ISMLA Multilingual Session 1: Important Standard Languages of the World

Johannes Dellert

Tübingen University

October 25, 2017

Plan

Introduction

Role of languages in our course:

- data from different major languages is used in the exercises
- also, we expect your term projects to deal at least partly with a language other than German or English or your native language

Dual goal of the course:

- familiarize yourself with current technology
- get a more comprehensive understanding of how diverse language data can be, but also an impression of how widely applicable simple solutions are

Why multilingual?

Reasons to venture beyond your native language and English:

- knowing how the world's major languages work helps to know beforehand whether you can expect your idea or your system to work for other languages
- some typical problems of NLP appear more clearly in some languages than in others, picking the most challenging language for a task will make your solution more generally applicable
- working with an unfamiliar language makes it easier to intuit how the computer sees the data (see Chinese Room Example)

Languages: The Natural State

For hundred of thousands of years, the linguistic landscape looked like this:

- about 1,000 speakers per language on average, i.e. every village evolves to speak its own language or dialect
  (still the case in some areas of the world)
- people know different languages for different purposes:
  - the language of village they grew up in
  - the language of the spouse (from another village)
  - the language of the village they live in
  - the language of the nearest marketplace
  - a supraregional lingua franca for communicating with strangers

$\Rightarrow$ Multilingualism is the norm, modern monolingualism is abnormal!

Languages in the Age of Standardization

Standardized written languages changed all that:

- every modern state uses some standard language(s) which are used in the media and for written communication
- nowadays, the vast majority of all people learns the standard language of their state at school, and the local languages are never written
- good result: critical mass for development of literature and science
- bad result: minority languages and local culture disappear under pressure of maximizing economic opportunities

The Engineer's Perspective

From the engineer's perspective, this means:

- computational linguistics as we know it would be almost impossible in the natural state (and we can be glad that standardization exists)
- the written standard languages form islands of predictability that you can build systems on (although we never should underestimate the potential of surprises)
- it is always worthwhile to keep in mind that the reality of what people actually speak is virtually intractable

Language Codes

Languages have received standardized codes as identifiers which are used almost universally:

- **ISO 639-1** defines two-letter codes for 209 written languages
- **ISO 639-3** by now distinguishes more than 6.900 “languages” using three-letter codes

Problems:

- for non-standardized languages, clear-cut distinctions are a fiction
- some important written languages have no ISO 639-1 codes (Cantonese, Tok Pisin)
- insufficient treatment of historical variants (e.g. different versions of Latin)

Ranking Languages by Importance

Possible criteria for deciding which standard languages are worth keeping in mind for multilingual applications:

- **number of speakers**? (unbiased and neutral criterion)
  - Punjabi (95M) more relevant than German (92M)?
  - Telugu (76M) more relevant than French (75M)?
- **Wikipedia size**? (as a proxy for presence on the internet)
  - Dutch (1.9M) twice as important as Chinese (960K)?
  - Sinugboanong Binisaya as important as English (5.4M)?
- my proposal: rank languages by **combined GDP of speakers**
  - distribute the GDP of each country according to ratios of speakers of the official language (going down to the state level in India)
  - GDP is a good proxy for commercial relevance on the world market
  - difficulty in treatment of regional official languages

#1: English (en, eng): Statistics

- GDP: $24.038 trillion (32.6% of world GDP)
- L1 speakers: 371M, L2 speakers: 611M
- coverage so far: 5.1% of world population, 32.6% of world GDP

There is debate about when Christianity was first introduced; it was no later than the 4th century, probably much earlier. According to Bede, missionaries were sent from Rome by Eleutherius at the request of the chieftain Lucius of Britain in 180 AD, to settle differences as to Eastern and Western ceremonials, which were disturbing the church. There are traditions linked to Glastonbury claiming an introduction through Joseph of Arimathea, while others claim through Lucius of Britain.

#1: English (en, eng): Interesting Features

- almost isolating morphology:
  - forms of the verb: 5
  - forms of the noun: 4 (only in written form)
- by far the largest lexicon of all languages
  - language of science produces new words all the time
  - colonial past has led to loanwords from across the globe
- mixed ancestry leads to peculiar structure of the lexicon:
  - *dog* $\rightarrow$ *canine*
  - *cat* $\rightarrow$ *feline*
  - *cow* $\rightarrow$?
  - *donkey* $\rightarrow$?

