk) branch of the Turkic language family (see [Wikipedia](#)). The ISO 639-3 code of the Uzbek *macrolanguage* is **uzb** (cf. [SIL.org](#)). There are two variations of Uzbek, with different ISO codes: one of the variants is Northern Uzbek, i.e. **uzn** (see [SIL.org](#)), while the other is Southern Uzbek, i.e. **uzs** (see [SIL.org](#)).

18.1.2 Geographic spread

The Uzbek language is spoken in Uzbekistan and elsewhere in Central Asia, mainly in Afghanistan, Pakistan, and Turkey. Figure 18.1 illustrates the Uzbek speaking territories in Uzbekistan and Turk-

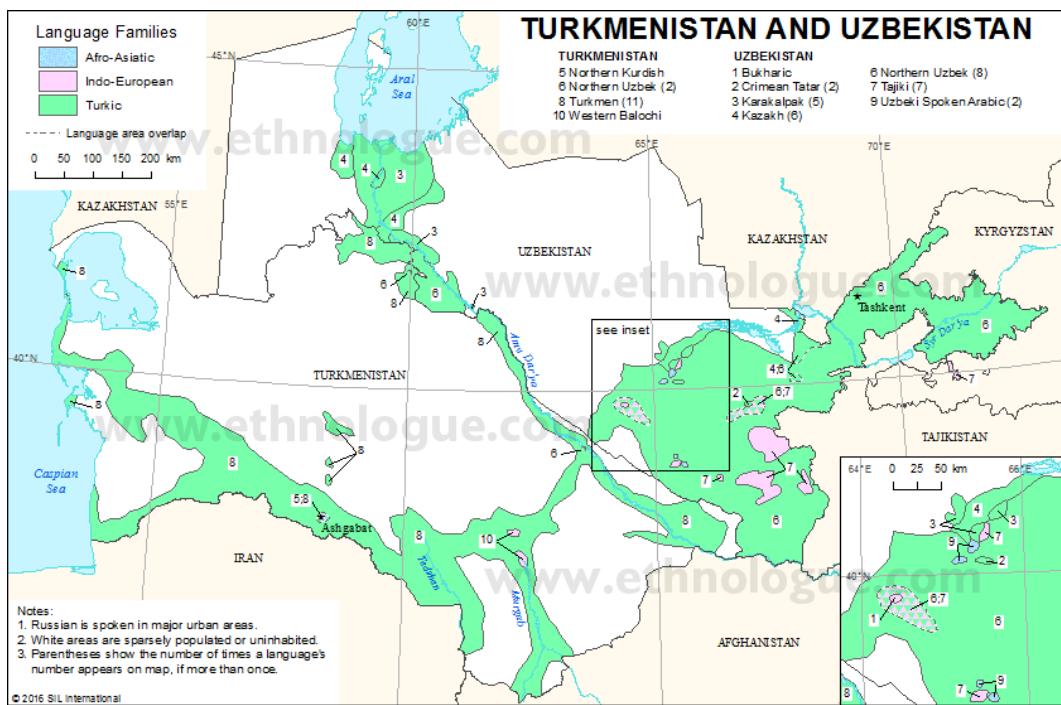


Figure 18.1: The linguistic diversity of Turkmenistan and Uzbekistan (source:Ethnologue)



Figure 18.2: The Afghanistan (source:The Joshua Project)

menistan. As it is seen, in these countries the Northern variety of the Uzbek language is spoken (see [Ethnologue](#)).

The Southern Uzbek language (uzs) is spoken in the Northern part of Afghanistan (see Map 18.2 and cf. [Ethnologue](#)).

18.1.3 Speaker populations

Uzbek is spoken by c. 27 million speakers. The Northern Uzbek (uzn) language is official in Uzbekistan, where its EGIDS (Expanded Graded Intergenerational Disruption Scale) level is 1 (National). Consequently, the language is used in education, work, mass media, and government at the national level in Uzbekistan. Figure 18.3. shows the position of the Northern Uzbek language within the cloud of languages, see the explanation in [section](#).

In Afghanistan, the Southern Uzbek language has a 2 (Provincial) EGIDS-status, i.e., the language is used in education, work, mass media, and government within major administrative subdivisions of a nation (see Figure 18.4).

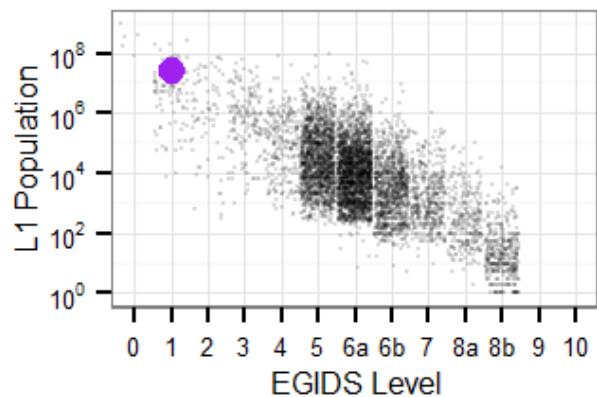


Figure 18.3: The EGIDS level for Northern Uzbek (source:[Ethnologue](#))

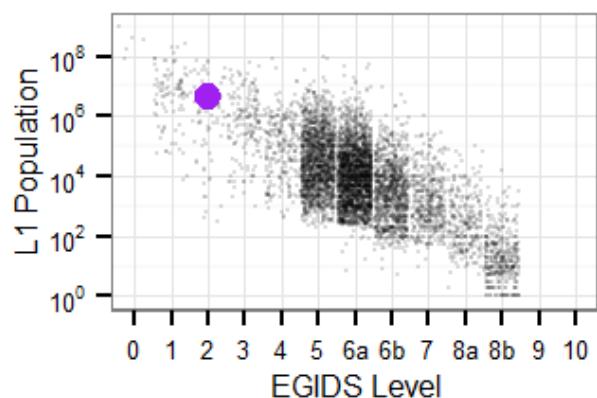


Figure 18.4: The EGIDS level for Southern Uzbek (source:[Ethnologue](#))

Table 18.1: The Uzbek speakers

Country	Number of L1 speakers
Uzbekistan	21,000,000
Afghanistan	3,400,000
Tajikistan	900,000
Kyrgyzstan	800,000
Kazakhstan	500,000
Turkmenistan	300,000
Russia	300,000

According to the corresponding [Wikipedia](#) entry, the estimated number of Uzbek speakers across countries is summarized in Table (18.1).

18.1.4 Dialect situation

The Uzbek language is further devided into many dialects that vary widely from region to region. According to [Wikipedia](#), the most-widespread dialects are the following:

- the **Tashkent** dialect
- the **Afghan** dialect
- the **Ferghana** dialect
- the **Khorezm** dialect
- the **Chimkent-Turkestan** dialect
- the **Surkhandarya** dialect

The Northern and the Southern Uzbek languages share many grammatical features. For instance, both variants are (basically) head-final, so their simple declarative clauses follow the SOV order. In these variants of the Uzbek language, the grammatical and adnominal relations are expressed by suffixes and postpositions. The ordering of the nominal phrases is also head-final. The verbal predicate agrees with the subject. In the Southern variation, object agreement is also reported.

The two variants have been exposed to different linguistic contacts whose consequences can be detected on their grammatical structures. For instance, there is a prefix in the Southern dialect which appeared due to the strong contact with Dari (Persian/Farsi). In addition, the historical vowel harmony is lost from the Northern Uzbek and its vowel system now resembles that of Tajiki.

18.2 Main typological and syntactic features

18.2.1 Linguistic typology

As previously mentioned, Uzbek belongs to the Turkic language family. The Turkic languages are usually associated with rich suffix systems and strongly agglutinative grammatical characteristics. Normally, their structure follow a head final order both in phrase and clause levels.

Phonological level The phoneme inventory of Uzbek includes 7 vowels (high, mid, and low front vowels and high, mid, low-mid, and low back vowels) and 27 consonants. The [Descriptive Phonetics of Uzbek](#) provides information about the phonetics of Uzbek in detailed.

Morphological level Uzbek is an agglutinative language in which the grammatical relations are mainly expressed by suffixes attached to the nominal/verbal stems.

Morphosyntactic level The nouns in Uzbek are specified for number (singular and plural), case (nominative, accusative, genitive, dative, locative, ablative, equitative) and possession. All of these categories are indicated by suffixes. The verbs take agreement, mood, voice and inseparable tense-aspect markers (cf. [The Language Gulper-Uzbek](#)).

Syntactic level The basic word order in Uzbek clauses follows the SOV order. This order, however, can be changed due to informational structural changes, i.e. the topicalized element typically occupies the sentence initial position. In phrases, the modifiers precede their heads. The heads and their modifiers optionally agree in number. Subordination is expressed by non-finite and/or nominalized verb forms that can take nominal suffixes, such as case and possessive markers.

18.2.2 Predication

The verbal predicate appears clause finally and takes the agreement markers. In transitive clauses, the indefinite direct objects do not take the accusative marker.

- (112) *Abbos olma yeydi.*

Abbos apple eat

'Abbos eats an apple.'

- (113) *Abbos olma-ni yeydi.*

Abbos apple-ACC eat

'Abbos eats the apple.'

In the nominal predication, there is no overt copula in present tense. The predicate noun takes the bound form of the corresponding personal pronouns that shows agreement with the subject.

- (114) *Men o'zbek-man.*

1SG Uzbek-1SG.PRON

'I am Uzbek.'

18.2.3 Possession

In the adnominal possessive constructions, the possessor taking the suffix *-ning* precedes the possessed item. The possessed item is marked by the possessive suffix. The possessor can be pronominal and lexical.

Table 18.2: The paradigm of the Uzbek imperative

Pronouns	Imperative paradigms	English translation
Sen	ket	go (informal)
Siz	keting	go (formal)
U	ketsin	let him/her go
Ular	ketsinlar	let him/her go (respect)
Sizlar	ketinglar	go (plural)
Ular	ketishsin	let them go

- (115) *Ular-ning kitob-i*
 3PL.PRON book-3PL
 ‘their book’

If the possessed item is in plural, it takes the plural suffix.

- (116) *Ular-ning kitob-lar-i*
 3PL.PRON book-PL-3PL
 ‘their books’

The predicate possession is expressed by the existential construction.

- (117) *Ular-ning kitob-i bor*
 3PL.PRON book-3PL exist
 ‘They have a book.’

18.2.4 Imperative

The imperatives are usually used with 2nd and/or 3rd person referents. Then, the verb takes inflectional markers (see 18.2).

18.2.5 Interrogative

In Uzbek polar interrogatives, there is a suffix (*-mi*) that appears sentence finally attached to the final word. There is only one exception: with 2nd-person (plural or singular, formal or informal) predicates the interrogative suffix precedes the pronominal suffix.

- (118) *Men o'zbek-man-mi?*
 1SG Uzbek-1SG.PRON-Q
 ‘Am I Uzbek?’

- (119) *Siz o'zbek-mi-siz?*
 2SG.FM Uzbek-Q-SG.FM
 ‘Are you Uzbek?’

The wh- (or content) interrogatives in Uzbek are described as being *in-situ*, which means that the interrogative phrases do not occupy a dedicated position within the clause. Instead, they appear in the position in which their non-interrogative counterparts would appear. Consequently, there is no wh-movement in the wh-interrogatives.

- (120) *Bu kim?*
this who
'Who is this?'

- (121) *Bu Aziz.*
this Aziz
'This is Aziz.'

18.3 Writing system, transcription

The Uzbek language was emerged as a literary language in the 14th century. The early form of Uzbek, i.e. *Chagatai* was written with the Arabic script. The Latin alphabet replaced the Arabic script in 1929, which was in turn replaced by the Cyrillic alphabet in 1940. In 1993, a new Latin alphabet of Uzbek was introduced, that was changed in the following two years. Currently, this Latin alphabet is used in Uzbekistan for educational purposes and in the mass media. Additionally, the Cyrillic script is still used in Kazakhstan, Kyrgyzstan, Tajikistan, and Turkmenistan. Similarly, the Uzbek minority in Afghanistan uses the Arabic alphabet for writing the languages (see e.g. [About World Languages Uzbek](#)).

The character charts for both alphabets used with the Uzbek language is available at the [Omniglot](#).

18.4 Previous research on the language

The phonological system of the Uzbek language is discussed by e.g. [Sjoberg \(1962\)](#) in detail. The morphology of the language is addressed in [Matlatipov and Vetulani \(2009\)](#), a.o. Additionally, basic and structural grammars are provided by the following sources: [Sjoberg \(1963\)](#), [Reshetov \(1966\)](#), [Xodzhiev \(1997\)](#), [Boeschoten \(1998\)](#). For a detailed bibliography, see the OLAC entry for [Uzbek](#). Additionally, socio-linguistical attempts and results concerning the language planning of Uzbek are summarized by ([Fierman \(1991\)](#)). Finally, there is an onilne course of Uzbek called [Teach yourself Uzbek](#) that provides downloadable sources of Uzbek, such as coursebooks, dictionaries, flashcards, etc.

Furthermore, there are (ongoing) projects that attempt to describe and document the language. For instance, the [300 Languages](#) subproject of The Rosetta Project aims at archiving both variation, i.e. Uzbek as the macrolanguage, Northern Uzbek and Southern Uzbek, as well as, constructing a universal corpus of human language by collecting parallel texts and audio recordings. The recordings and translations cover Holy Quran.

Additionally, there is an archived text available at the [archive.org](#), which was designed for classroom and self-study of Uzbek by Peace Corps volunteers training to serve in Uzbekistan. The data consists of language and culture lessons on 11 topics: personal identification; classroom communication; conversation with hosts; food; getting and giving directions; public transportation; social situations; the communications system; medical needs; shopping; and speaking about the Peace Corps.

Research and control bodies There are philological studies concerning the Uzbek language at the [Academy of Sciences Republic of Uzbekistan](#), additionally the following Universities in Uzbekistan have Linguistic departments dealing with Uzbek: the [Karshi State University](#), the [Bukhara State University](#).

18.5 Data and sources

This section presents a collection of vocabularies, dictionnaires (both paper editions and online ones), and texts of Uzbek, as well as, it provides an overview of the news portals on the Uzbek language.

18.5.1 Basic vocabulary

The An Crúbadán project provides Uzbek sources written in Latin and in Cyrillic. The available data written in [Latin script](#) covers character trigrams, urls, word bigrams, word frequency lists based on 182,836 words. The [Cyrillic](#) version contains the same tools and outputs, however, it was created on the basis of a collection consisting of 40,893,319 words.

Additionally, there is an online word list available on the homepage of the 101 languages (see [Uzbek 101](#)).

The website entitled [Introduction to the Uzbek Language](#) provides wordlists and grammatical descriptions concerning certain parameters of the language.

Besides, the UCLA Phonetics Laboratory provides material, i.e. recordings, word lists, phrase-books, etc. that are gathered for phonetic and phonological research purposes.

Furthermore, a [basic vocabulary](#) is available, and the [1,000 most common words of Uzbek](#) are also provided.

[Sketch Engine](#)

18.5.2 Dictionaries

This section provides dictionnaires of the Uzbek language.

Traditional (paper) dictionaries

The following printed dictionnaires of Uzbek are available: [Khakimov \(1994\)](#) (it contains 8,000 entries with phonetic pronunciation and a brief pronunciation guide), [Radjabov \(2015\)](#) (it contains over 20,000 entries both in the Latin and Cyrillic alphabets, as well as, a concise grammar and pronunciation sections, and [Awde et al. \(2002\)](#) (includes c. 5,000 entries, a basic Uzbek grammar section and a pronunciation guide).

Online dictionaries

The [Translatos.com](#) provides easily scrapeable dictionaries for the following language pairs: Russian–Uzbek dictionary (\approx 108000 words and phrases) Uzbek–Russian dictionary (\approx 82000 words and phrases), English–Uzbek dictionary (\approx 17000 words and phrases), Uzbek–English dictionary (\approx 27000 words and phrases). The entries contain translations, contexts and examples sentences.

Additionally, there are English-Uzbek online dictionnaires:

- [Online Pocket Uzbek Dictionary \(UZ-EN, EN-UZ\)](#) contains pronunciation, English translation.

- Ismanov Uzbek Dictionary (UZ-EN, EN-UZ) includes English translations and POS tags.
- Indiana University Uzbek Dictionary (UZ-EN, EN-UZ) provides English translations and contexts for the entries (it is a BETA version).

18.5.3 Corpora

Monolingual corpora

Uzbek is one of the many languages whose text corpora are included in [Sketch Engine](#). The corpus is available to both trial users as well as paying subscribers. It contains 18,720,334 tokens (57K). The source texts were crawled by SpiderLing and postprocessed.

Additionally, European Language Resources Association offers a repository of Language Resources that includes Uzbek data: a [speech corpus of Uzbek](#) is collected from the general news radio broadcasting.

The [Northern Uzbek \(Cyrillic script\)Corpus of Corpora Collection of the Leipzig University](#) includes newspapers based on material crawled in 2011. It contains 134,076 sentences (1,842,697 tokens).

As of 03/10/2016, the Uzbek [wikipedia](#) contains 128,699 articles.

In addition, there are also translations available:

- the Uzbek Bible written in [Cyrillic](#) is also available
- both the [Cyrillic](#) and the [Latin](#) version of the *Universal Declaration of Human Rights* is available
- there is the Uzbek translation of the [Quran](#) written in the Cyrillic script

The Uzbek data of the [European Corpus Initiative Multilingual Corpus](#) collected from an Usbek Novel 'Ärk Freedom' with English interlineal translation. The corpus contains 72K of Uzbek words.

Bilingual corpora

[Li et al. \(2016\)](#) introduce a model in which morpheme-level aligned data have been used in corpus building. The Uzbek-English and Turkish-English data have been manually aligned at the morpheme level.

The [Deltacorpus](#) includes c. 1 million tokens web-crawled Uzbek texts tagged by a DELexicalized Tagger. The corpus is available from the <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1662>.

18.5.4 News portals

There is governmental censorship in Uzbekistan that restricts independent journalism. Licensing is the task of the State Press Committee and the Inter-Agency Coordination Committee. The state controls newspaper distribution and materials supply too. Additionally, no live programming is allowed in the country. There was only one news website called [Uznews.net](#), that operated since 2006, however, it was forced to suspend broadcasting in 2014.

While there are some news agencies that still have offices in Uzbekistan, such as [Trend News Agency](#) or the Reuters, their activities are restricted by media laws. Two foreign agencies has online broadcasting: the [BBC Uzbek](#) and the [Ozodlik Radio](#).

18.6 Computational tools

18.6.1 Language identification

The Basis technology provides a language identifier tool for Uzbek (see the [Rosette Language Identifier](#)). Additionally, a language identifier developed by the [Translated Labs](#) is also available. Finally, the tool named [Polyglot](#) is likewise compatible with Uzbek.

18.6.2 Tokenizer

[Polyglot](#) provides a tokenizer tool for Uzbek. The [MorphAdorner V2.0](#) supports Uzbek.

18.6.3 Stemmer

A description on the development of an [Uzbek stemmer](#) and the [stemmer itself](#) is available online.

18.6.4 Spell checker

The [stars21](#) developed a spell checker which can be used for the Latin scripts. This tool can also be used as language identifier. Furthermore, a free online [spell checker](#) is available for Uzbek. Finally, the HunSpell also provides an [Uzbek spell checker](#). The [GNU Aspell spell checker](#) supports Uzbek.

18.6.5 Phrase level and higher tools

As of 12/01/2017, no Uzbek chunker, named entity recognizer, sentence parser, and question answering machine is available for the Uzbek language.

Uzbek morphological analyzer There are attempts to develop morphological analyzers. The summary of the process is found in [Baisa et al. \(2012\)](#). In addition, [Polyglot](#) offers trained morfessor models to generate morphemes from words which support the Uzbek language. [Matlatipov and Vetulani \(2009\)](#) present the UZMORPP system of automatic morphological parsing for Uzbek.

Uzbek part-of-speech tagger [Yu et al. \(2016\)](#) describes a method of using Universal Dependencies tagset for Uzbek.

speech recognizer [Phonexia](#) provides a speech recognizer for Uzbek.

Uzbek machine translator The [Google translator](#) supports Uzbek (for a detailed description and the supported languages see the [Wikipedia entry of Google Translate](#)).

