COURSE OVERVIEW

Using speech technology in the real world

Course 1 (Monday): standard low-resource techniques

- I. Introduction to speech tech for under-resourced languages
- II. Domain adaptation (dealing with few labels)
- III. Semi-supervised techniques (creating pseudo-labels)
- IV. Distant supervision (unaligned speech and labels)

Course 2 (Tuesday): the case of unwritten languages

- V. Weakly supervised learning (dealing with bad labels)
- VI. Unsupervised learning (no labels)
- VII. Language emergence (no data)

Course 3 (Wednesday): Hot topic in the low-resource setting

- VIII. self supervised pretraining (CPC and beyond)

Course/reading Group 4 (Friday). Special topic: beyond sentences

On-line questionnaires

Note, for the courses 1 and 2, there will be a questionnaire (google form) associated with each of the video.

- You can fill in the questionnaire during the video (by pausing it) or shortly afterwards
- You can fill the questionnaire by discussing it on-line with your colleagues
- You should not spend more than a maximum allotted time on each questionnaire (about 50%-60% of the duration of the video)
- The questionnaire won’t be graded
- it is solely to help you
    - Understand and learn the materials
    - Prepare the project (some questions will be on the language you’ll choose for your dataset)
    - Prepare the Q&A session

Tutorials

Tuto 1 (Mon): How to build your own dataset

Tuto 2 (Tues): (continued)

Tuto 3 (Wed): Unsupervised pretraining with CPC

Tuto 4 (Thur): Fine tuning a pretrained model

Project

Build an ASR for your own language with only 1 hour of labeled data!

- Create your own 2h dataset
- Pretrain using CPC self-supervision
- Fine tune using CTC
- Compute the CER

(optional: use an LM)

The dataset will be constructed by you and put into your github
The base model will be provided as a colab notebook

Exam and evaluation

QUIZZ (Friday): short comprehension questions about the course

TUTORIALS: Datasets and notebooks will be evaluated

PROJECT: notebook will be evaluated

The week in brief

	Mon 22	Tues 23	Wed 24	Thur 25	Frid 26		Mon 29	Tues 30	Wed 1	Thur 2	Frid 3
9-10am	Course 1 Low res.	Course 2 unwritten	Course 3 Unsup pretrain		Reading group						
10-11am											
11-12am											
12am -2pm											
2-3pm	Q&A	Q&A	Q&A	Tuto 4 CPC	QUIZZ						Dead-line for project
3-4pm	Tuto 1 dataset	Tuto 2 dataset	Tuto 3 CPC		Project Q&A						
4-5pm											

!! Paris timezone=+2 hours !!