#1: English (en, eng): Challenges

- predicting pronunciation
- POS tagging
- usage by non-natives

#2: Chinese (zh, cmn): Statistics

- GDP: $11.608 trillion (15.7% of world GDP)
- L1 speakers: 897M, L2 speakers: 193M
- coverage so far: 17.4% of world population, 48.3% of world GDP

传说黄帝原系炎帝部落的一个分支的领袖，强大之后在阪泉之战中击败炎帝，成为新部落联盟首领，之后又与东南方的蚩尤部落发生冲突，在涿鹿之战中彻底击败对手，树立自己的霸主地位。后来黄帝的孙子颛顼和玄孙帝喾继续担任部落联盟的首领。帝喾的儿子尧继位，创立禅让制，传位给舜。在舜时期，黄河洪水泛滥，鲧采用堵塞的方法，结果洪水更厉害了，鲧被处决，他的儿子禹采用疏导的方法治水成功，因此受舜禅让继帝位。

#2: Chinese (zh, cmn): Interesting Features

- unique writing system unifying a country across language boundaries
  - in Classical Chinese, one syllable = one word = one symbol
  - in modern Mandarin, many words are disyllabic
- traditional language of science and culture for all of East Asia
  - much like Latin in Europe, Arabic in the Middle East, Sanskrit in India
- by far the most important tonal language (only about 430 possible syllables, tones bring that number to almost 1,300)

#2: Chinese (zh, cmn): Challenges

- tokenization
  - no spaces between words
  - many symbols are plurifunctional
- two competing script standards representing a political divide
  - traditional characters in Taiwan and Hongkong
  - simplified characters in mainland China
- detection of Classical Chinese
  - educated Mandarin is interspersed with quotes in Classical Chinese
  - these have a distinct syntax and a condensed style
- problems in representing loanwords and foreign names
  - there are no meaningless purely phonetic signs, i.e. most transcriptions will be ambiguous with a literal interpretation
  - transcriptions must stay within very limited phonetic possibilities:

#3: Spanish (es, spa): Statistics

- GDP: $5.531 trillion (7.5% of world GDP)
- L1 speakers: 436M, L2 speakers: 91M
- coverage so far: 23.3% of world population, 55.8% of world GDP

La particular posición de la península ibérica como «Extremo Occidente» del mundo mediterráneo determinó la llegada de sucesivas influencias culturales del Mediterráneo oriental, particularmente las vinculadas al Neolítico y la Edad de los Metales, proceso que culminó en las denominadas colonizaciones históricas del I milenio a. C. Tanto por su localización favorable para las comunicaciones como por sus posibilidades agrícolas y su riqueza minera, las zonas este y sur fueron las que alcanzaron un mayor desarrollo.

#3: Spanish (es, spa): Interesting Features

- largest language with complex verbal morphology
  (about 70 forms, not including object clitics)
- most complex tense system so far (more complex than e.g. French)
- only language where question and exclamation marks are brackets:
  ¡No exageres!
  Genial, ¿no?

#3: Spanish (es, spa): Challenges

- requires some morphological analysis, a few dozen irregular paradigms
- **pro-drop** language: verbal person marking is enough, automated methods have to analyse the verb form to infer the subject
- a variety of competing standards due to pluricentrism, many lexical differences which need to be bridged by a pan-Hispanic search engine:

	Spain	Mexico	Argentina
"jacket"	chaqueta	chamarra	campera
"glasses"	gafas	lentes	anteojos
"apricot"	albaricoque	chabacano	damasco
"cake"	tarta	pastel	torta

#4: German (de, deu): Statistics

- GDP: $4.209 trillion (5.7% of world GDP)
- L1 speakers: 76M, L2 speakers: 52M
- coverage so far: 24.4% of world population, 61.6% of world GDP

Die historisch erfassten germanischen Stämme der frühen römischen Kaiserzeit des ersten Jahrhunderts gliedern sich in drei Kulturgruppen auf: die sogenannten Rhein-Weser-Germanen, die Nordseegermanen und die Elbgermanen. Durch die makropolitischen Einflüsse des andauernden Konflikts mit dem Römischen Reich sowie innergermanische politische, soziale und wirtschaftliche Veränderungen kam es ab dem 2. Jahrhundert aus diesen Kulturgruppen heraus zur (nicht biologisch, sondern als historisch-sozialer Prozess verstandenen) Entstehung von neuen und größeren Stammesverbänden.

#4: German (de, deu): Interesting Features

- second-largest Germanic language
- one of the first non-Romance vernaculars of Europe to appear in writing, important for the creation of literary languages of Northern and Eastern Europe
- only major language to systematically reflect a word class distinction in orthography

#4: German (de, deu): Challenges

- complexity of syntax unique among major languages:
  - basic word order depends on sentence type
  - particle verbs induce very long-distance dependencies
  - added flexibility due to sentence-initial topic
- good lexical coverage presupposes treatment of compounding:
  Knusperwürfel, *Blutbiene*, *Ausweichkindergarten*, *Wildkatzendatenbank*, *Zimtcrumble*, *Einmessfunktion*, *Lidrandmassage*, *Kryptovaluta*
  (all observed first on October 19th by `wortwarte.de`)

#5: Japanese (ja, jpn): Statistics

- GDP: $4.209 trillion (5.7% of world GDP)
- L1 speakers: 128M, L2 speakers: 1M
- coverage so far: 26.1% of world population, 67.2% of world GDP