18.6.6 End-user support

The OS support for Uzbek is the following:

- Uzbek is not supported by the Mac OS X
- Microsoft Windows language pack is available, however it is only for Latin script
- there is a language pack of Linux for Uzbek

The Unicode character range for the Cyrillic script is 0400 — 04FF (and there is an additional Cyrillic supplementary code set: 0500 — 052F). UTF-8 converter tools are available both for the Latin and the Cyrillic scripts, developed by [HunSpell](#).

Online keyboards developed by [Lexilogos](#), [Gate2Home](#), and [Branah.com](#) are available for Uzbek.

Bibliography

Nocilas Awde, William Dirks, and Umida Hikmatullaeva. *Uzbek-English/English-Uzbek Dictionary and Phrasebook: Romanized*. Hippocrene Books, New York, 2002.

Vít Baisa, Vít Suchomel, et al. Large corpora for turkic languages and unsupervised morphological analysis. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey*. European Language Resources Association (ELRA), 2012.

Hendrik Boeschoten. Uzbek. In Lars Johansen and Éva A. Csató, editors, *The Turkic Languages*, pages 357–378. Routledge, London, 1998.

William Fierman. Language planning and national development. *The Uzbek experience*. Berlin: Mouton de Gruyter, 1991.

Kamran M. Khakimov. *Uzbek-English English-Uzbek*. Hippocrene Books, New York, 1994.

Xuansong Li, Jennifer Tracey, Stephen Grimes, and Stephanie Strassel. Uzbek-english and turkish-english morpheme alignment corpora. In <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2016-uzbek-english-turkish-english-morpheme-alignment.pdf>, 2016.

Gayrat Matlatipov and Zygmunt Vetulani. Representation of uzbek morphology in prolog. In *Aspects of Natural Language Processing*, pages 83–110. Springer, 2009.

Aleksey Radjabov. *Uzbek-English/English-Uzbek Practical Dictionary*. Hippocrene Books, New York, 2015.

V. V. Reshetov. Uzbechkij jazyk. In N. A. Baskakov, editor, *Jazyki narodov SSSR. Volume 2: Tjurkskie jazyki*, pages 340–362. Nauka, Moscow, 1966.

Andrée F. Sjoberg. The phonology of standard uzbek. In Nicholas Poppe, editor, *American Studies in Altaic Linguistics*, volume 13 of *Uralic and Altaic Series*, pages 237–262. Indiana University Press, Bloomington, 1962.

Andrée F Sjoberg. *Uzbek Structural Grammar*, volume 18 of *Uralic and Altaic Series*. Indiana University Press, Bloomington, 1963.

A. P. Xodzhiev. Uzbechkij jazyk. In E. R. Tenishev, editor, *Jazyki mira: Tjurkskie jazyki*, pages 426–437. Indrik, Moscow, 1997.

Zhiwei Yu, David Mareček, Zdeněk Žabokrtský, and Daniel Zeman. If you even don't have a bit of bible: Learning delexicalized POS taggers. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 96–103, 2016.

Chapter 19

Vietnamese (Imre Dávid András)

Contents

19.1 Demography and ethnography	285
19.2 Main typological and syntactic features	287
19.3 Writing system, transcription	291
19.4 Previous research on the language	291
19.5 Data and sources	291
19.6 Computational tools	292
Bibliography	294

Introduction

The present chapter aims at providing an introduction of Vietnamese. After a brief review of the demographic and ethnographic situation of the language and its speaking community, the typology, the writing system and the most notable researches about the grammar of the language will be outlined. The second part of this paper provides an account on the various available offline and online data sources of Vietnamese, and an outlook at the available tools for the processing of the language.

19.1 Demography and ethnography

19.1.1 Name variants

Vietnamese language can also be referred to as *Annamese*, *Ching*, *Gin*, *Jing*, *Kihn* and *Viet*. Annam was the name of the central region of Vietnam while it was a French protectorate. Vietnamese people were called Annamites, thus Annamese was the name of their language. Kihn and Viet both refers to the largest ethnic group of the region, it is an endonym, roughly meaning “Vietnamese people”. Gin is the name of a Vietnamese speaking ethnic group in southeastern China, Gin comes from Kihn, it is the endonym of the group, while Jing and Ching are exonyms used by the Chinese. (Paul et al., 2015)

ISO codes for the Vietnamese language are the following: *vi* ISO 639-1, *vie* ISO 639-2, and ISO 639-3.

Country	Number of speakers	Country	Number of speakers
Vietnam	73,594,427	China	30,000
United States	1,548,449	Philippines	27,600
Cambodia	600,000	Norway	18,333
France	350,000	Netherlands	18,000
Canada	220,425	Thailand	10,000
Australia	210,800	Denmark	8,575
Taiwan	200,000	Switzerland	8,173
Germany	150,000	Qatar	8,000
South Korea	143,000	Belgium	7,151
Japan	135,657	New Zealand	4,875
Laos	100,000	Finland	4,000
Malaysia	70,000	Ukraine	3,850
United Kingdom	70,000	Hungary	3,019
Czech Republic	60,931	Slovakia	3,000
Poland	50,000	Italy	3,000
Russia	36,225	Bulgaria	1500

Table 19.1: Number of Vietnamese speakers per country

19.1.2 Geographic spread

Being the national language, most native speakers of the language as well as Vietnamese speaking minorities are located throughout Vietnam and the surrounding countries. There are Vietnamese speaking minority groups in China, Cambodia, Laos and Taiwan. In the United States it is the sixth most spoken language: third most spoken in Texas, fourth most spoken in Arkansas and Louisiana, and fifth in California. There is a significant group of Vietnamese speakers in Australia, making up the seventh largest language group. There are Vietnamese minority groups throughout Europe, out of which the three largest are in France, Germany and the United Kingdom.

In the Czech Republic Vietnamese language has been granted some privileges: it can be used with public authorities and in courts, also in regions, where the at least 10% of the population is Vietnamese, they should be able to access public and legal information in Vietnamese. ([Wikipedia, 2015a](#))

19.1.3 Speaker populations

There are around 77 million Vietnamese speakers around the world, the majority of 73.5 million living in Vietnam, where the EGIDS level of the language is 1, meaning it is a national language. ([Paul et al., 2015](#)) The number of speakers per country can be found in the table 19.1. ([Wikipedia, 2015b](#))

19.1.4 Dialect situation

Vietnamese language has several, mutually intelligible dialects, which are regional varieties of the same language. Linguists traditionally divided Vietnamese into three dialects by regions: South, Central and North Vietnamese. These varieties differ mainly in their phonemes, but there are also minor grammatical differences. Although there is no standard among these dialects, most of the radio and

TV broadcasts, public and legal texts use the Hanoi variety (northern) of Vietnamese. (Paul et al., 2015)

19.2 Main typological and syntactic features

19.2.1 Linguistic typology

Vietnamese belongs to the Viet-Muong subgroup of the Mon-Khmer branch of the Austro-Asiatic language family. Vietnamese is a tonal language, which means that change in tone can change the meaning of a word. In the Vietnamese language, the minimal elements able to carry meaning are the syllables. The structure of Vietnamese syllables is the following: [initial consonant]+[labialization]+[nuclei vowel]+[final consonant/semi vowel] and a [tone]. The nuclei vowel and the tone are mandatory parts, the three others are optional (Bihm, 1999). There is no inflection in Vietnamese, nouns and verbs are not marked for gender, person, or tense. Because of the absence of inflections, word order and tone can carry significant meaning.

Sentences in Vietnamese follow the SVO word order. The next section gives information on specific syntactic constructions.

19.2.2 Predication

According to Thompson (1987), there are three different types of predication in Vietnamese. **Identificational predicates** contain the *identificational copula* **là**. Any independent word or phrase can occur after it, constituting such predicate, and in many cases those words or phrases would not be predicates otherwise, only in the presence of the copula. Identificational predicates can be heads of larger phrases and in this case they are often at the end of such phrases. They can also occur as descriptive complements and in this case they can be embedded in the phrase before its end. Some examples:

- (122) *Mai là sinh viên.*

‘Mai is (a) student.’

- (123) *Ông ấy là lính.*

‘He is (a) soldier.’

Temporal predicates can be distinguished by *tense markers*. There are two tense markers, **dā** ‘anterior’ and **sē** ‘subsequent’. Temporal predicates are of two types **verbal predicates** and **substantival predicates**. These predicates can also be found in the head position or embedded in a larger phrase. Some examples:

- (124) *Tôi sē di.*

‘I’ll go.’

(verbal)

(125) *Chị ấy đã quên.*

‘She’s forgotten’
(verbal)

(126) *Tháng tới tôi sẽ hai mươi lăm tuổi.*

‘Next month I’ll be twenty-five years old.’
(substantival)

For each temporal predicate there is a corresponding phrase only differing in the absence of the tense marker. Those phrases are called **unmarked predicates**. Speaking of usage, they are used in the same way as temporal predicates and they have also the same two types (verbal and substantive). The temporal aspect of unmarked verbal predicates are only clear in context. Some examples:

(127) *Tôi hai mươi lăm tuổi.*

‘I’m twenty-five years old.’
(substantive)

(128) *Tôi đi.*

‘I’m going.’

19.2.3 Possession

Possessive forms are constructed with the use of the **proposition** *của*. According to Bihn (1999), the syntactical rules regarding the proposition are the following. Any noun or pronoun following ‘của’ is modified by it, and the phrase expresses possession:

(129) *ba lô của tôi*

‘my backpack’

(130) *ông của Smith*

‘Smith’s grandfather’

(131) *Chiếc xe mới này là của anh ấy.*

‘This new motorcycle is his (belongs to him).’

If the noun following the proposition does not have any modifier (eg. adjective), *của* can be omitted. Any other cases the use of the proposition is necessary to avoid confusion or agrammatical sentences.

(132) *Xe anh ấy tốt lắm.*

‘His motorcycle is very good.’

(133) *Cái xe mới của anh ấy tốt lắm.*

‘His new motorcycle is very good.’

19.2.4 Imperative

There are several different words that can be used to indicate imperative mood, differing in strength and in the level of formality (Bihm, 1999). In the examples six of those words and their usage is presented. Please note that there are other such words and this list is not complete, and is only used to demonstrate the nature of Vietnamese imperatives:

- **hãy** is placed before a verb and it indicates rather formal and strong command or request:

(134) *Các anh hãy đọc bài này.*

‘Read this article.’

- **cứ** is placed before a verb and it suggests to do something without any hesitation:

(135) *Anh cứ hỏi.*

‘Go ahead, ask.’

- **đi** is placed after a verb, or if the word has some other modifiers, then after the sentence, and it conveys an informal suggestion to start doing something:

(136) *Đọc câu này đi.*

‘Go ahead and start reading this sentence.’

- **nhé** is placed at the end of a sentence and it means a weak, informal suggestion it is used in situations where the speaker expects no disagreement:

(137) *Chúng ta đi xem phim nhé.*

‘Let’s go watch a movie, shall we?’

- **mời** is placed at the beginning of a sentence to form a polite suggestion:

(138) *Mời anh ngồi.*

‘Take a seat, please.’

- **đừng** is placed before the verb to construct negative imperatives, prohibitions:

(139) *Đừng hút thuốc ở nơi công cộng.*

‘Do not smoke in public places.’

19.2.5 Interrogative

Basic interrogative sentences are formed with the help of the frame: *[subject] + có + [predicate] + không*. Its use resembles to the use of 'to be' or 'to do' in English interrogatives. Usually it frames the verb, but if there is a word modifying the verb, and adjective for example, it frames the modifier. Some basic examples:

(140) *Ahn có quen ai không?*

'Do you know anybody?'

(141) *Anh có rảnh không?*

'Are you free?'

If the base of the question is a verbal predicate, in which the verb is có, then there is only one có in the output sentence. Changes in the temporal perspective can be made by changing the temporal aspect of the verb có.

The other syntactic construction to form questions is through the use of **interrogative words**::
Interrogative pronouns:

- *gi* - **what**
- *ai* (at the beginning of a sentence) - **who**
- *ai* (at the end of a sentence) - **whom**

Interrogative adjectives:

- *gi* - **what (kind of)**
- *[noun] + nào* - **which [noun]**
- *hôm nào* - **which day (when)** - at the beginning of the sentence: the question is regarding a future event; at the end of the sentence: the question is regarding a past event.
- *bao nhiêu* - **how much/many**

Interrogative adverbs:

- *dâu [verb]* - **where** - for verbs of motion (eg. go, swim)
- *ở dâu* - **where** - for verbs of location (eg. work, live)
- *sao* (at the beginning of a sentence) - **why**
- *sao* (at the end of a sentence) - **how**

More on interrogatives can be found in [Bihn \(1999\)](#), [Thompson \(1987\)](#) and [Bac et al. \(2012\)](#).

19.3 Writing system, transcription

Although there were multiple writing systems in use until the 20th century, the one called *chữ nho* ('scholar's characters') which is mostly based only Chinese script was abandoned, and now it is only recognized by elderly people or scholars. The official script is called *chữ Quốc Ngữ* ('national language script') was codified by Alexandre de Rhodes, a French Jesuit missionary. Since he relied on the earlier works of Portuguese missionaries, the script reflects the Portuguese alphabet but it makes heavy use of diacritics. There are nine of these diacritics, five of them create additional phonemes and four to mark tone. Because it is common that there are two such diacritics on one letter it makes Vietnamese easily recognizable. The alphabet and the pronunciation of the letters and the explanation of the diacritics can be found on [Omniglot](#).

19.4 Previous research on the language

For a collected list of those researches and publications that are not mentioned in this report, including corpora, lexical resources, linguistic researches, dictionaries and more, the reader should visit the [the OLAC resources](#) site and the [the WALS resources](#) list.

19.5 Data and sources

19.5.1 Basic vocabulary

[The An Crúbadán Project](#) is the largest Vietnamese vocabulary available online with more than 42 thousand unique words extracted from more than 1700 documents. Besides the list of words, the downloadable package also contains word bigrams and character trigrams and the list of the online sources. Please note, since this vocabulary is made of using source codes of websites the bigrams and the trigams contain symbols which were not removed during the clean-up process and are part of the HTML mark-up.

For a meaningful list of words, which can be used for basic communication, even without any knowledge on syntax, the [Vietnamese Swadesh list](#) gives a good starting point. Many sites that aim to teach languages have a basic list of the more important words compiled. Memrise is an experimental language teaching site, which offers courses based on games and community. It teaches [the 500 most common Vietnamese words](#) through a memory game.

19.5.2 Dictionaries

[VDict](#) is a freely available online Vietnamese dictionary, offering translations between English-Vietnamese, French-Vietnamese, and Chinese-Vietnamese. It also has a monolingual Vietnamese dictionary. In each case besides giving the corresponding translation of a given word, it also gives the definition of the word in the target language. It also has an option to translate texts no longer than 200 characters.

[Cambridge Dictionaries](#) has over 40.000 entries of English-Vietnamese words, and besides translation, it also gives short definitions on both languages and some examples of usage.

At last but not least [Google Translate](#) is an outstanding tool for translating both words and texts. Google Translate supports Vietnamese since September, 2008. If one wants to translate Vietnamese to some lower level language, it is advised to translate it first to English, then to the desired language. More on the different levels of languages supported by Google Translate and on the method of

translation can be found on the [Wikipedia article](#) on Google Translate.

19.5.3 Corpora

The largest monolingual corpus in Vietnamese language one can find is the [Vietnamese Wikipedia](#). According to the English Wikipedia [article](#) about The Vietnamese Wiki, it had 1.150.000 articles in November, 2015, out of which almost 400.000 is manually created. The whole Vietnamese Wiki is downloadable through the [wikidump](#) site.

[HC Corpora](#) offers a downloadable monolingual corpora consisting of almost 50 million words.

For bilingual corpora the best candidates are usually sacred texts, literary and legal texts. Luckily, there is no need to list several individual texts here, because there are multiple large corpora available for download at [evbcorpus](#), where the largest English-Vietnamese corpus contains over 10 million words from 15 bilingual books, 350 parallel texts and thousand articles. It also has a detailed documentation about the process of building a bilingual corpus.

Also there is a project on GitHub called [bible-corpus](#), which has a collection of the Bible in more than 100 languages, including Vietnamese.

19.6 Computational tools

19.6.1 Language identification

There are several tools for identifying Vietnamese language. The [Compact Language Detector 2](#) and the project called [langid.py](#) on GitHub, are both written in Python and the [TextCat language guesser](#) is written in Perl, and all three are standalone, open source language identifiers. A standalone, but not open source tool able to identify Vietnamese is [Polyglot3000](#). There are online language identifiers as well: [Translated.net](#) offers a solution to identify 102 languages, including Vietnamese. [Google translate](#) is also able to identify Vietnamese language.

19.6.2 Tokenizer

A tokenizer is a tool used to divide a text into words. An open source tokenizer written in Python, which is also capable of recognizing Vietnamese, is [polygot](#). Another word segmentation tool is [JVnSegmenter](#), this is open source as well, but it was written in Java. Also a Java and Nodejs based open source word segmenter for Vietnamese is [vntokenizer](#) on GitHub, by a user called duyetdev.

19.6.3 Stemmer

As Vietnamese language is an isolating language, the concept of stemming is trivial, therefore there are no stemmers for it.

19.6.4 Spell checker

The widely used spell checker engine, Hunspell, has a Vietnamese dictionary available on GitHub, in a repository called [hunspell-vi](#). There are also pre-packaged spell checking add-ons for Mozilla based products which are also available there. There are also several web-based spell checker services online. [Vspell](#) only specializes in Vietnamese and besides having an input field it is also has an option to upload a file or check spelling on a website. [Stars21](#) is another online spell checker which also functions as a language identifier.

19.6.5 Phrase level and higher tools

One of the most advanced computational linguistic tool one can find is [JVnTextPro](#). It is an open source Java-based NLP tool capable of sentence segmenting, sentence tokenizing, word segmenting and POS (part-of-speech) tagging. There is also a dependency parsing tool-kit, [VnDP](#).

[Microsoft Translation](#) supports Vietnamese language. Its enterprise solution is capable of parsing, translating and manipulating speech, either in real time or using audio files. For more information on its capabilities and pricing, visit Microsoft Translate website.

19.6.6 End-user support

As of November, 2015, the Operating System level support for Vietnamese is the following:

- **Windows** - There is a [Language Pack](#) for every supported Windows operating system and also for Windows XP. Microsoft also provided [Language Interface Packs](#) for Microsoft Office 2007 and Microsoft Office 2010.
- **Mac OS X** - OS X does not support Vietnamese language, there are no available language packs in Vietnamese.
- **Ubuntu** - Ubuntu fully supports Vietnamese language; the language packs are downloadable via the [the Ubuntu packages](#) site.
- **Linux Mint** - For Linux Mint the level of support depends on the chosen desktop environment: while [Cinnamon](#) is fully translated, only 68% of the [MATE](#) environment has been done.
- **Debian** - There is a [package](#) to localise the desktop in Vietnamese in Debian.

Bibliography

Hoai Tran Bac, Minh Nguyen Ha, Duc Vuong Tuan, and Vuong Que. *Colloquial Vietnamese*. Routledge, 2012.

Nhu Ngo Bihn. *Elementary Vietnamese*. Tuttle Publishing, 1999.

L. M. Paul, G. F. Simons, and D. F. Charles. Ethnologue, 2015. URL <http://www.ethnologue.com/language/vie>. [Online; accessed 20-November-2015].

Laurence C. Thompson. *A Vietnamese Reference Grammar*. University of Hawaii Press, 1987.

Wikipedia. Vietnamese people, 2015a. URL https://en.wikipedia.org/wiki/Vietnamese_language. [Online; accessed 18-November-2015].

Wikipedia. Vietnamese language, 2015b. URL https://en.wikipedia.org/wiki/Vietnamese_people. [Online; accessed 18-November-2015].

Chapter 20

Wolof (Noémi Vadász)

Contents

20.1 Demography and ethnography	297
20.2 Main typological and syntactic features	300
20.3 Writing system, transcription	304
20.4 Previous research on the language	304
20.5 Data and sources	307
20.6 Computational tools	309
Bibliography	312

Introduction

Wolof (/wɔlf/) is a language of Senegal, the Gambia, and Mauritania, and the native language of the Wolof people. It belongs to the Senegambian branch of the **Niger–Congo language family**. Unlike most other languages of Sub-Saharan Africa, Wolof is not a tonal language. Wolof is originated as the language of the Lebou people. It is the most widely spoken language in Senegal, spoken natively by the Wolof people (40% of the population) but also by most other Senegalese as a second language. Wolof dialects vary geographically and between rural and urban areas. “Dakar-Wolof”, for instance, is an urban mixture of Wolof, French, and Arabic.