The teachers

Emmanuel Dupoux
Prof EHESS, Paris
Research Scientist, FAIR

Laurent Besacier
Prof U. Grenoble
Research Scientist, Naver Labs

Morgane Rivière
Research Engineer, FAIR

Resources:
https://github.com/besacier/AMMIcourse

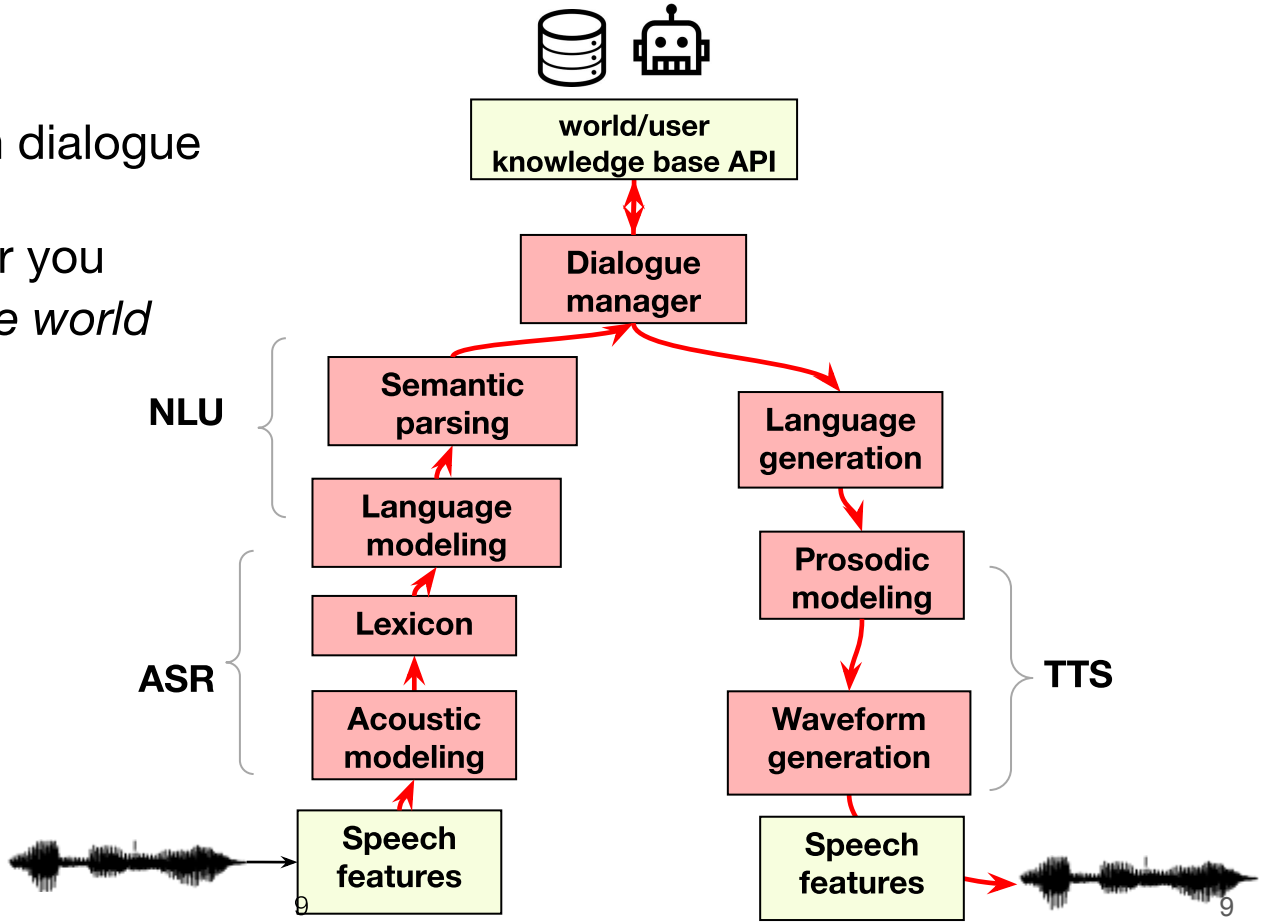
I. Introduction to speech processing for under-resourced languages

The dream application: the personal assistant

Overall objective

- Construct a whole spoken dialogue system
- That does useful things for you
- *For all the languages of the world*