紀元前 8 世紀頃以降、中国南部から稲作を中心とする文化様式を持つ弥生人が流入すると、各地に「クニ」と呼ばれる地域的政治集団が徐々に形成される。これらの地域的政治集団により、朝鮮半島南部から南西諸島までの範囲で海上交易で結びついた緩やかな倭人の文化圏が構成されていった。こうした文化圏の中で、勾玉などが紀元前 6 世紀以降日本から朝鮮半島へ伝搬したほか、紀元前 2 世紀頃に青銅器および鉄器の製造法が日本へ伝わった。1 世紀・2 世紀前後に各クニが抗争を繰り返し、各地に地域的連合国家を形成した。

#5: Japanese (ja, jpn): Interesting Features

- mixed writing system:
  - **Kanji** (Chinese characters) for most content words
  - **Hiragana** (a syllabary) for suffixes and function words
  - **Katakana** (a syllabary) for loans and foreign names
- unique system of particles e.g. for topic marking
- most important language with grammaticalized politeness distinctions

#5: Japanese (ja, jpn): Challenges

- Kanji pronunciation is extremely context-dependent:
  女 *onna* “woman”
  女神 *megami* “goddess”
  女権 *joken* “woman rights”
  女房 *nyōbō* “one’s wife”
- person and number typically need to be inferred from context:

  終わったのですか。
  owatta no desu ka.
  finish-PST PRT be Q
  “Have you finished?”

  終わった。
  owatta.
  “I am done.” / “It is over.”

#6: French (fr, fra): Statistics

- GDP: $3.589 trillion (4.9% of world GDP)
- L1 speakers: 76M, L2 speakers: 153M
- coverage so far: 27.2% of world population, 72.0% of world GDP

Au cours du IIIe siècle, l'Empire romain connaît une période de grave crise appelée l'Anarchie militaire. Aux raids barbares parfois à conséquences durables (pillage et accaparement de richesses transportables, prise d'otages ou d'esclaves) s'ajoutent une crise politique et économique qui se traduit par une dévaluation importante de la monnaie (à valeur beaucoup plus fiduciaire que réelle, comme les bronzes), une grande instabilité politique doublée de guerres civiles et de généralisation de bagaudes encore plus ravageuses que les incursions étrangères.

#6: French (fr, fra): Interesting Features

- the world's richest basic vowel inventory
- unlike for the other colonial languages, the official language is still very much dominated by the country of origin
- was the last lingua franca before English, leaving loanwords across Europe
- very marked phonological profile

#6: French (fr, fra): Challenges

- phoneme-grapheme correspondence regular, but very complex
- fast-changing slang a lot more widespread than in other languages:
  dodo “sleep”, *boulot* “work”, *fric* “money”
  are just as frequent in colloquial style as the official
  dormir, *travail*, *argent*

#7: Arabic (ar, ara): Statistics

- GDP: $2.379 trillion (3.2% of world GDP)
- L1 speakers: 290M, L2 speakers: 132M
- coverage so far: 31.1% of world population, 75.2% of world GDP

حسب الضئيل الذي عثر عليه بظاهر الأرض، كانت الممالك أو الإتحادات القبلية الأربع الرئيسية تتاجر بالبخور والبهارات والمر والذهب ويذكر العهد القديم قصة ملكة سبأ وزيارتها للملك سليمان وقدومها بقوافل محملة بالطيب والذهب وتعتبر مأرب مهد الحضارة اليمنية القديمة. بدأ السبئيون بالتوسع شيئا فشيئا والإستيلاء على الإمارات الصغيرة التابعة للممالك الأخرى قرابة القرن التاسع قبل الميلاد. لا توجد أنهار في اليمن كتلك الموجودة بمصر والعراق وطبيعة الأرض جبلية وعرة فظهرت ممالك متعددة، متحاربة ومتصارعة للسيطرة على الموارد المحدودة أشهرها وأقواها كانت مملكة سبأ.

#7: Arabic (ar, ara): Interesting Features

- unique among major languages: **root-and-pattern morphology**
  DRS → DaRaSa "he taught", maDRaSa "school", DaRS "lesson"
  RKB → RaKiBa "he drove', maRKaBa "verhicle", RaKB "platoon"
  MLK → MaLaKa "he ruled", maMLaKa "kingdom", MuLK "reign"
- plural forms are very unpredictable, must be modeled explicitly:
  dars "lesson" → *durūs* "lessons"
  milk "possession" → *amlāk* "possessions"
  ra'is "president" → *ru'asā'* "presidents"
- a major source of technical and religious terminology in all Islamic and many adjacent cultures (often mediated via Persian)

#7: Arabic (ar, ara): Challenges

- transcription: script does not represent short vowels
  (root always clear, pattern needs to be inferred from context)
  The word ملك can read:
  malik "king", *mulk* "reign", *milk* "possessions", *malaka* "he ruled",
  mallaka "he made king", *malak* "angel", *mullak* "owners"
- differences between Modern Standard Arabic (MSA)
  and Qur'anic style for embedded erudite phrasings
- prime example of **diglossia**:
  - spoken Arabic is not a single language, but a family
  - there are no native speakers of Standard Arabic