For a more detailed information see [Wikipedia](#).

The ISO 639-2 ([International Organization for Standardization](#)) code of Senegal Wolof is **wol**, of Gambian Wolof is **wof**. The top-level domain for Wolof websites is **.sn**. For a detailed classification of the language see the [classification of wolof](#).

20.1 Demography and ethnography

Wolof is spoken in Senegal, the Gambia and Mauritania. The official language of Senegal is French but it is regularly used only by those Senegalese who were educated in colonial-style French schools. Wolof is one of the country’s national languages along with Balanta-Ganja, Jola-Fonyi, Mandinka, Mandjuk, Mankanya, Noon, Pulaar, Serer-Sine, and Soninke. However, Wolof is rapidly becoming the *lingua franca* of the country, serving as a common language for speakers of its numerous languages, particularly in urban areas where they come in contact with each other.

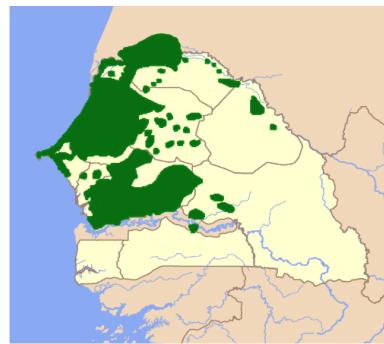


Figure 20.1: The Wolof-speaking territories

Wolof is **not a standardized** language since no single variety has ever been accepted as the norm. However, the variety spoken in Dakar, i.e. in the capital of Senegal, is the one that is most commonly used in radio broadcasts.

20.1.1 Name variants

/wɔlɔf/ is the standard spelling and may refer to the Wolof people or to the Wolof culture. Variants include the older French Ouolof and the principally Gambian *Wollof*. *Jolof*, *jollo* now typically refers either to the former Wolof state or to a common West African rice dish. Now-archaic forms include *Volof* and *Olof*. Some other name variants: Ouolof, Volof, Walaf, Waro-Waro, Yallof. The English exonym is Wolof.

20.1.2 Geographic spread

Wolof is spoken in Senegal and the Gambia. Significant numbers of Wolof speakers also are found in France, Mauritania, and Mali. In the whole region from Dakar to Saint-Louis, and also west and southwest of Kaolack, Wolof is spoken by the vast majority of the people. Typically when various ethnic groups in Senegal come together in cities and towns, they speak Wolof. It is therefore spoken in almost every regional and departmental capital in Senegal. Nevertheless, the official language of Senegal is French.

20.1.3 Speaker populations

Wolof is spoken by more than 10 million people and about the 40% (approximately 5 million people) of the population in Senegal speak Wolof as their native language. Increased mobility, and especially the growth of the capital Dakar, created the need for a common language: today, an additional 40% of the population speak Wolof as a second or acquired language.

In the Gambia, cc. the 20–25% of the population speak Wolof as the first language, but Wolof has a disproportionate influence because of its prevalence in Banjul, i.e. in the Gambian capital, where 75% of the population use it as a first language. Despite the fact that only a tiny minority are ethnic Wolofs in Serekunda, in the largest town of the Gambia, approximately the 70% of the population speaks and/or understands Wolof.

The official language of the Gambia is English; Mandinka (40%), Wolof (10%) and Fula (15%) are as yet not used in formal education.

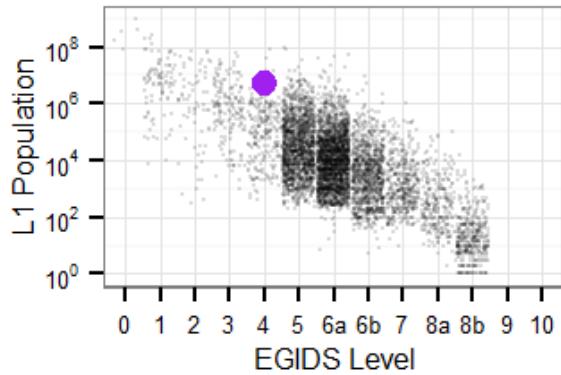


Figure 20.2: Senegal Wolof in language cloud

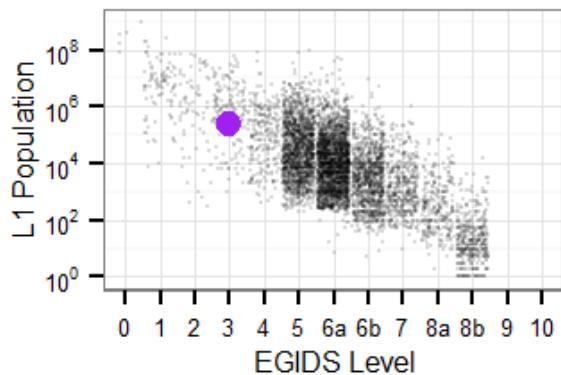


Figure 20.3: Gambian Wolof in language cloud

Figure 20.4: The place of Senegal and Gambian Wolof within the cloud of all living languages. Each language in the world is represented by a small dot that is placed on the grid in relation to its population (in the vertical axis) and its level of development or endangerment (in the horizontal axis), with the largest and strongest languages in the upper left and the smallest and weakest languages (down to extinction) in the lower right. Senegal and Gambian Wolof are represented by a purple dot which means that the languages have been developed to the point that they are used and sustained by institutions beyond the home and community.

In Mauritania, about 7% of the population (approximately 185.000 people) speak Wolof because of the river that is shared with Senegal. There, the language is used only around the southern coastal regions. Mauritania's official language is Arabic; France colonized the tribes and forced them all to speak their French as the official language but the most common language of all other tribes is the Wolof language.

According to [Ethnologue: wol](#) and [Ethnologue wof](#), the EGIDS-level of Senegal and Gambian Wolof is **3** (it is used in wider communication). It means that Senegal and Gambian Wolof has been developed to the point that it is used and sustained by institutions beyond the home and the community.

20.1.4 Dialect situation

Senegal Wolof and **Gambian Wolof** are distinct national standards: they use different orthographies and use different languages as their source for technical loanwords (i.e. Senegal Wolof loans

French words, Gambian Wolof loans English words). However, both the spoken and the written languages are **mutually intelligible**. Senegalese Wolof is the standard form of the language and it is a national language of Senegal (it is the most widely spoken language in Senegal). Within Senegal, the [Ethnologue](#) distinguishes five dialects: Baol, Cayor, Dylof, Lebou, and Jander. Lebou Wolof is unintelligible with standard Wolof: the distinction has, however, been obscured because all Lebu speakers are bilingual in standard Wolof. There are further differences between the Wolof spoken in urban areas as compared to that spoken in rural regions. For instance, Wolof spoken in Dakar has a larger number of French loanwords than other dialects.

20.2 Main typological and syntactic features

20.2.1 Linguistic typology

Wolof is a language of noun classes. Wolof divides nouns into 10 classes (8 singular and 2 plural), each with its own set of articles and pronouns. English has 2 classes (singular, plural), and French has 4 (masculine and feminine in singular and plural forms). The articles, pronouns and modifiers derived from these noun classes are further modified by the physical relation of the object in question to the speaker. That is to say how close or far away the object is. The verbal tense system is based on moods rather than time, with greater emphasis being placed on whether the action is completed or not, than on when the action occurred. In conjugation, only the person-mode-aspect markers change whilst the verb root remains invariable. There are no adjectives in Wolof, their function being replaced by relative clauses formed with adjectival verbs. The verb ‘to be’ does not exist either, although Wolof has a number of mechanisms for predicing a phrase.

The [Wolofresources1](#) and the [Wolofresources2](#) give detailed descriptions of the grammar of Wolof.

20.2.2 Predication

In Wolof, verbs are used for a number of purposes which require other grammatical devices in such Western European languages as English and French. For example, English and French have special grammatical devices for modifying nouns and verbs (i.e. adjectives and adverbs of manner). Wolof has no adjectives and few, if any, adverbs of manner as such. Instead, verbs and verb phrases are used to modify nouns and verbs. In order to understand how Wolof verbs are able to carry out the functions of English or French verbs, adjectives, and adverbs of manner, it will be helpful to consider Wolof verbs as being of two different types:

(1) Active verbs are those which indicate an action or process. Wolof active verbs almost always correspond to the verbs in English e.g. *wax* ‘to speak’, *dem* ‘to leave’, *lekk* ‘to eat’).

(2) Stative verbs are usually verbs which indicate being in a particular state of mind or static condition. Wolof stative verbs, which indicate a state of mind usually correspond to the verbs in English e.g. *xam* ‘to know’, *bëgg* ‘to want, desire’, *nob* ‘to love’, *gëm* ‘to believe in’.

Wolof appears to make some sort of distinction between states of mind, which are indicated by stative verbs, and certain actions or processes which, though of a mental nature, are indicated by active verbs, for instance *foog* ‘to think, believe (that)’ is a stative verb, but *xalaat* ‘to think about, ponder’ is an active verb. The semantic criteria which differentiate active and stative verbs may sometimes be further confused by inaccurate but established translation equivalents for Wolof and French or English. For example, ‘to hate, detest’ may be given as the translation of the verb *bañ*, so

that the verb would appear to be a stative verb. However, *bañ* literally means ‘to refuse, reject’, and is accordingly an active verb.

In English, a noun may be modified either by a simple adjective or by a predicated adjective, e.g. ‘the big house’ and ‘the house is big’. Since Wolof verbs carry out the function of English adjectives, the difference between the two kinds of English noun modification must be shown by some kind of verbal construction. Wolof uses a subject and a stative verb of the appropriate meaning for circumstances in which English has a clause with an adjective in the predicate. The pre-noun modifier construction in English is represented in Wolof by a relative-clause type of construction (e.g. *kér gu rēy gi* ‘the house which is big’).

20.2.3 Possession

When two nouns stand in a genitive relationship ('of', often showing possession), the marker has two forms: *-u* and *-i*. The possesee item (the head) bears the marker.

Although some speakers use the two markers indiscriminately, most preserve the distinction in which *-u* indicates that the possessed noun (it follows) is singular, while *-i* indicates it is plural.

(142) *xaritu Tapha*

‘Tapha’s friend’

doomu seef fi

‘The chief’s son; son of the chief’

bunti kér

‘The doors of the house’

The number with respect to the possessor is ambiguous. *Sama do omu xarit* could mean either ‘the son of my friend’ or ‘the son of my friends’. The *-u* marks the son as singular. When it is necessary to remove this ambiguity, the appropriate noun determiner can be added.

(143) *Sama doomu xarit bi*

‘The son of my friend’

Sama doomu xarit yi

‘The son of my friends’

Samay doomu xarit

‘The sons of my friends’

Possessive constructions consisting of a modified noun and the following noun modifier are usually kept intact. Thus, if the noun modifier (i.e. the last of the two nouns in the construction) is itself modified by one of the possessive pronouns *suma*, *sa*, *sunū*, *seen*, then these possessive pronouns precede the entire two-noun construction, even though it is the second of the two nouns which the possessive pronoun modifies.

(144) *suma doom-u xarit*

‘my friend’s son’

sa xarit-i doom

‘your (sg) son’s friends’

According to Gamble (1991b), there is a difference between Gambian and Senegalese Wolof possessive structures. In Senegalese Wolof possessive structures there is a distinction between singular and plural forms.

(145) *xarit u Mariyaama*

‘Mariama’s’ friend’

xarit i Mariyaama

‘Mariama’s’ friends’

(146) *fas u buur bi*

‘the king’s horse’

fas i buur bi

‘the king’s horses’

Unlike Senegalese Wolof, Gambian Wolof does not have the *u* suffix.

20.2.4 Imperative

In Wolof, different constructions are used for the singular and the plural imperatives. This difference relates strictly to number. There is no reference to politeness. The singular imperative is formed two ways: with an imperative marker and without an imperative marker, depending on the word following the imperative.

When the singular imperative is followed by a direct object pronoun, the base form of the verb is used without any further marker.

(147) *Dimbëli ma!*

‘(You sg) help me!’

Indi ko!

‘(You sg) bring it!’

Jël leen!

‘(You sg) take them!’

In most other cases, e.g. before a noun object or in cases where no word follows the imperative, the verb takes the singular imperative suffix, which is ...*l* if the verb has more than one syllable and ends in a vowel, and ...*al* (or its variant ...*ël*) if the verb ends in a consonant. Single-syllable verbs ending in a vowel may have these forms of the suffix, or may have the suffix in the forms ...*wal* or ...*wël*.

(148) *Dellul!*

‘(You sg) go back!’

Dimbélil Seex!

‘(You sg) help Cheikh.’

Indil ndox mi!

‘(You sg) bring the water.’

Demal!

‘(You sg) get out!’

Jëlöl mburu mi!

‘(You sg) take the bread!’

Jiël/Jiwël!

‘(You sg) plant [seeds]!’

Foal/Fowal!

‘(You sg) play!’

Wooal/Woowal!

‘(You sg) call!’

The Wolof plural imperative always has the same form regardless of whether or not the verb has a pronominal object. The imperative suffix is never used, but the subject *leen* always immediately follows the imperative verb form.

(149) *Dimbélí leen ma!*

‘(You pl) help me!’

Indi leen ko!

‘(You pl) bring it!’

The negative imperative, i.e. the prohibitive, is formed by placing *bul* before the verb (without any imperative suffix) for the 2nd person singular, and *bu leen* for the 2nd person plural.

(150) *Bul dellu!*

‘Don’t (you sg) go back!’

Bu leen dellu!

‘Don’t (you pl.) go back!’

20.2.5 Interrogative

The Wolof question words which can function both as interrogatives and as relative pronouns have the following meanings and forms:

In Wolof, a statement can be made into a question requiring a “Yes” or “No” answer in either of two ways. Firstly, the sentence, with the same word order as the declarative clause has, may be said

	Singular	Plural
thing('what?')	lu	yu
person ('who?')	ku	ñu
place('where?')	fu	
time('when?')	bu	
manner('how?')	nu	

with an interrogative intonation. The main characteristic of this interrogative intonation is a high start. That is, the sentence is begun with a much higher pitch than is normally used for declarative sentences.

Secondly, question-indicating words may be used (in which case the intonation is normal). *Ndax* (roughly ‘[is it] because...’) and *eske* (from French *est-ce-que*) simply make a sentence interrogative. The questioner has no preconceived notions about what the answer will be. Therefore verbs in questions with *ndax* are almost invariably in the affirmative form. A negative verb would indicate a presumed answer, and thus call for the use of *mbaa*, e.g.:

- (151) *Ndax wolof la? Eske wolof la?*

‘Is he a Wolof?’

Mbaa is used when the questioner wishes to indicate that an answer of agreement is expected (i.e. an affirmative answer to an affirmative question, a negative answer to a negative question), but that he is not quite sure. This form is more or less parallel to the English tag-question like: ‘She’s here, isn’t she?’, ‘You’ve finished it, haven’t you?’, ‘You don’t have the money, do you?’, ‘I’m not late, am I?’, etc., e.g.:

- (152) *Mbaa wolof la?*

‘He’s a Wolof, isn’t he?’

Xanaa is used in questions which occur to the questioner because he has gotten some idea from the situation; e.g., the questioner sees a person with a certain appearance, and hears the person mention that he has come from Dakar, he speaks Wolof, etc. The verb after *xanaa* is usually used with *la* (the complement predicator), *a* (the subject predicator) or *da* because by context there will be focus on one part or another of the sentence.

20.3 Writing system, transcription

Wolof was first written with a version of the Arabic script known as *Wolofal*, which is still used by many older men in Senegal. The Wolof orthography using the Latin alphabet was standardised in 1974 and is the official script for Wolof in Senegal.

20.4 Previous research on the language

Columbia University have a page of [resources for African languages](#) on the web including [Wolof materials](#).

A a	B b	C c	D d	E e	Ë ë	F f	G g	I i
J j	K k	L l	M m	N n	Ñ ñ	Dj	O o	P p
Q q	R r	S s	T t	U u	W w	X x	Y y	

Table 20.1: Wolof latin alphabet

Consonants

ا	ب	ٻ	ٻ	ٻ	ٻ	ٻ	ٻ	ٻ	خ
c	y	j	t	c	t	p̪	þ	p	b, p, gb
d	dhaad	t̪qaad	chiin	s	zaay	r	dzaal	d	haay
گ	ڪ	ڪ	ڦ	ڦ	ڦ	ڻ	ڻ	ڻ	ڻ
g	k	k	q	f	f	ŋ	gayn	ayn	tzaay
ل	ڦ	ڦ	ڦ	ڦ	ڦ	ڦ	ڦ	ڦ	ڦ
l	v	'	y, ñ, ڻ	w	h	ñ	n	m	l
									g

Vowels

ء	ء	ء	ء	ء	ء	ء	ء	ء	ء
o/o	ø	o	í	in	i	à	an	aa	a
									hamaz
ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ
ú	ë	é	e	en/en	e/e	un	u	on/on	

Figure 20.5: Wolofal (Arabic script for Wolof)

The French language journal [SudLangues](#) regularly publishes linguistic papers on the Wolof language.

[Becher \(2003\)](#) discusses how the Wolof express things that they have experienced (e.g. emotions, thoughts, perceptions, physical sensations etc.).

[Buell and Sy \(2006\)](#) and [Buell \(2005\)](#) discuss the use of verbal affixes ([PDF](#), [PDF](#)).

[Creissels \(2006\)](#) aims to propose some terminological clarifications in order to lay the foundations of a cross-linguistic study of morphosyntactic phenomena likely to be viewed as particular manifestations of the same type of mechanism as the construct state of Semitic languages ([PDF](#)).

([Creissels and Voisin-Nouguier, 2004](#)) is based on the analysis of Wolof valency changing derivations ([PDF](#)).

[Irvine \(1979\)](#) examines the analytical utility of the concept of “formality” in social-cultural anthropology, particularly the ethnography of communication ([PDF](#)).

[Irvine and Southwest Educational Development Lab. \(1975\)](#) has a book about the Wolof society as well ([PDF](#)).

With structural analysis [Ka \(1988\)](#) provides new evidence concerning the internal structure of the syllable in Wolof through examination of the secret code called Kall, spoken mainly in Senegal’s Ceneba area ([PDF](#)).

[Moore \(2007\)](#) applies conceptual metaphor theory to Wolof, a Niger-Congo language spoken in Senegal and The Gambia, West Africa. The paper examines a subset of the Wolof lexemes that can be used for talking about time in terms of space and motion. Many of the same metaphors that have been proposed for English and other languages also appear in Wolof, sometimes with interesting differences having to do with interactions between metaphor and lexical semantics. Additionally, a Wolof metaphor that does not occur in English is examined ([PDF](#)).

([Rialland and Robert, 2001](#)) is a brief excursion into one area of the Wolof lexicon. It uses a conceptual metaphor theory to structure a description of how certain Wolof words are used to talk about temporal relations in terms of space and motion, in motion metaphors of time ([PDF](#)).

([Ka, 1994](#)) outlines the language structure by analyzing such features as vowel length, complex segments and permissible syllables.

([Dione, 2014](#)) presents several techniques for managing ambiguity in LFG parsing of Wolof, a less-resourced Niger-Congo language. Ambiguity is pervasive in Wolof and This raises a number of theoretical and practical issues for managing ambiguity associated with different objectives. From a theoretical perspective, the main aim is to design a large-scale grammar for Wolof that is able to make linguistically motivated disambiguation decisions, and to find appropriate ways of controlling ambiguity at important interface representations. The practical aim is to develop disambiguation strategies to improve the performance of the grammar in terms of efficiency, robustness and coverage ([PDF](#)).

Some other linguistic papers about the language are available [Wolof Language Acquisitions at Columbia University Libraries](#). More linguistic papers are accessible in French at the page called [Resources for Learning Wolof](#) available for downloading.

The World Atlas of Language Structures ([WALS](#)) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars). There are materials about [Wolof](#).

20.4.1 Grammars

([Gamble, 1991b](#)) is a full description of the language. It is available in [PDF](#). ([Gambia, 1995](#)) is another full description of the language. It is available in [PDF](#). [Wolofresources](#) is a PDF grammar manual on Wolof. [Ngom \(2003\)](#) provides an account of the phonological, morphological and grammatical traits of Wolof as spoken in Senegal, and a Wolof text with an interlinear translation. [Camara \(2006\)](#) provides a detailed grammar of Wolof and a Wolof-English-Wolof dictionary. [Nussbaum and Center for Applied Linguistics \(1970\)](#) basic course in Dakar Wolof is designed to be taught audiolingually by a native speaker of the language. It is available in [PDF](#).