Amazon Alexa,
Google Home,
Baidu Raven, etc



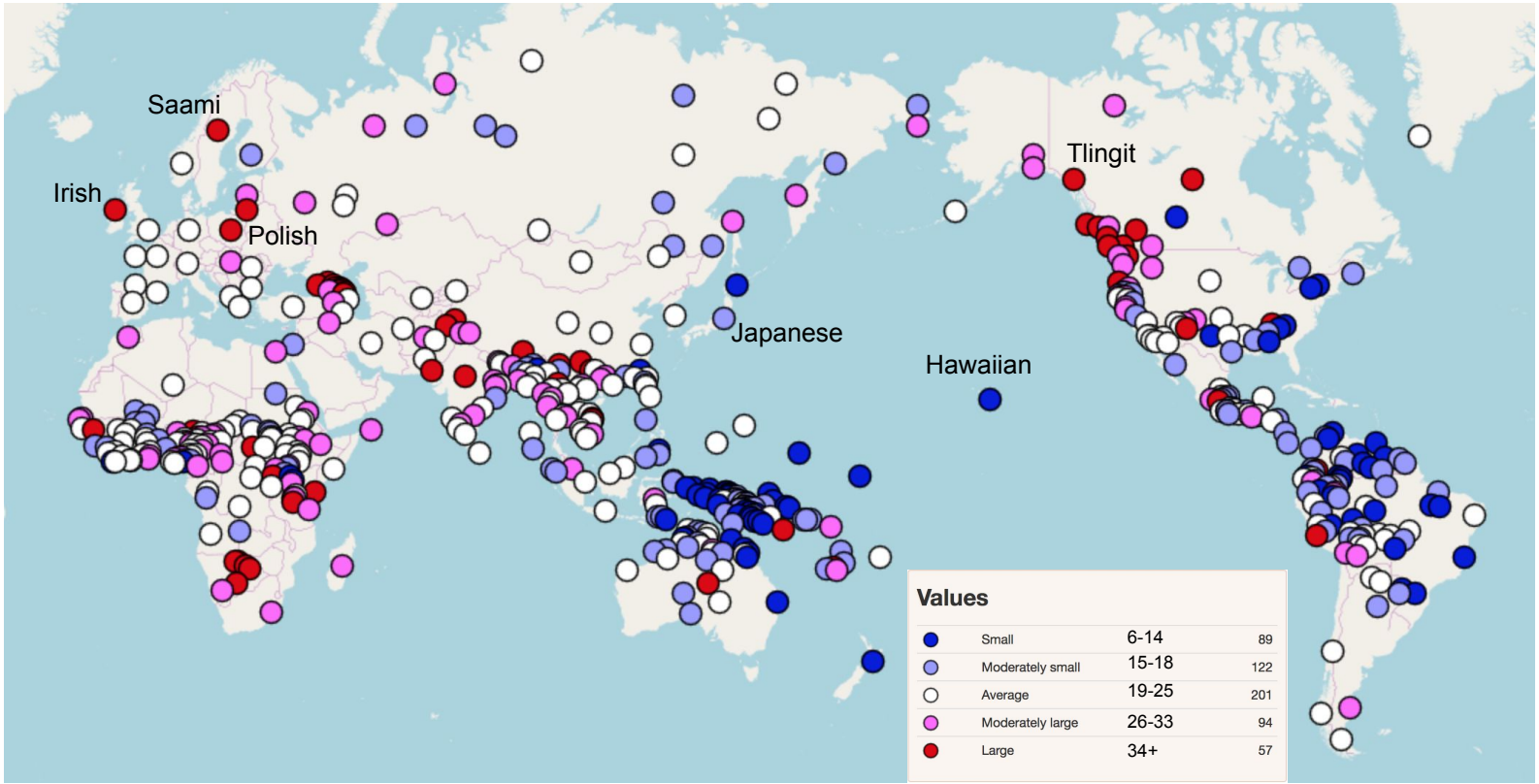
9

3 major challenges for the scaling up of speech technologies to the languages of the world

- Language **diversity**
- Language **variation**
- Language **sparsity**

Language diversity

AM: Phonological diversity



AM: Phonological diversity

Central Rotokas	Bilabial	Alveolar	Velar
Voiceless	p	t	k
Voiced	b ~ β	d ~ ɾ	g ~ ɣ

East !Xoon dialect: Non-click consonants (Traill 2018)

		Labial	Dental	Alveolar	Palatal	Velar	Uvular	Glottal
Plosive/ affricate	voiced	b	d̪	dz	ɟ [ʝ ~ j]	g	ɢ [ᴺɢ]	
	tenuis	p	t̪	ts		k	q	ʔ
	voiceless aspirated	pʰ	t̪ʰ	tsʰ		kʰ	qʰ	
	voiced aspirated		d̪tʰ	dtsʰ		gkʰ	ɢqʰ [ᴺɢqʰ]	
	velarized	px	t̪x	tsx				
	voiced velarized[16]		d̪t̪x	dtsx				
	voiceless ejective		t̪’	ts’		k’	q’	
	ejective cluster[17]	p’kx’	t̪’kx’	ts’kx’		kx’		
	voiced ejective[18]		d̪t̪’kx’	dts’kx’		gkx’		
Fricative	voiceless	f		s		x		h
Nasal	voiced	m	n̪		-ɲ-	-ŋ		
	glottalized	ˀm		ˀn				
Other		-β-		-l-	-j-			