#8: Portuguese (pt, por): Statistics

- GDP: $2.100 trillion (2.8% of world GDP)
- L1 speakers: 218M, L2 speakers: 11M
- coverage so far: 34.1% of world population, 78.1% of world GDP

A economia Ibérica tinha uma agricultura rica, forte exploração mineira e uma metalurgia desenvolvida. A língua Ibérica, uma língua não Indo-europeia continuou a ser falada durante a ocupação romana. Ao longo da costa Este, utilizou-se uma escrita Ibérica, um sistema de 28 sílabas e caracteres alfabéticos, alguns derivados dos sistemas fenício e grego, mas de origem desconhecida. Ainda sobrevivem muitas inscrições dessa escrita paleohispânica, mas poucas palavras são compreendidas, excepto alguns nomes de locais e cidades do século III, encontradas em moedas.

#8: Portuguese (pt, por): Interesting Features

- very similar to Spanish to the point of mutual intelligibility between the written languages
- the least Europe-centered of all former colonial languages

#8: Portuguese (pt, por): Challenges

- pronunciation not trivial to predict from orthography
- two extremely divergent pronunciation standards:
  - European tends to delete unstressed vowels:
    cidade [sðað], *real* [ʁjaɬ]
  - Brazilian with some extreme sound shifts:
    cidade [si'daʤi], *real* [he'aw]

#9: Italian (it, ita): Statistics

- GDP: $1.873 trillion (2.5% of world GDP)
- L1 speakers: 63M, L2 speakers: 3M
- coverage so far: 35.0% of world population, 80.6% of world GDP

I sepolcreti, infatti, testimoniano la presenza di un antico stanziamento. Isolati, se mai, sembrano rimanere gli ambiti dell'Etruria interna, nelle regioni più inospitali, mentre i villaggi in vicinanza del mare o di vie di comunicazione fluviale si rivelano molto attive. Le principali città costiere sorgono a pochi chilometri dalla costa, l'unica città stato etrusca sul mare è stata probabilmente Populonia, mentre le altre città costiere sembrano di solito dotate di insediamenti marittimi come Regisvilla per Vulci, l'insediamento etrusco presso la colonia romana di Gravisca.

#9: Italian (it, ita): Interesting Features

- standardized much later than the other three major Romance languages (due to later unification)
- closely related local variants persist, spoken Italian can deviate very far from the standard

#9: Italian (it, ita): Challenges

- due to different rules of orthography, some analysis steps require more work than in Spanish:
  lunghe: *lunga* + *-e*
  guance : *guancia* + *-e*
- complex interaction of prepositions and articles:
  del, *della*, *delle*, *dei*,
  al, *alla*, *alle*, *ai*,
  nel, *nella*, *nelle*, *nei*
- wealth of dialects reflected by non-standard forms in private messages

#10: Russian (ru, rus): Statistics

- GDP: $1.456 trillion (2.0% of world GDP)
- L1 speakers: 153M, L2 speakers: 113M
- coverage so far: 37.1% of world population, 82.6% of world GDP

До V века славяне жили при родовом строе. Во главе каждой родовой общины стоял родовой старейшина, обладающий неограниченной властью. Земля являлась собственностью общины, часть сельскохозяйственных работ осуществлялась коллективно. С V века началось разложение родового строя, родовая община начала заменяться территориальной общиной (вервью), управление общиной наряду со старейшинами стало осуществлять вече —общее собрание членов общины.

#10: Russian (ru, rus): Interesting Features

- by far the most relevant language with a complex case system (six cases, i.e. more complex than Latin)
- deviates from other Slavic and IE languages in some syntactic features, probably due to Uralic substrate influence:
  - null copula in present tense
  - "to have" not expressed by a verb

#10: Russian (ru, rus): Challenges

- some important semantic distinctions are expressed only through case, a problem for statistical machine translation
- very fine-grained lexicalized distinctions in verbal aspect make accurate translation difficult
- unpredictable accent with huge effects on pronunciation; generating correct pronunciation requires accent marks, but these are not written

#11: Korean (ko, kor): Statistics

- GDP: $1.395 trillion (1.9% of world GDP)
- L1 speakers: 77M, L2 speakers: 1M
- coverage so far: 38.1% of world population, 84.5% of world GDP

붕당 정치가 변질되고 그 폐단이 심화되면서, 노론과 남인 위주의 일당전제화 경향이 두드러졌다. 조선 후기 사회에서 서민은 점차 경제적 변화에 적극적으로 대응하고 생산력도 증가하였다. 이후 조선에서는 영조, 정조 시대에 다시 중흥하였다. 이때 실학이 융성하였고, 천주교가 일부 남인에 의해 학문의 일부로서 전래되었다. 그리고 양명학이 전래되었으며 천문학과 의학, 농업과 상업 분야에서의 기술적 성과가 산업 발전을 촉진하였다. 한편으로는 양반층이 증가하고 농민의 분화가 이루어지는 등 반상제의 신분제가 동요하였다. 세도 정치 시기에는 삼정의 문란으로 민란이 일어나기도 하였다.