A grammar primer of Gambian Wolof was produced by Sierra Dem, (1995) of the U.S. Peace Corps in the Gambia in PDF format.

20.5 Data and sources

[Omniglot](#) is an encyclopedia of writing systems and languages. It can be used to learn about languages, to learn alphabets and other writing systems, and to learn phrases in many languages. There is also advice on how to learn languages. There are some information about [Wolof](#).

[OLAC](#), i.e. the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources. OLAC Archives contain over 100,000 records covering resources in half of the world's living languages including [Wolof](#).

The [Wolofresources](#) lists some important resources of the language such as courses, dictionaries and lexicons, grammar manuals and linguistic papers.

Other resources of the language can be found here: <http://library.columbia.edu/>.

20.5.1 Basic vocabulary

[Swadesh lists](#) were originally devised by the linguist Morris Swadesh. In the 1940s to 1950s, Swadesh created word lists of body parts, verbs, natural phenomena, in order to compute the relationships of languages, and in particular their age. A Swadesh list may also be useful to achieve knowledge of some universal terms in other languages. This is because, for basic communication, knowledge of vocabulary is more important than knowledge of grammar and syntax. Sometimes it is even possible to achieve (very) basic communication skills with no knowledge of the target language syntax whatsoever. [Niger-Congo Swadesh lists](#) including Wolof is provided as well.

The [Wolofresources](#) provides a basic vocabulary sorted into categories. Important words to learn are distinguished.

The [Omniglot](#) collects the useful phrases. Some of the phrases have audion recording with the correct pronunciation.

Learning Wolof ([Gaye and Togo, 1980](#)) can be used by those who would like to learn Wolof individually or with the help of an informant or tutor. A typical chapter includes an introduction of new material, presentation of dialogue, grammar, cultural information, written exercises, and vocabulary words. A Wolof-English glossary of 2.500 words is appended. It is available in [PDF](#).

[Trainee Wolof Manual](#) is a book for Wolof learners.

20.5.2 Dictionaries

Paper editions

There are some traditional paper edition dictionaries. ([Gamble, 1991a](#)) is a Gambian Wolof – English dictionary.

([Munro and Gaye, 1997](#)) has two major parts: the Wolof–English section includes full entries for each Wolof word they define, with many bound grammatical morphemes, as well as cross-references to variant forms and words which occur only as parts of longer expressions. The English–Wolof index relates English words to entries in the main Wolof-English secion. It is available in [PDF](#).

[The Peace Corps dictionary of Gambian Wolof](#) was produced by Sierra Dem, (1995) of the U.S. Peace Corps in the Gambia in PDF format.

([Kantorek, 2005](#)) with its more than 3.000 total dictionary entries is a comprehensive phrasebook with an easy-to-use pronunciation guide.

Find some more paper edition dictionaries [Wolof Language Acquisitions at Columbia University Libraries](#).

Online dictionaries

There are some online Wolof – English dictionaries such as [resourcepage.gambia](#), [afroweb](#), and [xlingua](#).

There are dictionaries with other target language, [Freelang](#) is a French – Wolof dictionary, [home2.swipnet.se](#) is a Swedish – Wolof – English dictionary.

There are some collections of Wolof proverbs. [Léebuy Wolof yi - lu dajale 2750](#) léebuy wolof is a collection of 2.750 proverbs without translation or explanation. [Wisdom of the Wolof Sages](#) is a collection of about 750 Wolof proverbs translated and explained in English. [Burton \(1865\)](#) is a collection of proverbs from West African people groups including a chapter of Wolof proverbs.

Scraping Since [resourcepage.gambia](#) is a traditional dictionary digitalized in PDF format it can be easily transferred to a plain text file with az OCR application.

[Afroweb](#) is a scrapable dictionary due to its relative big size and transparent structure. The words are collected into letters which contain pages. An entry always includes the Tagalog word form, its part-of-speech (or more part-of-speech categories) and the English translation. Sometimes an expanded or varied word form or an example is included as well, and if there are more translations they are listed.

Searchbox dictionaries with an easily parametrizable search interface like [xlingua](#) are scrapable with the help of a world-list wich contains the lexical forms of Wolof words. If no more completed list is available, [Niger-Congo Swadesh lists](#) can be used for this purpose. A more detailed list is available on [Lego](#).

20.5.3 Corpora

The [African Languages Materials Archive](#) includes a number of electronic books in Wolof (mostly in PDF). The [Universal Declaration of Human Rights](#) is available in Wolof as well.

The [ANNIS](#) is a web-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation. It is an open source, cross platform (Linux, Mac, Windows), web browser-based search and visualization architecture for complex multi-layer linguistic corpora with diverse types of annotation. Since complex linguistic phenomena, such as information structure

interact on many levels, ANNIS addresses the need to concurrently annotate, query and visualize data from such varied areas as syntax, semantics, morphology, prosody, referentiality, lexis and more. For projects working with spoken language, support for audio / video annotations is also required.

The [Indigenous Tweets](#) is a website that records minority language Twitter messages to help indigenous speakers contact each other. There are 39.292 entries by 13 users in [Wolof](#). There are blogs in [Wolof](#) listed as well containing 3 blogs altogether with 3 posts (1.093 words).

Wikipedia The [Wikipedia](#) offers free copies of all available content to interested users. The [Wiki-media](#) is a global movement whose mission is to bring free educational content to the world. The content of Wikimedia is dumped and available as well. The [Wolof Wikipedia](#) contains 1.055 articles as of 21/10/2016.

These dumps are available in Tagalog in 21/10/2016: [Wolof Wikipedia](#), [Wolof Wiktionary](#) and [Wolof Wikiquote](#).

Bible The Bible is available in Wolof from different sources as [gospelgo](#), [paul timothy](#) and [jehovah's witnesses](#).

Bilingual United Nations Human Rights Office of the High Commissioner [OHCHR](#) worldwide collection of materials on the Universal Declaration of Human Rights (UDHR), including various resources developed by governmental and non-governmental organizations both on the occasion of the Declaration's 50th Anniversary (1998) and prior to/after the Anniversary year. The collection is unique in the world and comprises more than 400 items. Since UDHR is the most translated document – and is available in [Wolof](#) – it can be used as a multilingual corpus.

20.5.4 News portals

The [Xibaaryi.com](#) news website contains video, audio and text sources of news. The [Wolofresources.com](#) contains a collection of Seleganese radio stations, podcasts and TV stations on Wolof.

Some Senegalese radio stations stream on the internet like the [SudFM](#). It has talksback or news in Wolof. The [Radio Santati](#) and the [Rewmi FM](#) mostly play music but they have some Wolof programs at various times of the day.

20.5.5 Contact person

The [Wolofresources](#) lists a lot of opportunities to study Wolof including universities and institutions in the USA and in Europe. All universities and departments offering Wolof language courses are supplied with address, and most of them with a contact e-mail address as well.

20.6 Computational tools

According to [Dione et al. \(2010\)](#), their work is the first effort in building publicly available NLP resources for the Wolof language. More generally, there has recently been a growing interest in NLP technologies for African languages ([PDF](#)). In recent times, a number of researchers and institutions, both from Africa and elsewhere, have come forward to share the common goal of developing capabilities in language technologies for the African languages. The goal of [De Pauw et al. \(2009\)](#), then, is to provide a forum to meet and share the latest developments in this field. It also seeks to attract

linguists who specialize in African languages and would like to leverage the tools and approaches of computational linguistics, as well as computational linguists who are interested in learning about the particular linguistic challenges posed by the African languages.

20.6.1 Language identification

[Compact Language Detector 2](#) probabilistically detects over 80 languages including Wolof in Unicode UTF-8 text, either plain text or HTML/XML.

20.6.2 Tokenizer

[Cheikh Bamba \(2013\)](#), presents an LFG-analysis of Wolof valency-changing affixes, focusing on those morphemes found in the applicative and causative constructions. The analysis is implemented using the XLE parsing tool. The relevant components of the system include a tokenizer, a finite-state morphological analyzer, annotated phrase structure and sub-lexical rules.

20.6.3 Stemmer

No separate tokenizer tool for Wolof has been found so far. The [Verbix](#) has an online Wolof verb conjugator.

20.6.4 Spell checker

[Wolof Spell Check Engine](#) is a free spell checker.

20.6.5 Phrase level and higher tools

POS-tagger Since no NLP resources were available for Wolof, [Dione et al. \(2010\)](#) had to design a tagset and create a POS-annotated gold standard from scratch. The paper presents the process of creating a semi-automatically annotated gold standard, exploiting available lexical resources and using purpose-built heuristic tools for stemming and guessing of word forms (see [PDF](#)).

Speech recognition [Kouawa et al. \(2011\)](#) developed a model Automatic Speech Recognition (ASR) system and a Speech Synthesis or Text-to-Speech (TTS) system on keywords of the vernacular language Wolof (see [PDF](#)).

[Gauthier et al. \(2016a\)](#) used Hidden Markov Models for building acoustic models for spoken language Wolof for the task of ASM (see [PDF](#)).

[Gauthier et al. \(2016b\)](#) introduces an experiment which shows that how the vowel length contrast can be used for ASM in the case of two African languages, Hausa and Wolof (see [PDF](#)).

20.6.6 End-user support

GNOME GNU computing platform for Unix-like operating systems has [Wolof Language packs](#).

For Mac OS X no Tagalog language pack is available ([support.apple.com](#)).

The Alf@net project of [ANAFIA](#) has translated [Firefox](#) and [AbiWord](#) into Wolof. They also included several [tutorials](#) on using a computer and a [typing program](#) in Wolof.

There is a project to translate Debian (Linux) Installer in [Wolof](#).

All the Wolof characters not found in English are contained in the Unicode Latin supplement 1 and Latin Extended A character sets. Refer to the [Chart of special Wolof character Unicode codes](#) for more information.

[Mozilla Firefox](#) provides a Wolof language pack add-on for its browser.

Bibliography

Jutta Becher. Experiencer constructions in Wolof. *Hamburger afrikanistische Arbeitspapiere (HAAP)*, 2:1–89, 2003.

Leston Buell. A fixed hierarchy for Wolof verbal affixes, 2005. PDF available.

Leston Buell and Mariame Iyane Sy. Affix ordering on Wolof applicatives and causatives. In John Murrath Mugane, John Priestley Hutchison, and Dee A. Worman, editors, *Selected proceedings of the 35th annual conference on African linguistics: African languages and linguistics in broad perspective*, pages 214–224. Somerville, MA, USA: Cascadilla Proceedings Project, Somerville MA, 2006.

R.F. Burton. *Wit and Wisdom from West Africa: Or, A Book of Proverbial Philosophy, Idioms, Enigmas, and Laconisms*. Tinsley brothers, 1865.

S. Camara. *Wolof Lexicon and Grammar*. NALRC Press, 2006. ISBN 9781597030120. URL <https://books.google.hu/books?id=ApgLAQAAMAAJ>.

Dione Cheikh Bamba. Valency Change and Complex Predicates in Wolof: An LFG Account, 2013. A submission to Lexical Functional Grammar Conference.

Denis Creissels. The construct form of nouns in African languages: a typological approach, 2006. 36th Colloquium on African Languages and Linguistics August 28 – 30 2006 Leiden University.

Denis Creissels and Sylvie Voisin-Nouguier. The verbal suffixes of Wolof coding valency changes and the notion of co-participation. Paper presented at the Workshop on Reciprocity and Reflexivity, FU Berlin, 1-2 October 2004, 2004.

Guy De Pauw, L. Levin, and G. M. de Schryver, editors. *Proceedings of the workshop on Language Technologies for African Languages (AfLaT 2009)*, Athens, Greece, 2009. Association for Computational Linguistics. ISBN 1-932432-25-6.

Cheikh M. Bamba Dione. LFG Parse Disambiguation for Wolof. *Journal of Language Modelling*, 2(1):105–165, 2014.

Cheikh M. Bamba Dione, Jonas Kuhn, and Sina Zarrieß. Design and Development of Part-of-Speech-Tagging Resources for Wolof (Niger-Congo, spoken in Senegal). In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, LREC'10, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

Peace Corps The Gambia. *Wollof Grammar Manual*. Peace Corps, The Gambia, 1995.

- David Percy Gamble. *Gambian Wolof-English dictionary*, volume 23 of *Gambian studies*. Department of Anthropology, San Francisco State University (SFSU), San Francisco, revised edition edition, 1991a.
- David Percy Gamble. *Elementary Gambian Wolof grammar*, volume 25 of *Gambian studies*. Department of Anthropology, San Francisco State University (SFSU), San Francisco, 1991b.
- Elodie Gauthier, Laurent Besacier, and Sylvie Voisin. Automatic Speech Recognition for African Languages with Vowel Length Contrast. *Procedia Computer Science*, 81:136 – 143, 2016a. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.
- Elodie Gauthier, Laurent Besacier, and Sylvie Voisin. Automatic Speech Recognition for African Languages with Vowel Length Contrast. *Procedia Computer Science*, 81:136 – 143, 2016b. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.
- A. Gaye and Peace Corps (U.S.). Togo. *Practical course in Wolof: an audio-aural approach, student's manual*. U.S. Peace Corps, Regional Training Resource Ofice, 1980. URL <https://books.google.hu/books?id=TBIkAQAAMAAJ>.
- Judith T. Irvine. Formality and Informality in Communicative Events. *American Anthropologist*, 81 (4):773–790, 1979.
- Judith T. Irvine and TX. Southwest Educational Development Lab., Austin. *Wolof Speech Styles and Social Status. Working Papers in Sociolinguistics Number 23 [microform]* / Judith T. Irvine . Distributed by ERIC Clearinghouse [Washington, D.C.] , 1975.
- O. Ka. *Wolof Phonology and Morphology*. University Press of America, 1994. ISBN 9780819192882. URL <https://books.google.hu/books?id=e3pkAAAAMAAJ>.
- Omar Ka. Wolof Syllable Structure: Evidence from a Secret Code. In *Proceedings of the Eastern States Conference on Linguistics*, 1988.
- Nyima Kantorek. *Wolof Dictionary & Phrasebook*. Hippocrene Books, New York, 2005.
- James Tamgno Kouawa, Morgan Richomme, Claude Lishou, Aristide Thomas Mendo'o, Pascal Uriel Elingui, and Seraphin D. Oyono Obono. Speech Recognition and Text-to-speech Solution for Vernacular Languages. In *ThinkMind // ICDT 2011, The Sixth International Conference on Digital Telecommunications*, pages 56–63, april 2011.
- Kevin Ezra Moore. Lexical Resources in Wolof and English for Talking of Time in Terms of Space. In Doris L. Payne and Jaime Peña, editors, *Selected Proceedings of the 37th Annual Conference on African Linguistics*, pages 111–124, 2007.
- P. Munro and D. Gaye. *Ay Baati Wolof: A Wolof Dictionary*. UCLA occasional papers in linguistics. Department of Linguistics, University of California, Los Angeles, 1997. URL <https://books.google.hu/books?id=3V9spwAACAAJ>.
- Fallou Ngom. *Wolof*. Languages of the World/Materials 333. Lincom GmbH, Muenchen, 2003. URL get-book.cfm?BookID=6596.

Loren V. Nussbaum and DC. Center for Applied Linguistics, Washington. *Dakar Wolof [microform] : A Basic Course / Loren V. Nussbaum and Others* . Distributed by ERIC Clearinghouse [Washington, D.C.] , 1970. URL <http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED051686>.

Annie Rialland and Stéphane Robert. The intonational system of Wolof. *Journal of Linguistics*, 39 (5):873–939, 2001.

Chapter 21

Yoruba (Noémi Vadász)

Contents

21.1 Demography and ethnography	318
21.2 Main typological and syntactic features	322
21.3 Writing system, transcription	324
21.4 Previous research on the language	325
21.5 Data and sources	326
21.6 Computational tools	328
Bibliography	333

Introduction

Yoruba (English pronunciation: /'jɒrʊbə/; Yor. èdè Yorùbá) is a language spoken in West Africa mainly in Nigeria. The number of speakers of Yoruba is approaching 30 million. It is a pluricentric language spoken principally in Nigeria and Benin, with communities in other parts of Africa, Europe, and the Americas. A variety of the language, Lucumi, is the liturgical language of the Santería religion of the Caribbean. Yoruba is most closely related to the Itsekiri language (spoken in the Niger Delta) and to Igala (spoken in central Nigeria).

Yoruba is classified within the Edekiri languages, which together with Itsekiri and the isolate Igala form the Yoruboid group of languages within the Volta-Niger branch of the Niger-Congo family. The linguistic unity of the Niger-Congo family dates to deep prehistory, estimates ranging around 15 kya (the end of the Upper Paleolithic). In present day Nigeria, it is estimated that there are over 40 million Yoruba primary and secondary language speakers and several other millions of speakers outside Nigeria making it the most widely spoken African language outside Africa.

(source of this section: [Wikipedia](#))

The ISO 639-3 ([International Organization for Standardization](#)) code of Yoruba is **yor**. The top-level domain for Yoruba websites is **.ng**.

21.1 Demography and ethnography

21.1.1 Name variants

Yoruba (english exonym: Yoruba, endonym: èdè Yorùbá) has some name variants like Yariba and Yooba. As an ethnic description, the word ‘Yoruba’ was first recorded in reference to the Oyo Empire in a treatise written by the 16th-century Songhai scholar Ahmed Baba. It was popularized by Hausa usage and ethnography written in Arabic and Ajami during the 19th century, in origin referring to the Oyo exclusively. The extension of the term to all speakers of dialects related to the language of the Oyo (in modern terminology North-West Yoruba) dates to the second half of the 19th century. It is due to the influence of Samuel Ajayi Crowther, the first Anglican bishop in Nigeria. Crowther was himself a Yoruba and compiled the first Yoruba dictionary as well as introducing a standard for Yoruba orthography.

The alternative name Akú, apparently an exonym derived from the first words of Yoruba greetings (such as È kú àáro? ‘good morning’, È kú ale? ‘good evening’) has survived in certain parts of their diaspora as a self-descriptive, especially in Sierra Leone.

(source of this subsection: [Wikipedia](#))

21.1.2 Geography

The Yoruba people (Yoruba: àwọn ọmọ Yorùbá) are an ethnic group of Southwestern and North central Nigeria as well as Southern and Central Benin in West Africa. The Yorùbá constitute over 40 million people in total; the majority of this population is from Nigeria and make up 21% of its population, according to the CIA World Factbook, making them one of the largest ethnic groups in Africa. The majority of the Yoruba speak the Yoruba language which is tonal, and is the Niger-Congo language with the largest number of native speakers.

The Yorùbá share borders with the Borgu in Benin; the Nupe and Ebira in central Nigeria; and the Edo, the Esan, and the Afemai in mid-western Nigeria. The Igala and other related groups are found in the northeast, and the Egun, Fon, Ewe and others in the southeast Benin. The Itsekiri who live in the north-west Niger delta are related to the Yoruba but maintain a distinct cultural identity. Significant Yoruba populations in other West African countries can be found in Ghana, Togo, Ivory Coast, Liberia and Sierra Leone.

The Yoruba diaspora consists of two main groupings, one of them includes relatively recent migrants, the majority of which moved to the United States and the United Kingdom after major economic changes in the 1970s; the other is a much older population dating back to the Atlantic slave trade. This older community has branches in such countries as Cuba, Brazil, and Trinidad and Tobago.

(source of this section: [Wikipedia](#))

21.1.3 Speaker population

In the [21.1](#) table for counting the L1 ratio I used the latest census of the countries. The numbers of L1 and L2 speakers are from [Ethnologue](#).

According to [Ethnologue](#) the EGIDS level for this language in its primary country – Nigeria – is **2** (Provincial) – The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.



Figure 21.1: Where Yoruba is spoken

country	# L1	EGIDS level	L1 ratio	# L2
Nigeria	18900000	2	9,2%	2000000
Benin	465000(L2?)	3	23,4%?	N/A
total	19380800			N/A

Table 21.1: L1 and L2 speakers

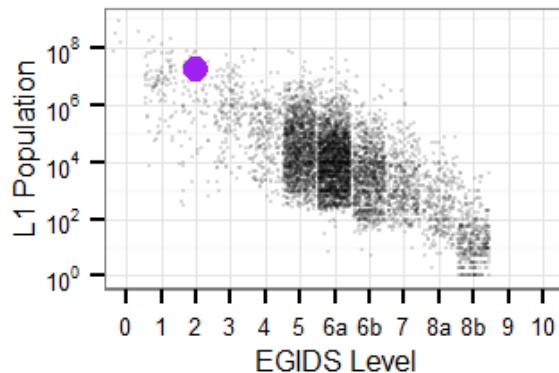


Figure 21.2: Yoruba in language cloud, see the explanation in section .