bilabial clicks	dental clicks	lateral clicks	alveolar clicks	palatal clicks
ʘ	ǀ	ǁ	ǃ	ǂ
gʘ	gǀ	gǁ	gǃ	gǂ
ʘq	ǀq	ǁq	ǃq	ǂq
ʘɢ	ǀɢ	ǁɢ	ǃɢ	ǂɢ
ʘʰ	ǀʰ	ǁʰ		ǂʰ
gʘh	gǀh	gǁh	gǃh	gǂh
ʘqʰ	ǀqʰ	ǁqʰ	ǃqʰ	
	ɢǀqʰ	ɢǁqʰ	ɢǃqʰ	ɢǂqʰ
ʘx	ǀx	ǁx	ǃx	ǂx
gʘx	gǀx	gǁx	gǃx	gǂx
ʘkx’	ǀkx’	ǁkx’	ǃkx’	ǂkx’
gʘkx’	gǀkx’	gǁkx’	gǃkx’	gǂkx’
kʘˀ	kǀˀ	kǁˀ	kǃˀ	kǂˀ
ʘk’	ǀk’	ǁk’		ǂk’
ʘq’	ǀq’	ǁq’	ǃq’	ǂq’
n̥ʘ	n̥ǀ	n̥ǁ	n̥ǃ	n̥ǂ
nʘ	nǀ	nǁ	nǃ	nǂ
’nʘ	’nǀ	’nǁ	’nǃ	’nǂ
ŋ̊ʘh	ŋ̊ǀh	ŋ̊ǁh	ŋ̊ǃh	ŋ̊ǂh

AM: Phonological diversity

Syllables are formed of phoneme sequences

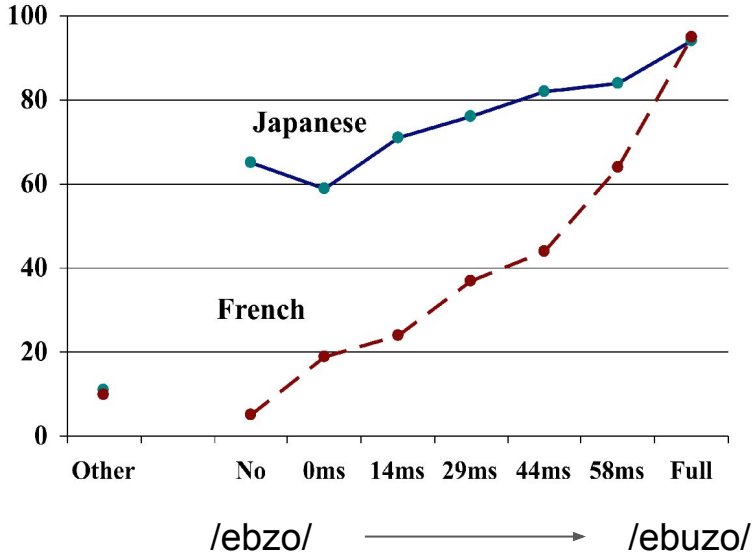
In most languages, some syllables are valid, some are not

Japanese: only V, CV, VN, CVN allowed

> phonological adaptation of borrowings:

sphinx > /sufiNkusu/

Christmas > /kurisumasu/



14

AM: Phonological diversity

Different vowel/consonant frequencies and cluster usage:

Georgian /gvbrdɣvnis/ ‘he's plucking us’

Nuxalk (“Bella Coola”) *clhp’xwlhtlhplhhskwts’* /xɬpʼχʷɬtʰɬpʰɬːskʷʰt͡sʼ/
‘he had possessed a bunchberry plant’

Hawaiian *He aha kēia?* ‘What is it?’