#11: Korean (ko, kor): Interesting Features

- structurally very similar to Japanese, but not provably related
- Hangul, the only alphabetic script based on first principles:
  ㅡ [ɯ], ㅜ [u], ㅠ [ju], ㅗ [o], ㅛ [jo]
  ㅣ [i], ㅏ [a], ㅑ [ja], ㅓ [ʌ], ㅕ [jʌ]
- use of space between words makes tokenization much easier than in other East Asian languages
- Chinese characters completely abolished in North Korea, of decreasing importance in South Korea

#11: Korean (ko, kor): Challenges

- due to East Asian block arrangement, every possible syllable has its own codepoint in Unicode (almost 12,000 codepoints)
- this makes morphological rules with effects across syllable boundaries very difficult to model
- just as many homophones as Japanese, but they are also homographs in Korean (disambiguation difficult):
  간 *gan* is written for a range of hanja (Chinese characters):
  間 as in 간극 (間隙) “gap”
  肝 as in 간장 (肝臟) “liver”
  簡 as in 간결 (簡潔) “conciseness”
  奸 as in 간교 (奸巧) “cunningness”
  看 as in 간과 (看過) “overlooking”

#12: Indonesian (id, ind): Statistics

- GDP: $1.175 trillion (1.6% of world GDP)
- L1 speakers: 77M, L2 speakers: 204M
- coverage so far: 39.2% of world population, 86.1% of world GDP

Kondisi pertanian yang ideal memungkinkan upaya bercocok tanam padi lahan basah mulai berkembang sekitar abad ke-8 SM. memungkinkan desa dan kota kecil mulai berkembang pada abad pertama Masehi. Kerajaan ini yang lebih mirip kumpulan kampung yang tunduk kepada seorang kepala suku, berkembang dengan kesatuan suku bangsa dan sistem kepercayaan mereka. Iklim tropis Jawa dengan curah hujan yang cukup banyak dan tanah vulkanik memungkinkan pertanian padi sawah berkembang subur.

#12: Indonesian (id, ind): Interesting Features

- in most typological variables near the average, arguably the major Asian language that is easiest to learn (not only for Europeans)
- three times more L2 than L1 speakers (most extreme ratio among major languages)
- phonology and morphology very streamlined, as can be expected for an ancient lingua franca

#12: Indonesian (id, ind): Challenges

- two competing national standards:
  - Malaysia: more native speakers, loanwords from English
  - Indonesia: more speakers overall, loanwords from Dutch
- lexical influence of local languages (Javanese, Balinese, etc.) in common every-day usage
- standardization is not very strong due to a suboptimal school system

#13: Dutch (nl, nld): Statistics

- GDP: $1.008 trillion (1.4% of world GDP)
- L1 speakers: 24M, L2 speakers: 4M
- coverage so far: 39.5% of world population, 87.5% of world GDP

Door de introductie van nieuwe technologieën als metaalbewerking kon voedsel efficiënter vergaard worden, wat een elite vrij maakte die zich bezighield met andere zaken. De belangrijkste specialisatie was die van spirituele leiders die gevaren konden duiden of zelfs afwenden. Rondom hen ontstonden centra van toenemende welvaart, aanvankelijk om de goden tevreden te stellen. Ter bescherming werden tijdelijk krijgers als leider aangesteld die echter gaandeweg meer macht verkregen. Relatief egalitaire samenlevingen werden op die manier adellijk met een aristocratisch bestuur. Dit gold onder meer voor de Kelten voor wie het latere Nederland een randgebied was en de La Tène-cultuur daarna.

#13: Dutch (nl, nld): Interesting Features

- arose from a different point in the same dialect continuum as Standard German, still mutually intelligible with Low German
- has lots of amusing false friends with German:
  - *durven* "to dare" vs. *mogen* "to be allowed"
  - *kachel* "stove" vs. *tegel* "tile"
  - *malen* "to grind" vs. *schilderen* "to paint"
  - *schattig* "cute" vs. *schaduwrijk* "shadowy"
  - *verzoeken* "to request" vs. *proberen* "to try"
  - *vuilnis* "trash" vs. *rotting* "decay"

#13: Dutch (nl, nld): Challenges

- lemmatization requires phonetic analysis, cutting of endings will not be enough: *leven* "to live", *ik leef* "I live"
- hard for automated methods to tell apart from its descendant Afrikaans (which has a simplified grammar and differs in some orthographic details)
- some syntactic peculiarities shared with German
- compound analysis necessary just as for German

#14: Hindi (hi, hin): Statistics

- GDP: $739 billion (1.0% of world GDP)
- L1 speakers: 260M, L2 speakers: 120M
- coverage so far: 43.1% of world population, 88.5% of world GDP