21.1.4 Dialect situation

The Yoruba dialect continuum itself consists of several dialects. The various Yoruba dialects in the Yorubaland of Nigeria can be classified into three major dialect areas: Northwest, Central, and Southeast. Of course, clear boundaries can never be drawn and peripheral areas of dialectal regions often have some similarities to adjoining dialects.

- North-West Yoruba (NWY): Abeokuta, Ibadan, Oyo, Ogun and Lagos (Eko) areas
- Central Yoruba (CY): Igbomina, Yagba, Ilésà, Ifé, Ekiti, Akure, Efón, and Ijebu areas.
- South-East Yoruba (SEY): Okitipupa, Ilaje, Ondo, Owó, Ikare, Sagamu, and parts of Ijebu.

North-West Yoruba is historically a part of the Oyo empire. In NWY dialects, Proto-Yoruba /gh/ (the velar fricative [y]) and /gw/ have merged into /w/; the upper vowels /i/ and /u/ were raised and merged with /i/ and /u/, just as their nasal counterparts, resulting in a vowel system with seven oral and three nasal vowels. Ethnographically, traditional government is based on a division of power between civil and war chiefs; lineage and descent are unilineal and agnatic.

South-East Yoruba was probably associated with the expansion of the Benin Empire after c. 1450 AD. In contrast to NWY, lineage and descent are largely multilinear and cognatic, and the division of titles into war and civil is unknown. Linguistically, SEY has retained the /gh/ and /gw/ contrast, while it has lowered the nasal vowels /in/ and /un/ to /en/ and /on/, respectively. SEY has collapsed the second and third person plural pronominal forms; thus, àn áñ wá can mean either ‘you (pl.) came’ or ‘they came’ in SEY dialects, whereas NWY for example has e wá ‘you (pl.) came’ and wón wá ‘they came’, respectively. The emergence of a plural of respect may have prevented coalescence of the two in NWY dialects.

Central Yoruba forms a transitional area in that the lexicon has much in common with NWY, whereas it shares many ethnographical features with SEY. Its vowel system is the least innovating (most stable) of the three dialect groups, having retained nine oral-vowel contrasts and six or seven nasal vowels, and an extensive vowel harmony system.

(source of this section: [Wikipedia](#))

Ethnography The Yoruba group is assumed to have developed out of undifferentiated Volta–Niger populations by the 1st millennium BC. Settlements of early Yoruba speakers are assumed to correspond to those found in the wider Niger area from about the 4th century BC, especially at Ife. As the North-West Yoruba dialects show more linguistic innovation, combined with the fact that Southeast and Central Yoruba areas generally have older settlements, suggests a later date of immigration for Northwest Yoruba.

(source of this section: [Wikipedia](#))

The history of the Yoruba people The documented history of the Yoruba people begins with the Oyo Empire, which became dominant in the early 17th century. Older traditions of the formerly dominant Ife kingdom are sparse and unreliable.

The peoples who lived in Yorubaland, at least by the seventh century BC, were not initially known as the Yoruba, although they shared a common ethnicity and language group. The historical Yoruba develop in situ, out of earlier (Mesolithic) Volta-Niger populations, by the 1st millennium B.C.E.

Oral history recorded under the Oyo Empire derives the Yoruba as a race from the population of the older kingdom of Ile-Ife. Archaeologically, the settlement at Ife can be dated to the 4th century BC, with urban structures appearing in the 12th century (the urban phase of Ife before the rise of Oyo, ca. 1100-1600, is sometimes described as a “golden age” of Ife).

Ife was surpassed by the Oyo Empire as the dominant Yoruba military and political power between 1600 CE and 1800 CE. The nearby kingdom of Benin was also a powerful force between 1300 and 1850 CE.

Most of the city states were controlled by Obas, elected priestly monarchs, and councils made up of Oloyes, recognised leaders of royal, noble and, often, even common descent, who joined them in ruling over the kingdoms through a series of guilds and cults. Different states saw differing ratios of power between the kingship and the chiefs' council. Some, such as Oyo, had powerful, autocratic monarchs with almost total control, while in others such as the Ijebu city-states, the senatorial councils were supreme and the Qba served as something of a figurehead.

In all cases, however, Yoruba monarchs were subject to the continuing approval of their constituents as a matter of policy, and could be easily compelled to abdicate for demonstrating dictatorial tendencies or incompetence. The order to vacate the throne was usually communicated through an aroko or symbolic message, which usually took the form of parrots' eggs delivered in a covered calabash bowl by the senators.

The Yoruba eventually established a federation of city-states under the political ascendancy of the city state of Oyo, located on the Northern fringes of Yorubaland in the savanna plains between the forests of present Southwest Nigeria and the Niger River.

Following a Jihad led by Uthman Dan Fodio and a rapid consolidation of the Hausa city states of contemporary northern Nigeria, the Fulani Sokoto Caliphate invaded and annexed the buffer Nupe Kingdom. It then began to advance southwards into Oyo lands. Shortly afterwards, its armies overran the Yoruba military capital of Ilorin, and then sacked and destroyed Oyo-Ile, the royal seat of the Oyo Empire.

Following this, Oyo-Ile was abandoned, and the Oyo retreated south to the present city of Oyo (formerly known as “Ago d'Oyo”, or “Oyo Atiba”) in a forested region where the cavalry of the Sokoto Caliphate was less effective. Further attempts by the Sokoto Caliphate to expand southwards were checked by the Yoruba who had rallied in defence under the military leadership of the ascendant Ibadan clan, which rose from the old Oyo Empire, and of the Ijebu city-states. However, the Oyo hegemony had been dealt a mortal blow. The other Yoruba city-states broke free of Oyo dominance, and subsequently became embroiled in a series of internecine conflicts that soon metamorphosed into a full scale civil war. These events weakened the southern Yorubas in their resistance to British colonial and military invasions. Maria Lugones observes that among the Yoruba people there was no concept of gender and no gender system at all before colonialism. She argues that colonial powers used a gender system as a tool for domination and fundamentally changing social relations among the indigenous. In 1960, greater Yorubaland was subsumed into the Federal Republic of Nigeria. The historical records of the Yoruba, which became more accessible in the nineteenth century with the more permanent arrival of the Europeans, tell of heavy Jihad raids by the mounted Fulani warriors of the north as well as of endemic intercity warfare amongst the Yoruba themselves. Archaeological evidence of the greatness of their ancient civilization in the form of, amongst other things, impressive architectural achievements like Sungbo's Eredo that are centuries old, nevertheless abound.

(source of this subsection: [Wikipedia](#))

Literary yoruba Literary Yoruba, also known as Standard Yoruba, Yoruba koiné, and common Yoruba, is a separate member of the dialect cluster. It is the written form of the language, the standard variety learned at school and that spoken by newsreaders on the radio. Standard Yoruba has its origin in the 1850s, when Samuel A. Crowther, the first African Bishop, published a Yoruba grammar and started his translation of the Bible. Though for a large part based on the Oyo and Ibadan dialects, Standard Yoruba incorporates several features from other dialects. It also has some features peculiar to itself, for example the simplified vowel harmony system, as well as foreign structures, such as calques from English which originated in early translations of religious works.

Because the use of Standard Yoruba did not result from some deliberate linguistic policy, much controversy exists as to what constitutes “genuine Yoruba”, with some writers holding the opinion that the Oyo dialect is the most “pure” form, and others stating that there is no such thing as genuine Yoruba at all. Standard Yoruba, the variety learnt at school and used in the media, has nonetheless been a powerful consolidating factor in the emergence of a common Yoruba identity.

(source of this section: [Wikipedia](#))

21.2 Main typological and syntactic features

21.2.1 Linguistic typology

Yoruba is a highly isolating language. Its basic constituent order is subject–verb–object (SVO), as in *ó nà Adé* ‘he beat Adé’. The bare verb stem denotes a completed action (often called perfect); tense and aspect are marked by preverbal particles such as *ń* “imperfect/present continuous”, *ti* “past”. Negation is expressed by a preverbal particle *kò*. Serial verb constructions are common, as in many other languages of West Africa.

As in [Bamgbose \(1966\)](#) and in [Rowlands \(1969\)](#) Yoruba is a SVO-languaise. The negation-element – if there is any – is positioned between the subject and the verb (SNegVO), as in the case of the English language as well. The order of constituents within phrases correlates with the basic word order. The noun precedes the adjective(s) and the relative clause as well. Yoruba is characterized by using prepositions and Noun-Genitive, Noun-Adjective, Noun-Demonstrative and Noun-Numeral order.

Although Yoruba has no grammatical gender, it does have a distinction between human and non-human nouns; probably a remainder of the noun class system of proto-Niger–Congo, the distinction is only apparent in the fact that the two groups require different interrogative particles: *tani* for human nouns (‘who?’) and *kini* for non-human nouns (‘what?’). The associative construction (covering possessive/genitive and related notions) consists of juxtaposing nouns in the order modified-modifier as in *inú àpótí* inside box ‘the inside of the box’, *filà àkàndé* ‘Akande’s cap’ or *àpótí aṣo* ‘box for clothes’. More than two nouns can be juxtaposed: *rélùweè abé ilè* (railway under ground) ‘underground railway’, *inú àpótí aṣo* ‘the inside of the clothes box’. In the rare case where this results in two possible readings, disambiguation is left to the context. Plural nouns are indicated by a plural word.

There are two prepositions: *ní* ‘on, at, in’ and *sí* ‘onto, towards’. The former indicates location and absence of movement, the latter encodes location/direction with movement. Position and direction are expressed by these prepositions in combination with spatial relational nouns like *orí* ‘top’, *ápá* ‘side’, *inú* ‘inside’, *etí* ‘edge’, *abé* ‘under’, *ilè* ‘down’, etc. Many of these spatial relational terms are historically related to body-part terms.

(source of section: [Wikipedia](#))

21.2.2 Predication

According to Wetzer (2013), in Yoruba, verbal predicates are encoded by means of zero marking. Nominal predicates can be distinguished from verbal predicates by the presence of an overt copula. For the expression of nominal predicates, different copular items can be used, the most frequent ones being *jé* and *se*. Predicate adjectivals in Yoruba are considered to be verby; they are not marked for person and appear without an overt copula. The following examples are from Welmers (1973) and Rowlands (1969):

- (153) ó *lo*
he go

‘He went.’ Welmers (1973)[257]

- ó *ga*
he tall

‘He is tall.’ Welmers (1973)[257]

- ó *jé* ènià
he COP person

‘He is a human being (i.e. not a ghost, animal etc.)’ Rowlands (1969)[153]

21.2.3 Possession

According to Bisang (2007), in contrast with the English possessive structures, there is only one possible word order in Yoruba, the possessed-possessor order.

English	NGen	the car of my father
	GenN	My father’s car
Yoruba	NGen	<i>mòtò bába mi</i>
		car father I

21.2.4 Imperative

Find some examples of Yoruba imperatives here: [Yoruba imperatives](#).

- (154) *Ko si le!*

‘Write it down!’

21.2.5 Interrogative

Find some examples of Yoruba interrogatives here: [Yoruba interrogatives](#).

- (155) *Kí lorúko e?*

‘What is your name?’

A	B	D	E	È	F	G	Gb	H	I	J	K	L	M	N	O	Ò	P	R	S	Ù	T	U	W	Y
a	b	d	e	é	f	g	gb	h	i	j	k	l	m	n	o	ó	p	r	s	ú	t	u	w	y
á	à	Á	é	è	É	E/ È	é / é	è / è	É / È	í	ì													
Í	ó	ò	Ó	Q / Q	ó / ó	ò / ò	Ó / Ó	ú	ù	Ú	S / S													
á	à	á	é	è	é	e / e	é / é	è / è	é / é	í	ì													
í	ó	ò	ó	ó / ó	ó / ó	ò / ò	ó / ó	ú	ù	ú	ú													

21.3 Writing system, transcription

In the 17th century Yoruba was written in the Ajami script, a form of Arabic. Modern Yoruba orthography originated in the early work of CMS missionaries working among the Aku (Yoruba) of Freetown. One of their informants was Crowther, who later would proceed to work on his native language himself. In early grammar primers and translations of portions of the English Bible, Crowther used the Latin alphabet largely without tone markings. The only diacritic used was a dot below certain vowels to signify their open variants [E] and [O], viz. ⟨e⟩ and ⟨ó⟩. Over the years the orthography was revised to represent tone among other things. In 1875 the Church Missionary Society (CMS) organised a conference on Yoruba Orthography; the standard devised there was the basis for the orthography of the steady flow of religious and educational literature over the next seventy years.

The current orthography of Yoruba derives from a 1966 report of the Yoruba Orthography Committee, along with Ayo Bamgbose's 1965 Yoruba Orthography, a study of the earlier orthographies and an attempt to bring Yoruba orthography in line with actual speech as much as possible. Still largely similar to the older orthography, it employs the Latin alphabet modified by the use of the digraph ⟨gb⟩ and certain diacritics, including the traditional vertical line set under the letters ⟨e⟩, ⟨ó⟩, and ⟨ú⟩. In many publications the line is replaced by a dot ⟨e⟩, ⟨ó⟩, ⟨ú⟩. The vertical line had been used to avoid the mark being fully covered by an underline.

The Latin letters ⟨c⟩, ⟨q⟩, ⟨v⟩, ⟨x⟩, ⟨z⟩ are not used.

The pronunciation of the letters without diacritics corresponds more or less to their International Phonetic Alphabet equivalents, except for the labial-velar stops [kp] (written ⟨p⟩) and [gb] (written ⟨gb⟩), in which both consonants are pronounced simultaneously rather than sequentially.

The diacritic underneath vowels indicates an open vowel, pronounced with the root of the tongue retracted (so ⟨e⟩ is pronounced [ɛ] and ⟨ó⟩ is [ɔ]). ⟨ú⟩ represents a postalveolar consonant [ʃ] like the English ⟨sh⟩, ⟨y⟩ represents a palatal approximant like English ⟨y⟩, and ⟨j⟩ a voiced palatal plosive [ʃ], as is common in many African orthographies.

In addition to the vertical bars, three further diacritics are used on vowels and syllabic nasal consonants to indicate the language's tones: an acute accent ' for the high tone, a grave accent ` for the low tone, and an optional macron ¯ for the middle tone. These are used in addition to the line in ⟨e⟩ and ⟨ó⟩. When more than one tone is used in one syllable, the vowel can either be written once for each tone (for example, *⟨òó⟩ for a vowel [o] with tone rising from low to high) or, more rarely in current usage, combined into a single accent. In this case, a caron ^ is used for the rising tone (so the previous example would be written ⟨óó⟩) and a circumflex ^ for the falling tone.

In Benin, Yoruba uses a different orthography. The Yoruba alphabet was standardized along with other Benin languages in the National Languages Alphabet by the National Language Commission in 1975, and revised in 1990 by the National Center for Applied Linguistics.

(source of this subsection: [Wikipedia](#))

A	B	D	E	È	F	G	Gb	H	I	J	K	Kp	L	M	N	O	Ò	P	R	S	Sh	T	U	W	Y
a	b	d	e	È	f	g	gb	h	i	j	k	kp	l	m	n	o	Ò	p	r	s	sh	t	u	w	y

Table 21.2: Benin alphabet

Tones According to [Bamgbose \(1966\)](#), Yoruba (Defoid, Niger-Congo; Nigeria) has lexical distinctions between three level tones, as in the three verbs /bí/ ‘give birth’, /bī/ ‘ask’, /bi/ ‘vomit’.

According to [Wikipedia](#), Yoruba is a tonal language with three level tones: high, low, and mid (the default tone). Every syllable must have at least one tone; a syllable containing a long vowel can have two tones. Contour tones (i.e. rising or falling tone melodies) are usually analysed as separate tones occurring on adjacent tone bearing units (morae) and thus have no phonemic status. Tones are marked by use of the acute accent for high tone (⟨á⟩, ⟨ní⟩), the grave accent for low tone (⟨à⟩, ⟨ñ⟩); Mid is unmarked, except on syllabic nasals where it is indicated using a macron (⟨á⟩, ⟨ñ⟩)).

First, the use of the subdots and tone marks otherwise known as diacritic markings are not represented on the conventional keyboards. Therefore, most Yoruba documents are computer coded without the marks. Secondly, revealed that the use of the diacritics affect the retrieval of Yoruba documents by popular search engines.

When a word precedes another word beginning with a vowel, assimilation or deletion ('elision') of one of the vowels often takes place.[18] In fact, since syllables in Yoruba normally end in a vowel, and most nouns start with one, this is a very common phenomenon, and indeed only is absent in very slow, unnatural speech. The orthography here follows speech in that word divisions are normally not indicated in words that are contracted as a result of assimilation or elision: *ra ejá* → *reja* ‘buy fish’. Sometimes however, authors may choose to use an inverted comma to indicate an elided vowel as *in ní ilé* → *n'ilé* ‘in the house’.

Long vowels within words usually signal that a consonant has been elided word-internally. In such cases, the tone of the elided vowel is retained, e.g. *àdirò* → *ààrò* ‘hearth’; *koríko* → *koóko* ‘grass’; *òtító* → *òótó* ‘truth’.

(source of this subsection: [Wikipedia](#))

21.4 Previous research on the language

[Voeltz \(2006\)](#) and [Mugane \(2003\)](#) provide a detailed description about the typology of African languages.

[Jones \(2006\)](#) is an article about the semantics and pragmatics of focus in Yoruba.

21.4.1 Grammars

In this section some descriptions of the Yoruba language are listed. The first grammar of Yoruba is written by [Rowther \(1852\)](#).

[Ward \(1952\)](#) gives an introduction of Yoruba with a detailed description of the grammar.

[Bamgbose \(1966\)](#) is a full description of the language. This descriptive grammar derives from a large body of written and spoken texts in Standard Yoruba. In order to avoid the faults of traditional grammars, this study has been deliberately based on a structural theory, using Halliday's Scale and Category model. [Bamgbose \(1966\)](#) was the first full-length exemplification of this theory to be published, and will continue to be of interest to general linguists as well as to specialists in West African languages and Yoruba scholars.

However Adéwolé (2001a), Adéwolé (2001b) and Yetunde and Schleicher (2006) are made for language learners one can get a complete description of the grammar of Yoruba.

21.5 Data and sources

The World Atlas of Language Structures ([WALS](#)) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars). There are materials about [Yoruba](#).

[Omniglot](#) is an encyclopedia of writing systems and languages. It can be used to learn about languages, to learn alphabets and other writing systems, and to learn phrases in many languages. There is also advice on how to learn languages. There are some information about [Yoruba](#).

[OLAC](#), the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources. OLAC Archives contain over 100,000 records, covering resources in half of the world's living languages including [Yoruba](#).

There are some webpages provide general informations about the Yoruba language as [Africa.uge.edu](#) or [African Yoruba Pages](#) and [LanguagesGulper](#).

A very interesting resource of spoken Yoruba is [Online Yoruba Radio](#).

21.5.1 Basic vocabulary

[Swadesh lists](#) were originally devised by the linguist Morris Swadesh. In the 1940s to 1950s, Swadesh developed word lists of body parts, verbs, natural phenomena, in order to compute the relationships of languages, and in particular their age. A Swadesh list may also be useful to achieve knowledge of some universal terms in other languages. This is because, for basic communication, knowledge of vocabulary is more important than knowledge of grammar and syntax. Sometimes it is even possible to achieve (very) basic communication skills with no knowledge of the target language syntax whatsoever. [Niger-Congo Swadesh lists](#) including Yoruba is provided as well.

On the site of [mylanguages: yoruba vocabulary](#) the user can find the fundamental words translated to English.

There are more online resources for Yoruba phrases like [mylanguages: yoruba phrases](#) and [omniglot:yoruba phrases](#).

Learning Yoruba There are a lot of webpages helping the language learners. Here some examples provided.