15

Lexicon: Morphological diversity

- Analytic and isolating languages
  - Each word carries exactly one meaning
  - Ex.: Chinese /ɥɔ^{214} mən^{4} tʰa̲n^{35} kɑŋ^{55}ʧɕʰin^{35} lə^{5}/ (1st_pers plur PLAY PIANO past) ‘we played the piano’

- Synthetic languages
  - Agglutinative
    - Each word can have several morphs, each carrying one meaning
    - Ex.: Turkish *el-ler-imiz-in* (HAND-pl-poss1pl-genitive) ‘of our hands’
  - Fusional
    - Each word can have several morphs, each carrying one or more meanings, of which (generally) only one lexical morph (ex.: inflectional morphology, i.e. conjugation, declension…)
    - Ex.: Latin *rexistis* /rek-s-is-tis/ (RULE-perf-perf-perf.2sg) ‘you$_{PLUR}$ ruled’
  - Polysynthetic
    - Each word can have several lexical or grammatical morphs
    - Ex.: Island Halkomelem (Salish) *hwpulqwitth’a’ustum*
      (locative-GHOST/DEATH(?)-blanket/cloth-face-transitive-passive)
      ‘to be adversely affected by a spirit entering the body through the face’

Lexicon: Morphological diversity

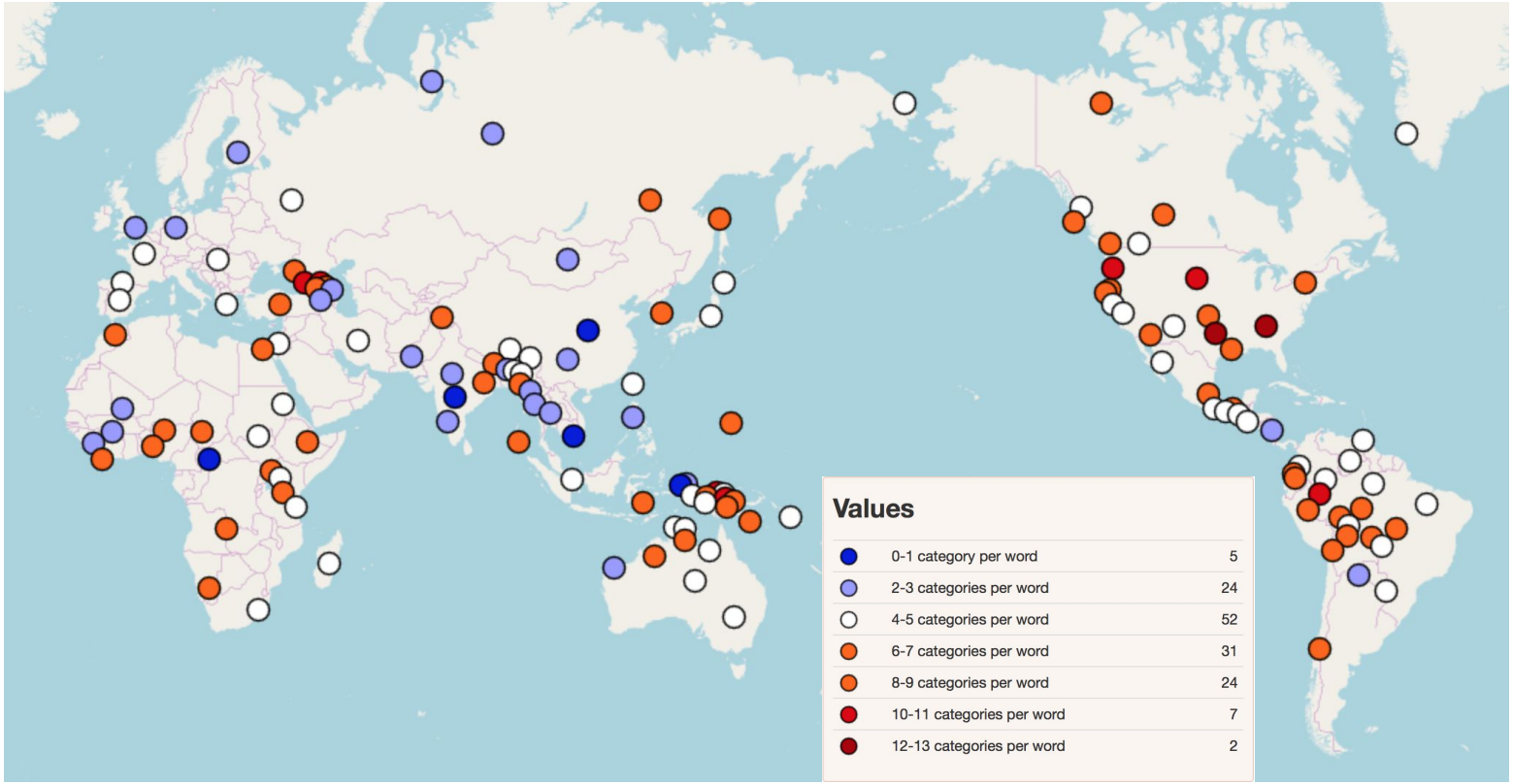
Most languages show elements of different morphological types