ऋग्वैदिक काल में आर्यों का निवास स्थान सिंधु तथा सरस्वती नदियों के बीच में था। बाद में वे सम्पूर्ण उत्तर भारत में फ़ैल चुके थे। सभ्यता का मुख्य क्षेत्र गंगा और उसकी सहायक नदियों का मैदान हो गया था। गंगा को आज भारत की सबसे पवित्र नदी माना जाता है। इस काल में विश् का विस्तार होता गया और कई जन विलुप्त हो गए। भरत, त्रित्सु और तुर्वस जैसे जन् राजनीतिक हलकों से ग़ायब हो गए जबकि पुरू पहले से अधिक शक्तिशाली हो गए। पूर्वी उत्तर प्रदेश और बिहार में कुछ नए राज्यों का विकास हो गया था, जैसे – काशी, कोसल, विदेह, मगध और अंग।

#14: Hindi (hi, hin): Interesting Features

- dominant language of Northern India, shares status as official language of the Indian Union with English
- not an official language in every region (there are 21 such languages, most of them with more native speakers than Italian)
- many languages of Northern India (Bihari, Rajasthani) are very similar to Hindi, and Hindi is used as the written language in these regions
- largest language using an **abugida**, i.e. a script style where vowels are added as diacritics to a consonant backbone

#14: Hindi (hi, hin): Challenges

- many synonyms even for common terms (due to a wealth of registers): “blood”: लहू (lahū), रक्त (rakt), ख़ून (xūn), रुधिर (rudhir), लोह (loh), लोहू (lohū), अस्र (asra)
- large differences between “street Hindi” and literary Hindi

#15: Turkish (tr, tur): Statistics

- GDP: $723 billion (0.98% of world GDP)
- L1 speakers: 71M, L2 speakers: 17M
- coverage so far: 44.1% of world population, 89.4% of world GDP

Şehir devletleri tarafından yönetilen bu bölgenin müstahkem şehirleri, kral mezarları, hazineleri, Hatti kültürünün simgeleridir. MÖ 2000 yılları sonlarında büyük savaşlar sonucunda çıkan yangınlarla sona eren bu çağı, Asur Ticaret Kolonileri Çağı izler. Yazılı kaynaklardan Hititlerin, Anadolu'ya MÖ 3. binin son yıllarında, 2. binin başında küçük gruplar halinde, girmeye başladıkları ihtimali çıkmaktadır. Hititler'in Anadolu'ya Kuzey Karadeniz üzerinden veya kuzeydoğudan, Kafkaslar üzerinden geldikleri ve Kızılırmak kavisinin kuzey kesimine yerleşmiş oldukları değerlendirilmektedir.

#15: Turkish (tr, tur): Interesting Features

- by far the highest amount of synthesis among the major world languages (**agglutinating** language):
  değer-len-dir-il-mek-te-dir
  value-NtoV-CAUS-PAS-INF-LOC-be.3SG
  "is (in the process of) being evaluated"
- **vowel harmony**, i.e. suffixe vowels adapt to stem in some features:
  sınıf-lan-dır-ıl-mak-ta-dır
  class-NtoV-CAUS-PAS-INF-LOC-be.3SG
  "is (in the process of) being classified"
- as in most agglutinating languages, morpheme boundaries are quite clear, i.e. some of this complexity can be treated by tokenization below the word level

#15: Turkish (tr, tur): Challenges

- there is no way around morphological analysis
  - basic noun paradigm has 84 forms
  - basic verb paradigm has 120 forms
  - copula suffixes add a factor of 24 to noun forms
  - each verb has dozens of participle forms which again act like nouns, amounting to tens of thousands of forms for each verb
- Ottoman Turkish was a lot more influenced by Persian in both vocabulary and syntax, and quotes from that era still mark educated literary language

#16: Urdu (ur, urd): Statistics

- GDP: $527 billion (0.71% of world GDP)
- L1 speakers: 65M, L2 speakers: 94M
- coverage so far: 45.0% of world population, 90.2% of world GDP

ابھی ہم متاخر حجری عہد سے گزرتے ہی سندھ کی وادی میں ہماری نظر تمدن کے ایسے آثاروں پر پڑتی ہے کہ ہم ٹھٹھک کر رہ جاتے ہیں۔ ہم ابھی اجتماعی زندگی کی بنیاد پڑنے، بستیاں بسنے، صنعت میں کس قدر مشق و صفائی پیدا کرنے کا ذکر کر رہے تھے۔ اب یک بارگی ہمیں عالی شان شہر دیکھائی دیتے ہیں۔ ان کے مکانات پختہ اور مظبوط، دو دو تین تین منزلہ اونچے ہیں۔ ان میں سڑکیں ہیں، بازار ہیں۔ ان کے باشندوں کی زندگی و رواج اور عادات سانچے میں ڈھلی ہوئی معلوم ہوتی ہے۔