- [MyLanguages](#)
- [LearnYoruba](#)
- [Polymath](#)
- [Transparent Language](#)
- [Countries and their cultures](#)

21.5.2 Dictionaries

Paper editions

There are a lot of paper dictionaries between Yoruba and an other language (mostly English) like [Abraham \(1958\)](#), [Wakeman \(1950\)](#), [Delano \(1958\)](#) and [Sachnine \(1997\)](#).

Online dictionaries

There are some online dictionaries for Yoruba. [yorubadictionary.com](#) is an English-Yoruba-English online dictionary with distinct search interface for Yoruba names and numerals. [aroadeyorubadictionary.com](#) is an English-Yoruba-English online dictionary as well, but it has a more traditional interface, it can be used as a paper edition (the search can be done manually, by turning a page, not typing in the word). [nigeriadictionary.com](#) contains other languages beside Yoruba (Igbo, Pidgin English, Hausa), and it has more functions than a simple online dictionary. For example the user can translate a less than thousand character long text. [Freelang](#) provides a Yoruba-English searchbox dictionary.

[Global Yoruba Lexical Database v. 1.0](#) is a set of related dictionaries providing definitions and translations for over 450,000 words from the Yoruba language and its variants: Standard Yoruba (over 368,000 words), Gullah (over 3,600 words), Lucumí (over 8,000 words) and Trinidadian (over 1,000 words).

Scraping [yorubadictionary.com](#) is a scrapable dictionary due to its relative big size and transparent structure. The words are collected into letters which contain pages. An entry always includes the Tagalog word form, the part-of-speech (or more part-of-speeches) and the English translation. Sometimes an expanded or varied word form or an example is included as well, and if there are more translations they are listed.

21.5.3 Corpora

The [Corplinguistics](#) blog has an entry about African language corpora. The blogger collected the most suitable sources and ideas finding (and making) big corpora for less common languages like Yoruba.

[Fagbolu et al. \(2015\)](#) introduces the Digital Yorùbá Corpus, which is a large and structured set of texts usually stored and processed in electronic form. The corpora can be used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language. The corpus aid in Natural Language Processing and Machine Translation, programming tools.

[Yoruba web as corpus](#) was compiled in June 2015 with encoding in UTF-8 and is not tagged yet. The corpus contains 2.8 million words.

[Indigenous Tweets](#) is a website that records minority language Twitter messages to help indigenous speakers contact each other. There are 285.655 entries by 2239 users in [Yoruba](#). There are blogs in [Yoruba](#) listed as well containing 10 blogs altogether with 139 posts (58.264 words).

There is a speech corpus for Yoruba as well. [The Lagos-NWU Yoruba Speech Corpus](#) consists of 16 female speakers and 17 male speakers. Their speech was recorded in Lagos, Nigeria for the purpose of speech recognition research. Each speaker recorded about 130 utterances read from short texts selected for phonetic coverage. Recordings were done using a microphone connected to a laptop computer in a quiet office environment. The corpus is available for download.

Wikipedia Wikipedia offers free copies of all available content to interested users. Wikimedia is a global movement whose mission is to bring free educational content to the world. The content of Wikimedia is dumped and available as well. Yoruba Wikipedia contains 31.473 articles as of 21/10/2016.

These dumps are available in Yoruba in 21/10/2016:

- [Yoruba wikipedia](#)
- [Yoruba wikibooks](#)
- [Yoruba wiktionary](#)

Bible The various translations of Bible and other religious texts are great sources for collecting data and corpora from a language. Furthermore these texts are perfect for making parallel corpora which is very useful for the task of machine translation. In this subsection find some webpages that provide Yoruba translation of the Bible.

The [bible.com](#) webpage provides hundreds of languages and translations of the Bible for reading and downloading including Yoruba. There are three versions of the Yoruba Bible:

- [BIBELI MIMÓ \(BM\)](#)
- [Bíbélí Mímó ní èdè Yorùbá òde òní \(BMY\)](#)
- [Yoruba, Bible \(YCE\)](#)

The [Jehovah's Witnesses](#) webpage is available in 772 languages including [Yoruba](#). Among articles and other contents the Bible is available in these languages.

Bilingual United Nations Human Rights Office of the High Commissioner [OHCHR](#) worldwide collection of materials on the Universal Declaration of Human Rights (UDHR), including various resources developed by governmental and non-governmental organizations both on the occasion of the Declaration's 50th Anniversary (1998) and prior to/after the Anniversary year. The collection is unique in the world and comprises more than 400 items. Since UDHR is the most translated document – and it is available in [Yoruba](#) – it can be used as a multilingual corpus.

21.5.4 News portals

[Alaroye](#) and [Oláyemí Oníròyìn](#) are online newspapers.

21.5.5 Contact person

A lot of universities provide opportunity to learn Yoruba studies like [The University of Texas at Austin](#), [University of Florida](#), [University of Wisconsin – Madison](#) and [Harward University](#). To contact the departments all websites provide address and e-mail address.

21.6 Computational tools

The [African Language Technology](#) contains a steadily growing collection of bibliographic resources, web links and tools, provided by African Language Technology members. There are some tools for

african languages, among others there is a demonstraion system for a diacritic restoration method that is able to automatically restore diacritics on the basis of local graphemic context. It is based on the machine learning method of Memory-Based learning. They have applied the method to the African languages of Cilubà, Gíkúyú, Kíkamba, Maa, Sesotho sa Leboa, Tshivenda and Yoruba [De Pauw et al. \(2007\)](#).

[Olúgbéngá O. Akinadé \(2014\)](#) presents a tool which a computational system that is capable of converting cardinal numbers to their equivalent Standard Yorùbá number names. the process involved in formulating a Context-Free Grammar (CFG) to capture the structure of the Yorùbá numeral system was highlighted. The model was reduced into a set of computer programs to implement the numerical to lexical conversion process. System evaluation was done by ranking the output from the software and comparing the output with the representations given by a group of Yorùbá native speakers. The result showed that the system gave correct representation for numbers and produced a recall of 100% with respect to the collected corpus.

According to [Dione et al. \(2010\)](#), their work is, to our knowledge, the first effort in building a publicly available NLP resource for the Wolof language. More generally, there has recently been a growing interest in NLP technologies for African languages ([PDF](#)). In recent times a number of researchers and institutions, both from Africa and elsewhere, have come forward to share the common goal of developing capabilities in language technologies for the African languages. The goal of [De Pauw et al. \(2009\)](#), then, is to provide a forum to meet and share the latest developments in this field. It also seeks to attract linguists who specialize in African languages and would like to leverage the tools and approaches of computational linguistics, as well as computational linguists who are interested in learning about the particular linguistic challenges posed by the African languages.

In the proceedings mentioned above [Finkel and Odejobi \(2009\)](#) demonstrate the use of default default inheritance hierarchies to represent the morphology of Yorùbá verbs in the KATR formalism, treating inflectional exponences as markings associated with the application of rules by which complex word forms are deduced from simpler roots or stems. In particular, they suggest a scheme of slots that together make up a verb and show how each slot represents a subset of the morphosyntactic properties associated with the verb. They also show how we can account for the tonal aspects of Yorùbá, in particular, the tone associated with the emphatic ending.

21.6.1 Language identification

[Selamat and Akosu \(2015\)](#) defines under-resourced languages as those languages that do not have (or not enough) digital resources that can be employed for extensive research. The native speakers of such languages either do not use computers or if they do it is usually via a foreign language. This research is focused on languages with little or no digital resources, hence the name ‘under-resourced languages’ such az Yoruba.

[Compact Language Detector 2](#) probabilistically detects over 80 languages including Yoruba in Unicode UTF-8 text, either plain text or HTML/XML.

21.6.2 Tokenizer

The [Test Word Tokenizer service using GET](#) and the [MorphAdorner V2.0](#) are online tokeizers for a lot of languages including Yoruba.

[Kumolalo et al. \(2010\)](#) indtroduces a rule-based syllabicator, a tool that takes a word and returns the syllables that the word is composed of.

21.6.3 Stemmer

I have found no stemmer tool for Yoruba.

21.6.4 Spell checker

There are some online tools for spell checking like [SpellChecker](#).

21.6.5 Phrase-level and higher tools

Adeoye et al. (2013) developed a computational model for English language and Yoruba language noun-phrases involve a profound understanding of the syntactic and grammatical features of the two languages as well as their vocabularies since they are not related syntactically and grammatically. They also developed a bilingual lexicon which is made up of words in English language with their corresponding Yoruba counterparts and their equivalent part-of-speech. The model was implemented using PhP programming language and MySQL and was tested on one 160 randomly selected noun-phrases from daily news, and gives accuracy of 91% which is quite encouraging. The system if fully developed will go a long way in preventing the extinction threat of the Yoruba language.

POS-tagger Oluwatoyin (2015) proposes a POS-tagger for Yoruba and presents a new method by which the language can be transformed into a computer understandable language using it's morphological identification framework.

Sèmiyou A. et al. (2013) aimed to design a yoruba corpus. The main motivation of their work was to obtain training data for PoS taggers and to provide applications of Yoruba Language Processing (YLP) with a basic tool. The tagging was performed with SVMTool, a widely-used POS tagger. The corpus with 312,562 words, formed from the Web, was annotated with an accuracy of 98.04%. This annotated corpus might be used in translation system.

Speech recognition In Frank and James (2013) the speech recognition system proposed digitizes the isolated words spoken by a speaker and performs Mel Frequency cepstral analysis and other signal processing techniques on the digitized data. The processed speech signal is then passed on to a pattern recognition which takes action based on the type of command pattern received. Artificial Neural Network (ANN) is used as speech recognition engine. Two different corpora were collected of audio recordings of Yoruba, Igbo and Hausa language speakers, in which subjects read aloud different words. One of the collected corpora contained data with background noise and the other without background noise. The results obtained from simulation can be generalized to cater for larger vocabularies and for continuous speech recognition.

Text-to-speech application Afolabi et al. (2013) gives an account of Yoruba TTS system development using concatenation method. The paper describes the design, evaluation and the analysis of the result shows that 70% Respondents accepted its usability.

Machine translation O.b et al. (2015) looks at the various approaches to machine translations and future needs in order to provide more robust and sensible system in the area of natural language processing which will be resistant and impervious to failure regardless of users' inputs. It is hopeful that researchers in the area of language processing can make use of our valuable improvement and suggestions.

Adeoye et al. (2014) developed a computational model for English and Yoruba noun-phrases that involves a profound understanding of the syntactic and grammatical features of the two languages as well as their vocabularies, since they are not related syntactically and grammatically.

Akinwale et al. (2015) deals with the translation of English text to Yoruba text using rule based method. Twenty two rules were formulated for the translation which is specified using context free grammar. A bilingual dictionary dataset containing English words and the corresponding translation in Yoruba language was used.

Folajimi (2015) describes Statistical Machine Translation (SMT) system that translates English sentences to Yoruba sentences. The resulting software provides tools to tackle the problem of language translation between Yoruba (Nigeria language) and English language. The main challenge with the Yoruba language is that there is no English -Yoruba parallel corpus, hence we need to create English -Yoruba parallel corpus. The software employs a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. Design architecture and experimental results revealed that statistical machine translation is indeed a veritable tool for translating between English and Yoruba languages due to the fact that there is no parallel corpus between the English and Yoruba languages.

Using Statistical Machine Translation (SMT) as a Language Translation Tool for Understanding Yoruba Language ([PDF](#)).

Eludiora et al. (2011) discussed IFE-MT (<http://ifecisrg.org/ifemt>) a machine translation system for translating English statements into Yorùbá, the theoretical and practical challenges they encountered during the development of the system and how they were addressed. They have also discussed the design of the language resource and the system architecture in terms of its structure and configuration as well as the software design.

Eludiora (2014) PhD thesis also introduces a Yoruba Machine Translation System.

21.6.6 End-user support

[Symbolcodes](#) describes the basics of Unicode for foreign language and what kinds of utilities are needed for foreign language support. Among others it has a page about Yoruba.

Yoruba is written in the Roman alphabet but includes dotted vowels, and so requires special font keyboard support separate from languages like English, Spanish and French.

The following fonts include characters for dotted letters found in Yoruba spelling:

- Windows – Arial Unicode MS (Windows), Tahoma
- Macintosh OS X – Lucida Grande
- Doulos SIL – Includes Greek, Cyrillic
- Gentium – From SIL. Very readable
- TITUS Cyberbit – includes characters from many scripts such as Armenian, Cyrillic, Greek, Coptic and more.
- Aboriginal Serif – Includes Cherokee, Canadian Aboriginal, Central Asian Cyrillic

Operating Systems [Learn Yoruba](#) webpage provides a simple keyboard input software for typing Yoruba text in Windows OS. The font is available and downloadable from the webpage. Unlike

the regular English language alphabet, typing accented characters or the letter of other languages is usually a tedious process requiring multiple keystroke combinations to produce a single character. This usually involves depressing the Ctrl, Alt and Shift keys in combination with a letter character or a series of numbers which are the assigned ASCII code numbers for that particular “foreign” letter. Moreover, it is also necessary to have a chart of the ASCII code numbers handy or memorize the numbers to type in to produce each special character.

On Macintosh for dotted vowels the user can switch to the Extended Roman keyboard (10.2) or the U.S. Extended keyboard (10.3). For print work, there are also a number of freeware and shareware phonetics and classics fonts. [ScriptSource](#) has been designed to make it easy to locate software for particular languages and writing systems (fonts, keyboards, transliteration software etc.).

Browsers For dotted consonants the following browsers have the most consistent results:

- Firefox
- Mozilla
- Opera
- Safari 3+ (Note: Safari 2 does not support combining accents needed for Yoruba)

Internet Explorer for Windows may not display implosive consonants by default. Users who prefer Internet Explorer for Windows should set the Latin font to Arial Unicode MS or some other Unicode script with phonetic symbol support. Internet Explorer for Macintosh does not support implosive consonant symbols.

Bibliography

- Roy Clive Abraham, editor. *Dictionary of Modern Yoruba*. University of London Press, London, 1958.
- O. B. Adeoye, A. O. Adetunmbi, and A. Oguntimilehin. A Computational Model of English to Yoruba Noun-phrases Translation System. *FUTA Journal of Research Sciences*, 1:34–43, 2013.
- O. B. Adeoye, A. O. Adetunmbi, and K. A. Olatunji. A web-based english to yoruba noun-phrases machine translation system. *International Journal of English and Literature*, 5(3):71–78, 2014.
- Olóyé Lásún Adéwolé. *Beginning Yorùbá Part I*. Number 9 in Monograph Series. CASAS, Cape Town, 2001a.
- Olóyé Lásún Adéwolé. *Beginning Yorùbá Part II*. Number 10 in Monograph Series. CASAS, Cape Town, 2001b.
- Akin Afolabi, Elijah Omidiara, and Tayo Arulogun. Development of Text to Speech System for Yoruba Language. *Innovative Systems Design and Engineering* , 4(9), 2013.
- O. I. Akinwale, A. O. Adetunmbi, O. O. Obe, and A. T. Adesuyi. Web-Based English to Yoruba Machine Translation. *International Journal of Language and Linguistics*, 3(3):154–159, 2015.
- Ayo Bamgbose. *A Grammar of Yoruba*. The Cambridge University Press, Cambridge, 1966.
- Walter Bisang. Some general thoughts about linguistic typology and dialogue linguistics. In Marion Grein and Edda Weigland, editors, *Dialogue and Culture*, pages 53–72. John Benjamins Publishing Company, Mainz, 2007.
- Guy De Pauw, Peter W. Wagacha, and Gilles-Maurice de Schryver. Automatic diacritic restoration for resource-scarce languages. In *Proceedings of Text, Speech and Dialogue, Tenth International Conference*, Heidelberg, Germany, 2007. Springer Berlin / Heidelberg.
- Guy De Pauw, L. Levin, and G. M. de Schryver, editors. *Proceedings of the workshop on Language Technologies for African Languages (AfLaT 2009)*, Athens, Greece, 2009. Association for Computational Linguistics. ISBN 1-932432-25-6.
- Oloye Isaac Delano, editor. *Atúmò ede Yoruba (short dictionary and grammar of the Yoruba language)*. University of London Press, London, 1958.
- Cheikh M. Bamba Dione, Jonas Kuhn, and Sina Zarrieß. Design and Development of Part-of-Speech-Tagging Resources for Wolof (Niger-Congo, spoken in Senegal). In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language*

Resources and Evaluation, LREC'10, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

Safiriyu Ijiyemi Eludiora. *Development of an English to Yoruba Machine Translation System*. PhD thesis, Obafemi Awolowo University, Ile-Ife, Osun State, 2014.

Safiriyu Ijiyemi Eludiora, S. A. Salawu, Odetunji Ajadi Odejobi, and A. O. Agbeyangi. IFE-MT: An English-to-Yorùbá Machine Translation System. In *AGIS11 – Action Week for Global Information Sharing (AfLaT2011 Breakout Session)*, Addis Ababa, Ethiopia, 2011.

Olutola Fagbolu, Akinwale Ojoawo, Kayode Ajibade, and Boniface Alese. Digital Yorùbá Corpus. *International Journal of Innovative Science, Engineering & Technology*, 2(8):918–926, 2015.

Raphael Finkel and Odetunji Ajadi Odejobi. A Computational Approach to YorÙBÁ Morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*, AfLaT '09, pages 25–31, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Yetunde Folajimi. Using Statistical Machine Translation (SMT) as a Language Translation Tool for Understanding Yoruba Language. In *EIE's 2nd Intl' Conference on Computer, Energy, Network, Robotics and Telecommunication, At Otta, Nigeria*, 2015.

Ibikunle Frank and Katende James. Recognition of Nigerian Major Languages Using Neural Networks. *Journal of Computer Networks*, 1(2):32–37, 2013.

Susie Jones. Focus in Yorùbá: a semantic/pragmatic account. In *Papers on information structure in African languages*. Zentrum für Allgemeine Sprachwissenschaft, Sprachtypologie und Universalienforschung, Berlin, 2006.

F. O. Kumolalo, E. R. Adagunodo, and O. A. Odejobi. Development of a Syllabicator for Yorùbá Language. In *Proceedings of OAU TekConf*, pages 47–51, 2010.

John Mugane, editor. *Linguistic typology and representation of African languages*. Trenton, NJ: Africa World Press, 2003.

Abiola O.b, Adetunmbi A.o, and Oguntimilehin. A. Article: A review of the various approaches for text to text machine translations. *International Journal of Computer Applications*, 120(18):7–12, June 2015.

Odétúnjí A. Odéjóbí Olúgbéngá O. Akinadé. Computational modelling of Yorùbá numerals in a number-to-text conversion system. *Journal of Language Modelling*, 2(1), 2014.

Enikuomehin A. Oluwatoyin. A Computerized Identification System for Verb Sorting and Arrangement in a Natural Language: A Case Study of the Nigerian Yoruba Languages. *European Journal of Computer Science and Information Technology*, 3(1):43–52, 2015.

Evan Colyn Rowlands. *Teach Yourself Yoruba*. English Universities Press, London, 1969.

Samuel Ajayi Rowther. *The Grammar of the Yoruba Language*. Seeleys, London, 1852.

Michka Sachnine, editor. *Dictionnaire yorùbá-français, suivi d'un index français-yorùbâ*. Karthala, Paris, 1997.

- Ali Selamat and Nicholas Akosu. Word-length algorithm for language identification of under-resourced languages. *Journal of King Saud University - Computer and Information Sciences*, 2015. in press.
- Adedjouma Sèmiyou A., John O. R. Aoga, and Mamoud A. Igue. Part-of-Speech tagging of Yoruba Standard, Language of Niger-Congo family. *Research Journal of Computer and Information Technology Sciences*, 1(1):2–5, 2013.
- Erhard F. K. Voeltz, editor. *Studies in African Linguistic Typology*, volume 64. of *Typological Studies in Language*. John Benjamins Publishing Company, 2006.
- Canon C.W. Wakeman, editor. *A Dictionary of the Yoruba language*. University Press, Ibadan, 1950.
- Ida Caroline Ward. *An introduction to the Yoruba language*. W. Heffer, 1952.
- William E. Welmers. *African Language Structures*. University of California Press, Berkeley / Los Angeles, 1973.
- H. Wetzer. *The typology of Adjectival Predication*. De Gruyter Mouton, Berlin, Boston, 2013.
- Antonia Yetunde and Folarin Schleicher. *Colloquial Yoruba*. Taylor and Francis Ltd (Routledge), London, 2006.