- Ex.: English!
  - *the boy will play with the dog*
  - *John's cat eat**s** mice* (inflexional morphology*)
  - ***anti****dis**establish**ment**arian**ism*** (derivational morphology)
- Other example: creating words or word-like sequences from sentences
  - French: *je-m'en-foutisme*
  - English: *You know, I can't take all this let's-be-faithful-and-never-look-at-another-person routine, because it just doesn't work* (The Boys in the Band, 1970)

* additional morphemes don't change the meaning or the grammatical category

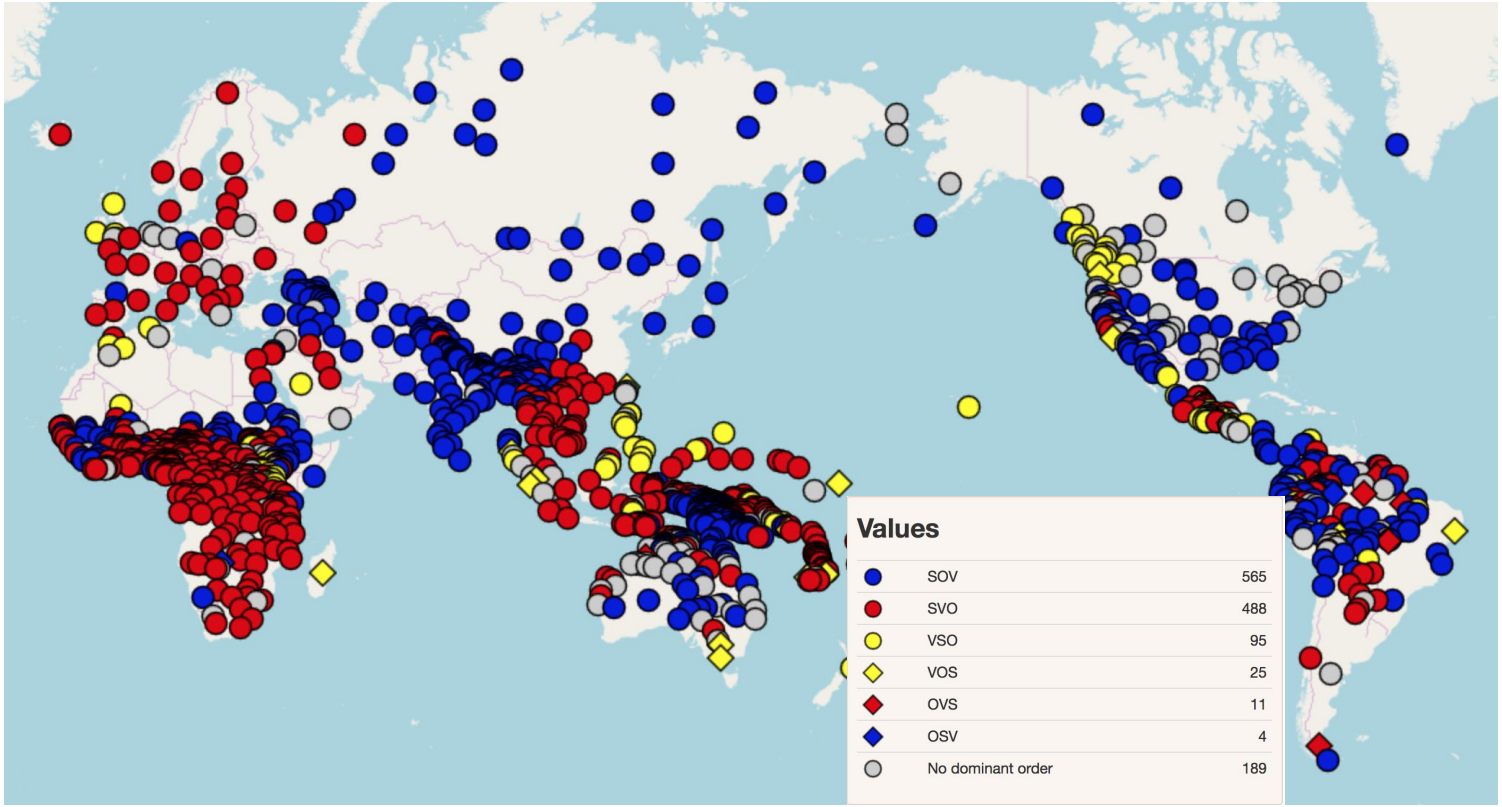
17

Lexicon: Morphological diversity



18

LM: Syntactic diversity



19

LM: Syntactic diversity

Levels of configurationality

- Free word order (often with very rich morphological marking)
  - Ex.: Warlpiri