#16: Urdu (ur, urd): Interesting Features

- Hindi and Urdu form a single spoken language called Hindustani:
  کونسے دانت میں درد ہے؟
  कौन से दाँत में दर्द है?
  kaun se dããt mẽ dard hai?
  “Which tooth hurts?”
- Hindi uses Sanskrit and Urdu uses Arabic for technical terms:
  ہر ایک ملک کا اپنا قومی پرچم ہوتا ہے۔
  har ek mulk kā āpnā qaumī parcam hotā hai.
  हर एक देश का अपना राष्ट्रध्वज होता है।
  har ek deś kā āpnā rāṣṭradhvaj hotā hai.
  “Every country has its national flag.”
- difference was originally one of religion (Islam vs. Hinduism)

#16: Urdu (ur, urd): Challenges

- Arabic script not an ideal match for mixed language:
  MLK is to be read as Arabic (malak? milk?)
  PRČM is to be read as Persian (perčem? paracam?)
  AYK is to be read as Aryan (ik? īk? ayaka?)
- script support is limited in most default configurations, making Urdu the most difficult among major languages to get to display correctly
- most difficult language for OCR (preference for elegant typography):

#17: Swedish (sv, swe): Statistics

- GDP: $504 billion (0.68% of world GDP)
- L1 speakers: 10M, L2 speakers: 1M
- coverage so far: 45.1% of world population, 90.8% of world GDP

Handeln tillsammans med de nya bronsvapnen gjorde hövdingarna mäktigare, vilket manifesteras i storhögarna. Kulturpåverkan var stark och snart uppfördes också imponerande storhögar i Sverige, som Kungagraven i Kivik. Handeln bestod förmodligen främst av import av metall och salt, medan bland annat bärnsten exporterades. Även i Sverige tycks hästen och vagnen haft stor betydelse, men framförallt var skeppen viktiga. De är återgivna på åtskilliga hällristningar och i skeppssättningar.

#17: Swedish (sv, swe): Interesting Features

- very similar to Danish and Norwegian, these are only different languages due to separate political history
- definiteness marked by suffixes:
  handel "trade" → *handeln* "(the) trade"
  hövdingar "chieftains" → *hövdingarna* "the chieftains"
- pitch accent is phonemic:
  brunnen "well" vs. *brunnen* "burnt"
  modet "the courage" vs. *modet* "the fashion"

#17: Swedish (sv, swe): Challenges

- hard to transcribe due to lack of standard pronunciation: *svartsjuk*, anything between [sva:rtʃʉ:k] and [svaʈɧʉk]
- some syntactic peculiarities shared with German
- compound analysis necessary just as for German

#18: Polish (pl, pol): Statistics

- GDP: $475 billion (0.64% of world GDP)
- L1 speakers: 55M, L2 speakers: 1M
- coverage so far: 45.8% of world population, 91.5% of world GDP

Istniały pewne strefy, gdzie ludność kultury pucharków lejkowatych podlegała wpływom kompleksu badeńskiego, będącego pośrednim między helladzką epoką brązu a późnoeneolitycznymi kulturami południowej Polski. Z tymi kontaktami wiąże się upowszechnienie ciałopalnego obrządku grzebalnego. W stylistyce form ceramicznych wpływy badeńskie oznaczały przejęcie form ceramicznych zdobionych ornamentem kanelowanym, często rozchodzącym się promieniście od dna naczynia.

#18: Polish (pl, pol): Interesting Features

- richest case system among major Indo-European languages (six cases plus fully productive vocative!)
- conservative among Slavic languages, no major substrate influence
- very exact correspondence between script and pronunciation
- loanwords are adapted to the native orthography very quickly: *landszaft*, *kindersztuba*, *grynszpan*, *kicz*

#18: Polish (pl, pol): Challenges

- more irregular case forms than in Russian
- non-trivial normalization steps necessary for morphological analysis:
  w biurze → *biuro* “office”
  w puszce → *puszka* “can”
  w szkole → *szkoła* “school”
- some inflectional morphemes behave more like clitics

#19: Persian (fa, fas): Statistics

- GDP: $437 billion (0.59% of world GDP)
- L1 speakers: 50M, L2 speakers: 60M
- coverage so far: 46.5% of world population, 92.1% of world GDP

آخرین پادشاه فهرست شاهان شوش، پوزور اینشوشینک، ابتدا شوش و سپس انشان را فتح کرد و به نظر می‌رسد که با مطیع‌کردن پادشاه سیماشکی توانست وحدت کوتاه‌مدتی در فدراسیون ایلامی به‌وجود آورد؛ اما جانشینان او نتوانستند که شوش را در قلمرو ایلام نگه‌دارند. پوزور اینشوشینک چندین کتیبه با نام خود در شوش باقی گذاشته‌است. برخی از آنان به اکدی نوشته‌شده‌اند و بقیه به خط ایلامی باستان یا ایلامی خطی؛ ایلامی خطی یک نظام نگارشی است که تنها معدودی از نشانه‌های آن با قطعیت رمزگشایی شده‌اند. شاید نشانه‌های ایلامی خطی از ایلامی ابتدایی گرفته شده باشند.