Chapter 22

Zulu (Ekaterina Georgieva)

Contents

22.1 Demography and ethnography	337
22.2 Main typological and syntactic features	340
22.3 Writing system, transcription	343
22.4 Previous research on the language	343
22.5 Data and sources	344
22.6 Computational tools	347
Bibliography	350

This chapter deals with Zulu, a representative of the Bantu language family, mainly spoken in South Africa. After a brief outline of the ethno-linguistic situation of the language, the main typological features of Zulu will be discussed. This is followed by an introduction to its writing system and a review of literature dealing with the Zulu grammar. The final sections round up the available linguistic data and other resources related to the language as well as deal with the available computational tools.

22.1 Demography and ethnography

22.1.1 Name variants

Zulu (endonym: *isiZulu*) is spoken by the Zulu people, a Bantu ethnic group and the largest ethnic group in South Africa. On the history and ethnography of the Zulu people the interested reader is pointed to the following articles of the [Encyclopædia Britannica](#): [Zulu ethnic group](#), [Zululand](#) and [Anglo-Zulu War](#).

As far as its linguistic classification is concerned, Zulu is a Bantu language belonging to the Benue-Congo subgroup of the Niger-Congo language family. Within the Bantu subgroup, Zulu belongs to the Nguni languages together with Xhosa, Swati and Ndebele.

The ISO 639-3 identifier of Zulu is **zul**.

22.1.2 Geographic spread

The vast majority of Zulu speakers reside in South Africa, predominantly in the KwaZulu-Natal province. The following map (22.1) illustrates the geographic distribution of Zulu in South Africa.

Additionally, Zulu is also spoken in Botswana, Lesotho, Malawi, Moyambique, Swaziland and Zimbabwe (source: [Joshua Project](#)).

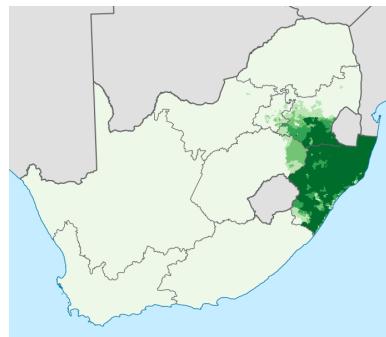


Figure 22.1: Proportion of the South African population that speaks Zulu as their first language, according to Census 2011 at electoral ward level (source: [Wikimedia Commons](#))

Country	Number of primary speakers (Joshua Project)	Number of L1 speakers (Ethnologue)	Number of L2 speakers	EGIDS level
South Africa	12,559,000	11,600,000	15,700,000	1
Lesotho	324,000	248,000	no data	5
Zimbabwe	167,000	no data	no data	no data
Swaziland	107,000	76,000	no data	5
Malawi	66,000	37,500	no data	5
Mozambique	6000	3000	no data	5
Botswana	5900	no data	no data	no data
Total	13,234,900	11,969,100	15,700,000	

Table 22.1: The Zulu speakers

22.1.3 Speaker populations

According to the [Ethnologue](#), Zulu has 11,600,000 native speakers in South Africa (the number of speakers is increasing), and thus, it is the most widely used language in domestic context in South Africa (24% of the population uses it at home). Moreover, there are 15,700,000 L2 users in South Africa alone. In total, there are 27,669,100 Zulu speakers in all countries (11,969,100 L1 and 15,700,000 L2 users).

Table 22.1 summarizes the data about the number of primary/L1 speakers according to two sources: [Joshua Project](#) and [Ethnologue](#). Additionally, it provides information about the number of L2 speakers and the EGIDS level of Zulu in the individual countries (based on the [Ethnologue](#)).

According to the [Ethnologue](#), the EGIDS level for the Zulu language is 1 (National), cf. Figure 22.2. This means that the language is used in education, work, mass media, and government at the national level. Additionally, the EGIDS level for the Zulu language in Lesotho, Malawi, Mozambique and Swaziland is 5 (dispersed). This means that the language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.

22.1.4 Dialect situation

According to the [Ethnologue](#), the main dialects of Zulu are the following: Lala, Qwabe, Cele and Transvaal Zulu. Alternatively, [Maho \(2009\)](#) offers a different classification with four dialects: central KwaZulu-Natal Zulu, northern Transvaal Zulu, eastern coastal Qwabe, and western coastal Cele. [Kubeka \(1979\)](#) distinguishes six dialects in KwaZulu-Natal (formerly, Natal and Zululand): Central Zululand dialect, Zululand Coast dialect, Natal Coast dialect, Lower Natal Coast dialect and South West Natal Coast dialect. The differences between these dialects are observed in the phonology and the morphology. Additionally, [Downing \(2001\)](#) discusses differences between the tonal patterns in the Zulu dialects, [Cook \(2013\)](#) deals with reduplication patterns and their dialectal variation.

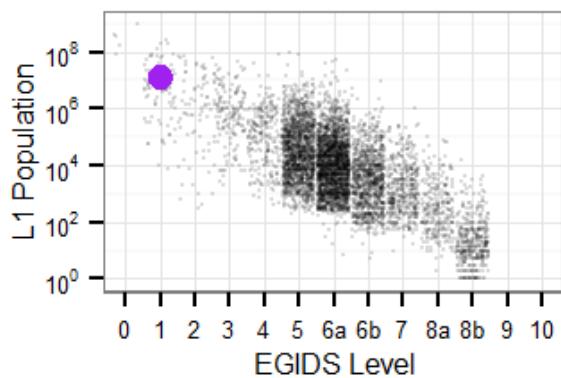


Figure 22.2: Zulu – EGIDS Level ([Ethnologue](#))

An additional distinction is made between standard Zulu, also called *deep Zulu* (*isiZulu esijulile*), and non-standard, i.e. urban Zulu (*isiZulu sasedolobheni*). Standard Zulu tends to be purist, using derivations from Zulu words for new concepts, whereas speakers of urban Zulu are more open to loan words, mainly from English (on the differences between standard and non-standard Zulu see [Magagula 2009](#)).

A remark is in order regarding the similarities between the Nguni languages. It is noteworthy that Zulu is reportedly similar and mutually intelligible with the Swati and Xhosa languages. Additionally, some studies assume that (Northern) Ndebele (spoken in Zimbabwe) is a dialect of Zulu, see [Kubeka \(1979, pp. 9\)](#) and [Doke and Vilakazi \(1972, pp. 13\)](#). However, other sources list it as a separate language, see for instance the [Ethnologue's entry](#) of Ndebele (its alternate names are: Isinde'bele, Ndebele of Zimbabwe, Northern Ndebele, Sindebele, Tabele, Tebele; its ISO 639-3 identifier is *nde*).

What makes matters even more complicate is the fact that the name ‘Ndebele’ is also used for another Nguni language that is spoken in South Africa, see [Ndebele](#) (its alternate names are: Isikhethu, IsiNdebele, Ndzundza, Nrebele, Southern Ndebele, Transvaal Ndebele; its ISO 639-3 identifier is *nbl*). A further complication is that Transvaal Ndebele (ISO 639-3 *nbl*) has two varieties, namely Southern and Northern Ndebele. It is a matter of debate whether these are distinct languages or dialects of one and the same language, see [Skhosana \(2009\)](#) for an in-depth discussion of the topic from historical, ethnocultural and linguistic perspectives. The conclusion of the author is that we are dealing with two different languages ([Skhosana, 2009, pp. 463](#)). Hence, the name ‘Ndebele’ applies to three different entities.

([Herbert and Bailey, 2002, pp. 66-68](#)) provide an overview of the sociolinguistic situation of the Nguni languages. They argue that the effect of language standartization on the development of the separate Nguni languages is considerable. Furthermore, given the wide-spread multilingualism in South Africa, they claim that language census data are unreliable since it is not taken into account what counts as a ‘speaker’. The question of the Ndebele languages is also addressed:

“[t]he zeal with which the Nationalist homeland policy was implemented lead to the elevation of several minor dialects to the status of official languages, e.g. Ndebele (Southern Transvaal Ndebele) and Swati, both North Nguni dialects. Until the 1980’s, these languages were classed as Zulu dialects, and Zulu materials were used without major difficulty. Ndrebele [sic] (Northern Transvaal Ndebele) was spoken over dispersed an area for a homeland to be consolidated while a history of widespread bilingualism with Pedi rendered it bureaucratically unnecessary. Thus, the ‘tenth’ indigenous Bantu

language of South Africa was rendered unnecessary and obsolete.” ([Herbert and Bailey, 2002](#), pp. 75, fn. 12)

22.2 Main typological and syntactic features

This section provides an overview of the main typological features of Zulu (further information can be found in the [WALS](#) entry of Zulu). Additionally, several constructions, such as predication, possession, imperative and interrogative clauses will be discussed.

22.2.1 Linguistic typology

With respect to **phonology**, Zulu has a moderately large consonant inventory and an average vowel inventory ([Maddieson, 2013a,d](#)). One of the most distinctive features of Zulu is the use of *click consonants* ([Maddieson, 2013b](#)). Like the great majority of other Bantu and African languages, Zulu is tonal (according to [Maddieson 2013c](#), Zulu has a simple tone system).

As far as its **morphology** is concerned, Zulu shows all typical features of the Bantu language family. More specifically, Zulu is an agglutinative language, it has a rich array of noun classes and an extensive system of inflections for person (both subject and object), tense and aspect marking.

Zulu nouns are classified into several morphological classes, each of which agrees with different singular and plural prefixes ([Poulos and Bosch 1997](#) distinguish 17 classes). Morphologically, Zulu nouns consist of two essential parts, namely the prefix and the stem. For ease of classification, prefixes are used to group nouns into noun classes. For example, the word *umfana* ‘boy’ contains the prefix *um-*, hence it belongs to Class 1 (see [Poulos and Bosch 1997](#), pp. 4). In Class 1, the plural form is derived by using the prefix *aba-*, and accordingly, *abafana* ‘boys’ is the plural form of the noun *umfana* ‘boy’. Furthermore, the class of the noun determines the form of other parts of speech, such as adjectives that show concord with the noun.

Zulu verbs show highly complex morphology. Full-fledged verb forms are realized by the addition of prefixes and suffixes to the verbal stem. These include subject prefixes (which agree with the subject of the sentence), object prefixes (which agree with the object of the sentence), tense affixes, modality affixes and negation. Furthermore, Zulu verbs can take several (derivational) affixes (causative, reflexive, and reciprocal, among others). Additionally, there are several non-finite verb forms, such as participle, infinitive and subjunctive. The interested reader is pointed to [Buell \(2005\)](#) who provides a detailed description of the Zulu verbal morphosyntax from a theoretical perspective.

With regards to **syntax**, Zulu shows SOV word order that is also typical of the Bantu languages in general. Furthermore, Zulu shows SV and OV word order, and it uses prepositions. (For additional information on other syntactic features consult the [WALS](#) entry of the Zulu language.)

22.2.2 Predication

Simple intransitive clauses are exemplified below ([Poulos and Bosch, 1997](#), pp. 17). In this example, the verb shows subject agreement (it inflects with the 3SG prefix *u-*) and present tense morphology (expressed by the prefix *ya-* and the suffix *-a*).

- (156) *Umfana uyasebenza.*
 umfana u-ya-sebenza-a
 CLASS1.boy CLASS1.SUBJ-PRS-work-PRS

‘The boy is working.’

If the verb is transitive (i.e. it has a direct object), it agrees not only with the subject but with the direct object as well. Similarly to subject agreement, object agreement must agree in class with the noun it refers to. In the following example, the verb inflects with the object agreement prefix *yi-* (Poulos and Bosch, 1997, pp. 24).

- (157) *Umfana uzoyishaya (inja).*
 umfana u-zo-yi-shay-a (inja)
 CLASS1.boy CLASS1.SUBJ-FUT-CLASS4.OBJ-hit-PRS (CLASS4.dog)

‘The boy will hit it (the dog).’

Zulu makes use of non-verbal predicates as well. In this case, Zulu does not employ a copula similar to English ‘to be’ but rather uses a special set of prefixes on the non-verbal predicate (Poulos and Bosch, 1997, pp. 34-38). The choice of prefix depends on the initial vowel of the noun and the class of the noun. Special forms of these prefixes are used in the first and second person. Stassen (2013) uses the term “pro-copula” for such affixes. In the example below, the nominal predicate *uthisha* ‘teacher’ is preceded by the prefix *ng-* that serves as a copula.

- (158) *Usipho nguthisha.*
 Usipho ng-uthisha
 Sipho COP-teacher

‘Sipho is a teacher.’

In the next example, the non-verbal predicate is an adjective. The prefix used on it is *mu-*.

- (159) *Lo mfana mude.*
 Lo mfana mu-de
 DEM boy COP-tall

‘The boy is tall.’

22.2.3 Possession

In Zulu, possessive constructions show the following word order: the *possessee* (or *possessed, posses-sum*) precedes the *possessor* (Dryer, 2013a). Additionally, possessive concord affixes are used (Poulos and Bosch, 1997, pp. 39). In the example below, the possessee is *inja* ‘dog’ and the possessor is *umfana* ‘boy’. The possessor inflects with the possessive concord prefix *ya-* (this prefix corresponds to Class 9 since the possessee belongs to Class 9) (source: Poulos and Bosch 1997, pp. 42).

- (160) *injā* *yomfana*
 injā *ya-umfana*
 CLASS9.dog CLASS9.POSS-boy
 ‘the boy’s dog’

Additional information on other possessive constructions in Zulu can be found in [Poulos and Bosch \(1997, pp. 39-43\).](#)

22.2.4 Imperative

In Zulu, imperatives can be formed in the second person singular and plural. The imperative verb forms contain the suffix *-a*, and, in the case of the plural forms, the pluralizer (*-ni*), but no subject agreement is required (for more examples see [Poulos and Bosch 1997, pp. 19](#)).

- (161) *Sebenza!*
 sebenz-a
 work-IMP
 ‘Work (sg)!’

- (162) *Sebenzani!*
 sebenz-a-ni
 work-IMP-PL
 ‘Work (pl)!’

Prohibitions are expressed in a different way. According to [van der Auwera et al. \(2013\)](#), the Zulu second-person singular prohibitive is different from the second singular imperative, and the negative strategy is different from the sentential negative found in declaratives. In Zulu, prohibitives use the infinitive form (but not the regular imperative verb form from shown above) ([van der Auwera et al., 2013](#)). The infinitive is combined with a special negative marker *musa*, different from the indicative declarative sentential negative marker. This is shown in the following example ([van der Auwera et al., 2013](#)):

- (163) *Musa* *ukushaya* *inga!*
 mus-a uku-shay-a inga
 NEG.IMP.AUX-2SG INF-hit-INF dog
 ‘Don’t hit the dog!’

22.2.5 Interrogative

Zulu uses both *content* and *polar* questions. Content questions are introduced by an interrogative phrases, such as *kuphi* ‘where’, and similarly to the way it is illustrated below, they do not occupy sentence initial position ([Dryer, 2013b](#)):

a	bh	b	tsh	d	e	f	gh	h
[a]	[b̥]	[b̥]	[tʂ̥]	[d̥]	[ɛ~e]	[f̥]	[g̥]	[h̥]
i	j	kh	k	kl	l	hl	dl	m
[i]	[ɸʒ̊ʱ]	[kʰ]	[k̊~g̊]	[kx̊]	[l̊]	[ɿ̊]	[ɿ̊]	[m̊]
n	ny	ng	o	ph	p	r	s	sh
[n]	[n̥]	[ŋ̥]	[ɔ~o]	[pʰ]	[p̊]	[r̊]	[s̊]	[ʃ̊]
th	t	u	v	w	hh	y	z	zh
[tʰ]	[t̊]	[u]	[v]	[w]	[ɸ̊]	[j̊]	[z̊]	[ʒ̊]
Click consonants								
c	ch	gc	nc	ngc				
[k̥]	[kʰ̥]	[ɸ̥̊]	[ŋ̥̥]	[ŋ̥̥̊]	<i>dental clicks</i>			
q	qh	gq	nq	ngq				
[k!̥]	[k!̥̥]	[ɸ!̥̥]	[ŋ!̥̥]	[ŋ!̥̥̊]	<i>(post)alveolar clicks</i>			
x	xh	gx	nx	ngx				
[k ̥]	[k ̥̥]	[ɸ ̥̥]	[ŋ ̥̥]	[ŋ ̥̥̊]	<i>alveolar lateral clicks</i>			

Figure 22.3: Zulu script – sample text (source: [Omniglot](#))

- (164) *uhlala* *kuphi?*
 u-hlal-a kuphi
 2SG.SUBJ-live-PRS where

‘Where do you live?’

Polar questions are formed with a question particle, such as *na* in the sentence below (see also Dryer 2013c):

- (165) *Ukhulumama* *isiZulu na?*
 u-khulum-a isiZulu na
 2SG.SUBJ-speak-PRS Zulu Q

‘Do you speak Zulu?’

22.3 Writing system, transcription

Historically, Zulu, like most indigenous South African languages, lacked a written form prior to the contact with missionaries from Europe, who documented the language using the Latin script. Like most of the Bantu languages, Zulu preserved this tradition and still uses Latin script. Zulu employs all 26 letters of the standard Latin alphabet and also marks additional phonemes by using sequences of multiple letters. Tone, stress and vowel length are not indicated in writing. The Zulu alphabet is illustrated in Figure 22.3.

22.4 Previous research on the language

The first grammar of the Zulu language was published in Norway in 1850 by the Norwegian missionary Hans Schreuder. Other descriptive grammars of the language include: Poulos and Bosch (1997),

Ziervogel et al. (1967), Cope (1984), Canonici (1996), Doke (1963), O’Neil (1912), Poulus and Msimang (1998) and Nyembezi (1957). The interested reader may also wish to consult the [WALS](#) and [OLAC](#) entry of the Zulu language.

22.4.1 Journals

Some of the major journals dealing with African languages are *African Studies* (formerly *Bantu Studies*), *Journal of African Languages and Linguistics*, *Journal of African Cultural Studies*, and *Southern African Linguistics and Applied Language Studies*.

22.4.2 Research and control bodies

The [Department of African Languages](#) at the University of South Africa and the University of Pretoria’s [Department of African Languages](#) have several experts working on the African languages, including Zulu.

Zulu is one of South Africa’s eleven official languages. The written form of Zulu was controlled by the Zulu Language Board of KwaZulu-Natal. This board has now been superseded by the [Pan South African Language Board \(PanSALB\)](#) which aims at promoting and creating conditions for the development and use of all eleven official languages of South Africa.

22.5 Data and sources

This section reviews the available sources for the Zulu language, such as vocabularies, paper-based and online dictionaries as well as various corpora. In addition, it provides an overview of the news portals available on the internet. Additionally, the interested reader may wish to consult the documentation of the [SeLa Project](#) which aims developing electronic dictionaries and terminology databases for the (South) African Bantu Languages.

22.5.1 Basic vocabulary

The *An Crúbadán* project provides [Zulu resources](#) (consisting of character trigrams, word bigrams and word frequency tables) compiled from 1714 documents, with a total of 3,256,104 words. Furthermore, a Zulu [Swadesh list](#) is also available.

Additionally, there are quite a few learners’ sources available for Zulu, such as vocabulary lists: [Zulu with Dingani](#) and [LearnZulu](#). Jason Wolfe’s [Useful Zulu Medical Dictionary](#) summarizes the most important medical terms and phrases in Zulu. [Ilman \(2014\)](#) is a useful English–Zulu phrasebook with pronunciation of the Zulu words. Additionally, [Mylanguages](#) provides useful language-learning resources for many languages, including Zulu (these include basic vocabulary, thematically grouped into word classes, as well as basic grammar).

22.5.2 Dictionaries

This section summarizes the Zulu dictionaries (both paper-based and online ones).

Traditional (paper-based) dictionaries

Beside the latest monolingual Zulu dictionary, i.e. [Mbatha \(2006\)](#) (1353 pp.), some of the bilingual English–Zulu dictionaries are listed below. (Note that there are several editions/reprints of some of

these dictionaries.)