- Relatively free word order
  - Often with rich morphological marking
  - And discontinuous constituents
  - Ex.: Polish ‘John went to the cinema’

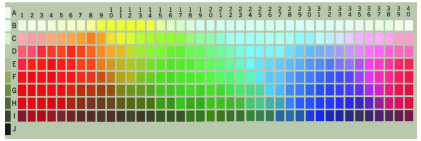
Jaś poszedł do kina.
Poszedł Jaś do kina.
Jaś do kina poszedł.
Poszedł do kina Jaś.
Do kina Jaś poszedł.
Do kina poszedł Jaś.

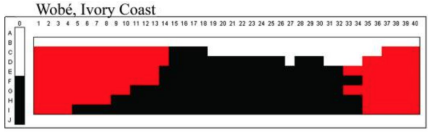
- Constrained word order (“configurational”)
  - Ex.: English, Chinese
  - Often with limited or no morphological marking
  - Discontinuous constituents are rare

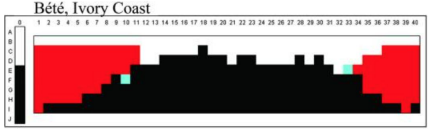
NLU: Semantic diversity

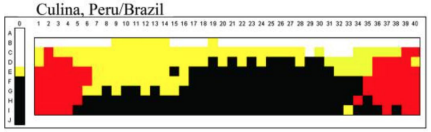
Words (fuzzily) partition the semantic space

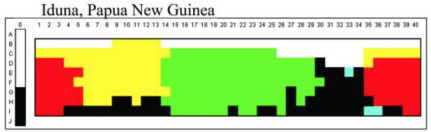
Partitions can differ from one language to another