#19: Persian (fa, fas): Interesting Features

- primary language of Islamic culture and education in Asia, source of many loans in Turkish, Hindi, and other languages of the region
- typologically quite far removed from its Iranian ancestors:
  - no grammatical gender
  - very rudimentary case system
  - copula as a clitic
- nearly every Arabic loanword has a purist alternative:
  "history": تاریخ *tārīx* vs. پیشینه *pišine*
  "celebration": عید *ejd* vs. جشن *ǧašn*
  "prayer": دعا *do'ā* vs. نیایش *nijāješ*
  "worship": عبادت *ebādat* vs. پرستش *parasteš*

#19: Persian (fa, fas): Challenges

- transcription: Arabic writing system is problematic for a language where vowels make a difference in meaning:
  "who?" *ki* کی
  "when?" *kej* کی
- an important grammatical ending (the *ezāfe*) is not written after consonants, and needs to be inferred from context:
  "name" نام *nām*
  "(him)self" خود *xod*
  "his own name" نام خود *nām-e xod*
- spacing not entirely standardized

#20: Thai (th, tha): Statistics

- GDP: $395 billion (0.54% of world GDP)
- L1 speakers: 20M, L2 speakers: 44M
- coverage so far: 46.8% of world population, 92.6% of world GDP

เศรษฐกิจของชุมชนทวารวดีคงจะมีพื้นฐานทางการเกษตรกรรม
มีการค้าขายแลกเปลี่ยนระหว่างเมือง
หรือการค้าขายแลกเปลี่ยนกับชุนชนภายนอก
ชุมชนทวารวดีเริ่มต้นแนวความเชื่อแบบพุทธศาสนา ในลัทธิเถรวาท
ควบคู่ไปกับการนับถือศาสนาพราหมณ์หรือฮินดู ทั้งลัทธิไศวนิกาย
และลัทธิไวษณพนิกาย โดยศาสนาพราหมณ์
หรือศาสนาฮินดูจะแพร่หลายในหมู่ชมชนชั้นปกครอง

#20: Thai (th, tha): Interesting Features

- a prototypical example of an isolating language
- very rich vowel and diphthong inventory
- script combines etymological spelling with pronunciation hints:
  อาจารย์ *ācāryX* [ʔaː˦˥.t͡ɕaːn˦] “professor”
  คริสต์มาส *khristXmās* [kʰrit̚˥˩.saˌ.maːt̚˧] “Christmas”

#20: Thai (th, tha): Challenges

- transcription is a challenge due to Brahmi-derived script
  - Indo-Aryan languages distinguish four types of plosives
  - Thai uses complex rules to distinguish tone by a combination of homophonic consonant symbols, coda structure, and diacritics
- one of the most complex languages to sort alphabetically
  - the vowel is written in front of the coda in some syllables
  - tone diacritics need to be ignored, but not vowel diacritics
- tokenization a challenge due to absence of spaces

Further Positions in the Language Ranking

#21	no	nor	Norwegian
#22	--	yue	Cantonese
#23	bn	ben	Bengali
#24	da	dan	Danish
#25	he	heb	Hebrew
#26	fi	fin	Finnish
#27	el	ell	Greek
#28	vi	vie	Vietnamese
#29	mr	mar	Marathi
#30	ro	ron	Romanian
#31	cs	ces	Czech
#32	tl	tgl	Tagalog
#33	kk	kaz	Kazakh
#34	hu	hun	Hungarian
#35	te	tel	Telugu

#36	sh	hbs	Serbo-Croatian
#37	kn	kan	Kannada
#38	ta	tam	Tamil
#39	ha	hau	Hausa
#40	sw	swa	Swahili
#41	sk	slk	Slovak
#42	uk	ukr	Ukrainian
#43	zu	zul	Zulu
#44	pa	pan	Punjabi
#45	my	mya	Burmese
#46	si	sin	Sinhalese
#47	am	amh	Amharic
#48	ml	mal	Malayalam
#49	az	aze	Azerbaijani
#50	as	asm	Assamese

Smaller Languages

Promises of computational linguistics on smaller languages:

- can be very interesting typologically
  - no polysynthetic language among the world's major languages
  - media language is quite uniform across the globe, texts from smaller more oral cultures tend to be much more diverse
- it can have a huge positive impact on the language's viability
  - spell checking encourages written usage,
    helps to kickstart exposure to the written language
  - comutational lexicography helps in standardization
    and fosters applicability across all domains
  - specialized search engines help to stay within the language's microcosm
- strategic consideration for researchers: if you are the only person working on that language, publishable progress is more likely

Smaller Languages

Challenges of computational linguistics on smaller languages:

- lack of documentation except for select minority languages (see below)
- even if standardized variants exist, text availability often very low
- necessity to deal with grammatical features about which little previous literature and no standard approaches exist
- no chance of commercial viability (external funding needed)

Attractive example languages due to availability of texts:

- Basque: unique outlier in Europe, vibrant community
- Greenlandic: a polysynthetic national language (with few speakers)
- Kabardian: large community of speakers, very interesting typology