- [Doke and Vilakazi \(1972\)](#) contains 927 pages (the entries include word-class labels as well). The words are grouped around word stems (but without the prefixes).
- [Bryant \(1905\)](#) contains about 20,000 entries with word-class information (altogether 778 pages). [Bryant's dictionary](#) is downloadable in various formats.
- [De Schryver \(2015\)](#) contains 672 pages;
- [Hartshorne et al. \(1985\)](#) contains 645 pages;
- [Doke et al. \(1958\)](#) contains 592 pages;
- [Africa \(1991\)](#) is a 5000-word dictionary that translates English and Afrikaans into Northern Sotho, Sesotho, Tswana, Xhosa and Zulu, and translates them back into Afrikaans and English (495 pages).
- [Dent and Nyembezi \(1969\)](#) contains 519 pages;
- [Roberts \(1895\)](#) contains Zulu translations of about 18,000 English words (267 pages). An [OCR version](#) of this dictionary is also available.
- [Roberts \(1895\)](#) contains 290 pages;
- [Dent \(1970\)](#) contains 150 pages;

Online dictionaries

There are a couple of online Zulu dictionaries as well:

- [Isizulu](#);
- [Isizulu2](#) (scrapeable);
- [MyMemory](#);
- [Zulu e-Dict](#) (first version of an integrated e-dictionary translating possessive constructions from English to Zulu, developed within the [SeLa Project](#), see below);
- [Zulu Wiktionary](#) (As of 09/01/2017, it contains 593 entries.)

22.5.3 Corpora

Monolingual

[Eiselen and Puttkammer \(2014\)](#) summarize the achievements of the NCHLT text project. This project aimed at developing multiple linguistic resources for ten of the official languages of South Africa, namely Zulu, Xhosa, Ndebele, Siswati, Setswana, Sesotho, Sepedi, Tswa-Ronga Xitsonga, Tshivenda and Afrikaans. The first step in this project was to collect unannotated, monolingual corpora for these ten languages. Due to the limited availability of electronic data for many of the languages, the corpus was aimed at collecting one million tokens per language. Most of the corpus data was sourced from South African government websites and documents, with some smaller sets of news articles, scientific

articles, magazine articles and prose. The [NCHLT isiZulu Text Corpus](#) contains 1,64 million tokens. The second step in the project was the annotation of a subset of data on four layers, i.e. the token, orthographic, morphological, and morphosyntactic layers. After the tokenization, each token was annotated with lemmatisation, part-of-speech and morphological analysis information. The [NCHLT isiZulu Annotated Text Corpus](#) comprises 44,324 tokens. The computational tools developed within this project are further discussed in Section 22.6.

As presented by [Spiegler et al. \(2010\)](#), there is a project for the developing of an open-source morphological Zulu corpus, i.e. the [Ukwabelana](#) corpus. This corpus contains about 100,000 common Zulu word types and 30,000 Zulu sentences compiled from fictional works and the Zulu Bible, from which the labeled words and sentences have been sampled. The Ukwabelana corpus contains 10,000 morphologically labeled words and 3,000 POS-tagged sentences.

Within the [SeLa Project](#), a morpho-syntactic database has been developed. This database contains an ontology of linguistic categories, containing linguistic units of South African Bantu languages (the beta version of the [Ontology database](#) is available online). It provides a list of morphemes labelled with morphosyntactic category. The goal of the database is to describe the morphemes of these languages in a single common database in order to outline and interpret commonalities and differences in more detail. Moreover, the relational database defines the underlying morphemic units (morphs) for these languages. It will be shown that the electronic part-of-speech ontology goes hand in hand with part-of-speech tagsets that label morphemic units. This database is designed as part of a forthcoming system providing lexicographic and linguistic knowledge on the official South African Bantu languages.

Additionally, Zulu is also represented in the [300 Languages Project](#), which is part of the Rosetta Project that aims at collecting materials in every variety of the 300 most widely-spoken languages and macrolanguages in the world.

As of 11/01/2017, [Zulu Wikipedia](#) has 885 content pages.

Bilingual

Some good candidates for building parallel corpora are listed below:

- the [Zulu](#) translation of the *Bible* is available online on the site of Yehovah's witnesses as well as on the homepage of the [Wordproject](#);
- the [Zulu](#) translation of the *Book of Mormon*;
- the *Quran* is also translated into [Zulu](#);
- the [Thai](#) translation of the *Universal Declaration of Human Rights* (available in .txt format on the web page of the [Unicode Consortium](#)).
- [Ubuntu](#) and [GNOME](#) localization files are also available aligned with different languages, including Zulu, in .xml format on the homepage of the OPUS corpus project.

Speech corpora

According to [Louw et al. \(2001\)](#), the *African Speech Technology project*'s goal is to develop telephone speech databases for five of South Africa's eleven official languages: South African English, Afrikaans, Zulu, Xhosa, and Southern Sotho (the [African Speech Technology isiZulu Speech Corpus](#) is available for download). These databases will be fully transcribed and will be used for the training and testing of phoneme-based, speaker-independent speech recognition systems.

Furthermore, the [NCHLT Speech Corpus](#) speech corpus contains wide-band speech from approximately 200 speakers per language, in each of the eleven official languages of South Africa, see [Barnard et al. \(2014\)](#) for more information about this corpus. The [NCHLT isiZulu Speech Corpus](#) contains 157,2 hours of Zulu speech collected from 210 informants. However, to date, only the NCHLT-clean subset of the corpora has been released. In the case of Zulu, the NCHLT-clean subset contains 56 hours and 14 minutes of speech (25,650 types and 130,866 tokens).

Additionally, [van der Westhuizen and Niesler \(2016\)](#) discuss a new English–Zulu code-switched speech corpus compiled from South African soap opera broadcasts. The corpus contains English–Zulu code-switched spontaneous speech, and each utterance is annotated orthographically and code-switching boundaries are delineated in time.

22.5.4 News portals

There are a couple of newspapers published in the Zulu language, such as [isoLezwe](#), *Ilanga* and *UmAfrika* (mainly available in Kwazulu-Natal province and in Johannesburg). [Eyethu](#) is the largest circulating black owned newspaper brand in KwaZulu-Natal, having started with the launch of Edendale Eyethu in 2008, the brand now publishes combined 301,400 copies in 11 titles weekly. A complete list of the [newspapers in South Africa](#) is also available.

The [Ukhozi FM](#) is a South African national radio station that offers radio broadcasts for the Zulu-speaking community.

22.6 Computational tools

This section summarizes the available computational tools developed for the Zulu language. There is a comprehensive description of the African languages from a computational point of view, see [Osborn \(2010\)](#).

22.6.1 Language identification

Zulu is supported by the [Polyglot3000](#), [LabsTranslated](#) and [saffsd/langid](#) tools.

22.6.2 Tokenizer

[MorphAdorner 2.0](#), a Java implementation, is also compatible with Zulu. The tools developed by [MarkLogic](#) provide basic language support for Zulu (basic language support enables basic full-text search and includes tokenization using whitespace-delimiters and punctuation).

22.6.3 Stemmer

Within the NCHLT Text project, a [NCHLT isiZulu Lemmatiser](#) and a [NCHLT isiZulu Morphological Decomposer](#) have been developed (see [Eiselen and Puttkammer 2014](#)). Furthermore, [Cotterell et al. \(2015\)](#) present the main properties of the [ChipMunk](#) morphological segmenter and stemmer, developed for six languages: English, Finnish, German, Indonesian, Turkish and Zulu.

22.6.4 Spell checker

There are a couple of open source spell checkers available for Zulu, such as [Stars21](#), [Online Spell Checker for Zulu](#), [HunSpell](#), and [Aspell](#). Additionally, the Mozilla dictionary includes a Zulu [spell](#)

checker as well.

22.6.5 Phrase level and higher tools

This section presents a collection of Zulu trained tools and, in the absence of such tools, scientific articles dealing with the relevant tools:

- **part-of-speech tagger.** There is a demo version of a POS-tagger developed for Zulu. POS taggers have been developed within two corpus projects: Ukwabelana and NCHLT (see Spiegler et al. 2010 and Eiselen and Puttkammer 2014, respectively). The interested reader may also wish to consult De Pauw et al. (2012) who discuss POS-tagging for four Bantu languages: Swahili, Northern Sotho, Zulu and Cilubà as well as Koleva (2013) for Zulu.
- **named entity recognizer.** Louis et al. (2006) discuss Named entity recognition in South African context. Despite the fact that their research is based on English texts, the NER system should recognize a wide range of non-English NE's, i.e. the ones coming from the eleven official languages of South Africa.
- **chunker.** As of 11/01/2017, no chunker is available for Zulu.
- **morphological analyzer.** As part of the Ukwabelana corpus project, a semi-automatic morphological analyzer will be provided (Spiegler et al., 2010). Furthermore, within the NCHLT text project, a morphological analyzer has been developed (Eiselen and Puttkammer, 2014). Additionally, Pretorius and Bosch (2003, 2009) discuss the *ZulMorph* morphological analyzer developed for Zulu. There are further studies on morphological analysis of Zulu, see Baumann and Pierrehumbert (2014), Spiegler et al. (2010), and Pretorius and Bosch (2003, 2009); on morphosyntactic databases see Faab et al. (2012) and Taljard et al. (2015).
- **sentence parser.** Chris and Smith (2014) present the GFL-Web tool, a web-based interface for syntactic dependency annotation with the lightweight FUDG/GFL formalism. According to this study, guidelines for Swahili, Zulu, and Mandarin are under development.
- **question answering system.** As of 11/01/2017, not available for Zulu.
- **speech recognizer.** The Speech API of Google Cloud Platform supports Zulu. Additionally, a general-purpose speech synthesizer has been developed for Zulu (Louw et al., 2005). The Lwazi Project's aim was to develop and apply several speech technologies for the official languages of South Africa. Within this project, freely-available speech resources, such as corpora for automatic speech recognition (ASR) and text-to-speech (TTS) as well as other linguistic and technology resources (e.g. pronunciation dictionaries, phoneme sets, etc.) have been developed for all official South African languages. All Zulu tools are available for download from the project's homepage.
- **machine translation.** Google Translator supports Zulu. Additionally, Wolff and Kotzé (2014) deal with machine translation for Zulu.

22.6.6 End-user support

Operation systems. A Windows language pack is available for Zulu. Ubuntu also supports Zulu.

Among the **browsers**, Zulu is supported by Google Chrome, [Mozilla Firefox](#), [Internet Explorer](#), and [Opera](#).

[Microsof Office 2013 proofing tools](#) are available for Zulu.

Additionally, the optical character recognition system [ABBYY FineReader Engine 10](#) recognizes Zulu (without dictionary support).

Bibliography

Reader's Digest Association South Africa. *South African multi-language dictionary and phrase book: English, Afrikaans, Northern Sotho, Sesotho, Tswana, Xhosa, Zulu*. Reader's Digest Association South Africa, 1991. URL <https://books.google.hu/books?id=yr00AAAAYAAJ>.

Etienne Barnard, Marelle H Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhorst. The NCHLT speech corpus of the South African languages. In *Proc. SLTU*, pages 194–200, 2014. URL <https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxuY2hsdHNwZWVjaGNvcnB1c3xneDo0YzgxNzhjOTIyMjQ5NDRI>.

Peter Baumann and Janet B Pierrehumbert. Using resource-rich languages to improve morphological analysis of under-resourced languages. In *LREC*, pages 3355–3359, 2014. URL http://faculty.wcas.northwestern.edu/~jbp/publications/Baumann_PierrehumbertLREC2014.pdf.

Alfred T. Bryant. *A Zulu-English Dictionary with Notes on Pronunciation: A Revised Orthography and Derivations and Cognate Words from Many Languages; Including Also a Vocabulary of Hlonipa Words, Tribal-names, Etc., a Synopsis of Zulu Grammar and a Concise History of the Zulu People from the Most Ancient Times*. The Mariannhill Mission Press, 1905.

Leston Chandler Buell. *Issues in Zulu verbal morphosyntax*. PhD thesis, University of California, 2005. URL <http://www.linguistics.ucla.edu/general/Dissertations/BuellDissertationUCLA2005.pdf>.

Noverino N Canonici. *Zulu grammatical structure*. Zulu Language and Literature, Univ. of Natal, 1996.

Michael T Mordowanec Nathan Schneider Chris and Dyer Noah A Smith. Simplified dependency annotations with GFL-Web. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 121–126, 2014. URL <http://www.aclweb.org/anthology/P14-5021>.

Toni Cook. *Morphological and phonological structure in Zulu reduplication*. PhD thesis, University of Pennsylvania, 2013. URL <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1903&context=edissertations>.

Anthony Trevor Cope. *Zulu: A comprehensive course in the Zulu language*. Department of Zulu Language and Literature, University of Natal, 1984.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. Labeled morphological segmentation with semi-Markov models. In *CoNLL 2015*, pages 164–174, 2015. URL <http://www.aclweb.org/anthology/K15-1017>.

Guy De Pauw, Gilles-Maurice De Schryver, and Janneke van de Loo. Resource-light Bantu part-of-speech tagging. In *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8-AFLAT 2012)*, pages 85–92. European Language Resources Association, 2012. URL <http://tshwanedje.com/publications/BantuPOS.pdf>.

Gilles-Maurice De Schryver. *Oxford Bilingual School Dictionary: isiZulu and English/Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi, Esishicilelwe abakwa-Oxford*. Oxford University Press Southern Africa, 2015. URL <https://www.oxford.co.za/book/9780199079544-oxford-bilingual-school-dictionary-isizulu-english-2#.WHVpRXOhEno>.

George Robinson Dent. *Compact Zulu Dictionary: English-Zulu, Zulu-English*. Shuter & Shooter, 3 edition, 1970.

G.R. Dent and C.L.S. Nyembezi. *Scholar's Zulu Dictionary; English-Zulu, Zulu-English*. Shuter and Shooter, 1969. ISBN 9780869850237. URL <https://books.google.hu/books?id=UPINAAAAYAAJ>.

Clement Martyn Doke. *Textbook of Zulu grammar*. Longmans Southern Africa, 1963.

Clement Martyn Doke, D McK Malcolm, and Jonathan MA Sikakana. *English-Zulu Dictionary*. Witwatersrand University Press, 2. rev. edition, 1958.

CM Doke and BW Vilakazi. *Zulu-English dictionary*. Witwatersrand University Press, 2nd edition, 1972. ISBN 0854940278, 9780854940271.

Laura J Downing. How ambiguity of analysis motivates stem tone change in Durban Zulu. *UBCWPL*, 4:39–55, 2001.

Matthew S. Dryer. Order of genitive and noun. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013a. URL <http://wals.info/chapter/86>.

Matthew S. Dryer. Position of interrogative phrases in content questions. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL <http://wals.info/chapter/93>.

Matthew S. Dryer. Polar questions. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013c. URL <http://wals.info/chapter/116>.

Roald Eiselen and Martin J Puttkammer. Developing text resources for ten South African languages. In *LREC*, pages 3698–3703, 2014. URL <https://pdfs.semanticscholar.org/1a77/15e8d41a228f8fc35f7e3f2580988eea9b24.pdf>.

Gertrud Faab, Sonja Bosch, and Elsabe Elizabeth Taljard. Towards a part-of-speech ontology: encoding morphemic units of two South African Bantu languages. *Nordic Journal of African Studies*, 21(3):118–140, 2012. URL http://www.njas.helsinki.fi/pdf-files/vol21num3/faas_bosch_taljard.pdf.

K.B. Hartshorne, J.H.A. Swart, and E. Posselt. *Dictionary of Basic English-Zulu*. Across the curriculum. Educum Publishers, 1985. ISBN 9780798009577. URL <https://books.google.hu/books?id=b1gIAQAAIAAJ>.

Robert K. Herbert and Richard Bailey. *Language in South Africa*, chapter The Bantu languages: sociohistorical perspectives, pages 50–78. Cambridge University Press, 2002.

Shirley Illman. *Illman's English / Zulu Dictionary and Phrase Book*. AuthorHouse, 2014. URL <https://books.google.hu/books?id=cTMkBQAAQBAJ>.

Mariya Koleva. Towards adaptation of NLP tools for closely-related Bantu languages: Building a part-of-speech tagger for Zulu. Master's thesis, Master's thesis, Saarland University, Germany, 2013.

Isaac Sibusiso Kubeka. *A preliminary survey of Zulu dialects in Natal and Zululand*. PhD thesis, University of Natal, Durban, 1979. URL <http://researchspace.ukzn.ac.za/handle/10413/5117>.

Anita Louis, Alta De Waal, and Cobus Venter. Named entity recognition in a South African context. In *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, pages 170–179. South African Institute for Computer Scientists and Information Technologists, 2006. URL https://www.researchgate.net/publication/30509720_Named_entity_recognition_in_a_South_African_context.

JA Louw, M Davel, and E Barnard. A general-purpose isiZulu speech synthesizer. *South African journal of African languages*, 25(2):92–100, 2005. URL <http://citeseervx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.3723&rep=rep1&type=pdf>.

Philippa H Louw, Justus C Roux, and Elizabeth C Botha. African speech technology (AST) telephone speech databases: corpus design and contents. In *INTERSPEECH*, pages 2055–2058, 2001. URL <http://perso.telecom-paristech.fr/~chollet/Biblio/Congres/Audio/Eurospeech01/CDROM/papers/page2055.pdf>.

Ian Maddieson. Tone. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013a. URL <http://wals.info/chapter/13>.

Ian Maddieson. Presence of uncommon consonants. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL <http://wals.info/chapter/19>.

Ian Maddieson. Consonant inventories. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013c. URL <http://wals.info/chapter/1>.

Ian Maddieson. Vowel quality inventories. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013d. URL <http://wals.info/chapter/2>.

Constance Samukelisiwe Magagula. Standard versus non-standard isiZulu : a comparative study between urban and rural learners' performance and attitude. Master's thesis, University of KwaZulu-Natal, 2009. URL <http://researchspace.ukzn.ac.za/xmlui/bitstream/handle/10413/273/Magagula/%2c/%20Constance/%20Samukelisiwe/%20thesis/%202009.pdf?sequence=1&isAllowed=y>.

- Jouni Maho. The online version of the New Updated Guthrie List, a referential classification of the Bantu languages. Unpublished manuscript, 2009. URL <http://goto.glocalnet.net/mahopapers/nuglonline.pdf>.
- MO Mbatha. *Isichazamazwi sesiZulu*. New Dawn Publishers, Pietermaritzburg, 2006.
- Cyril Lincoln Sibusiso Nyembezi. *Learn Zulu*. Shuter and Shooter, 1957.
- Joseph O’Neil. *A grammar of the Sindebele dialect of Zulu: with numerous examples and a key to the exercises*. E. Allen, 1912.
- Geōrgios Poulos and Sonja Elva Bosch. *Zulu*, volume 50. Lincom Europa, 1997.
- George Poulus and Christian T Msimang. *A linguistic analysis of Zulu*. Via Africa/Collegium Educational Publishers, 1998.
- Laurette Pretorius and Sonja Bosch. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 96–103. Association for Computational Linguistics, 2009. URL <http://www.aclweb.org/anthology/W09-0714>.
- Laurette Pretorius and Sonja E Bosch. Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation*, 18(3):195–216, 2003.
- Charles Roberts. *An English-Zulu dictionary: with the principles of pronunciation and classification fully explained*. Paul, Trench, Trübner, 1895.
- Philemon Buti Skhosana. *The linguistic relationship between Southern and Northern Ndebele*. PhD thesis, University of Pretoria, 2009. URL <http://repository.up.ac.za/bitstream/handle/2263/28563/Complete.pdf?sequence=8>.
- Sebastian Spiegler, Andrew Van Der Spuy, and Peter A Flach. Ukwabelana: An open-source morphological Zulu corpus. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1020–1028. Association for Computational Linguistics, 2010. URL <http://www.cs.bris.ac.uk/Publications/Papers/2001224.pdf>.
- Leon Stassen. Zero copula for predicate nominals. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL <http://wals.info/chapter/120>.
- Elsabé Taljard, Gertrud Faaß, and Sonja Bosch. Implementation of a part-of-speech ontology: Morphemic units of Bantu languages. *Nordic Journal of African Studies*, 24(2):146–168, 2015.
- Johan van der Auwera, Ludo Lejeune, and Valentin Goussov. The prohibitive. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL <http://wals.info/chapter/71>.
- Ewald van der Westhuizen and Thomas Niesler. Automatic speech recognition of English-isiZulu code-switched speech from South African soap operas. *Procedia Computer Science*, 81:121–127, 2016. URL <http://www.sciencedirect.com/science/article/pii/S1877050916300539>.

Friedel Wolff and Gideon Kotzé. Experiments with syllable-based Zulu-English machine translation. In *Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium*, pages 217–222, 2014. URL <http://www.prasa.org/proceedings/2014/prasa2014-38.pdf>.

Dirk Zier vogel, Jacobus Abraham Louw, and J Ngidi. *A handbook of the Zulu language*. van Schaik, 1967.