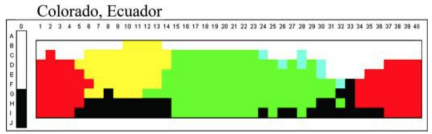


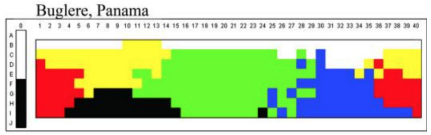


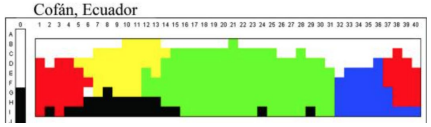


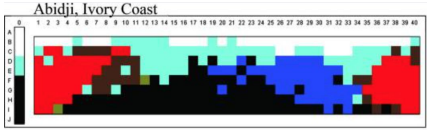


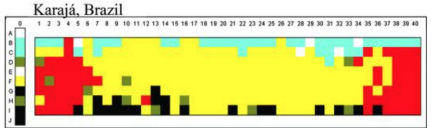


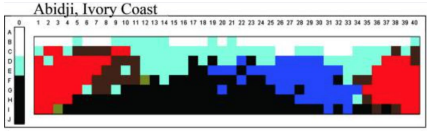


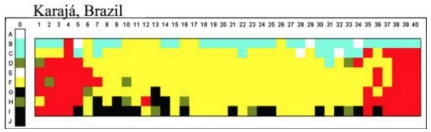








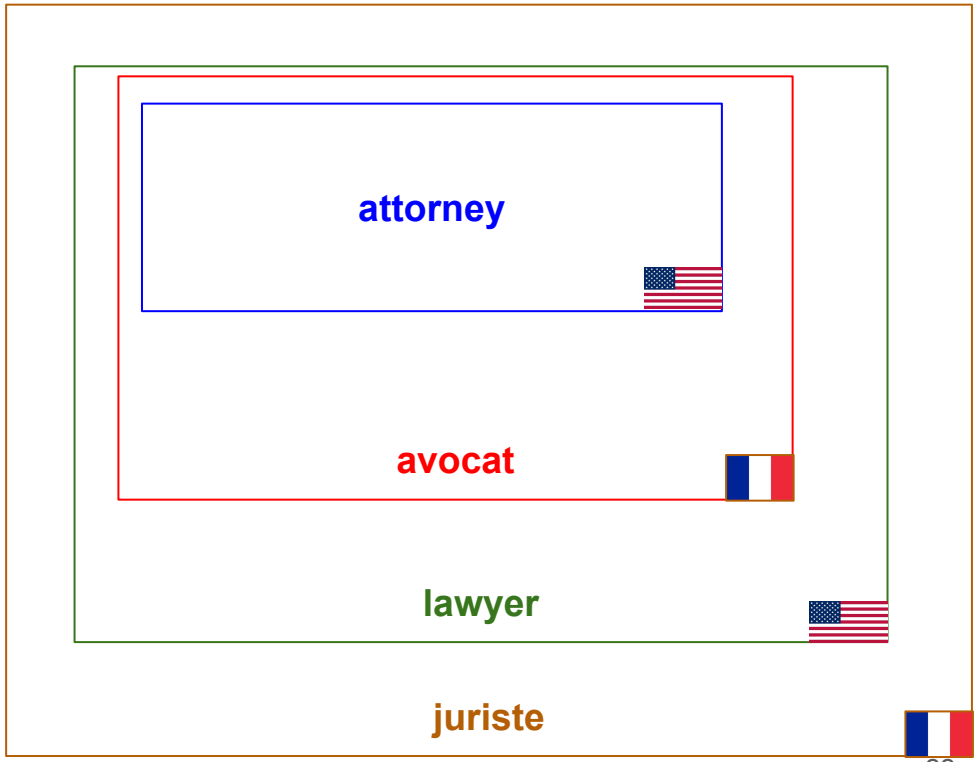




NLU: Semantic diversity

Words (fuzzily) partition the semantic space

Partitions can differ from one language to another



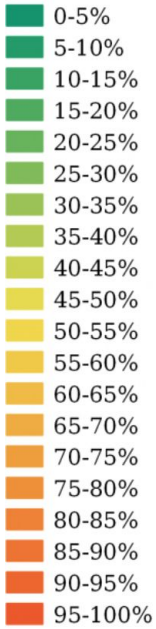
22

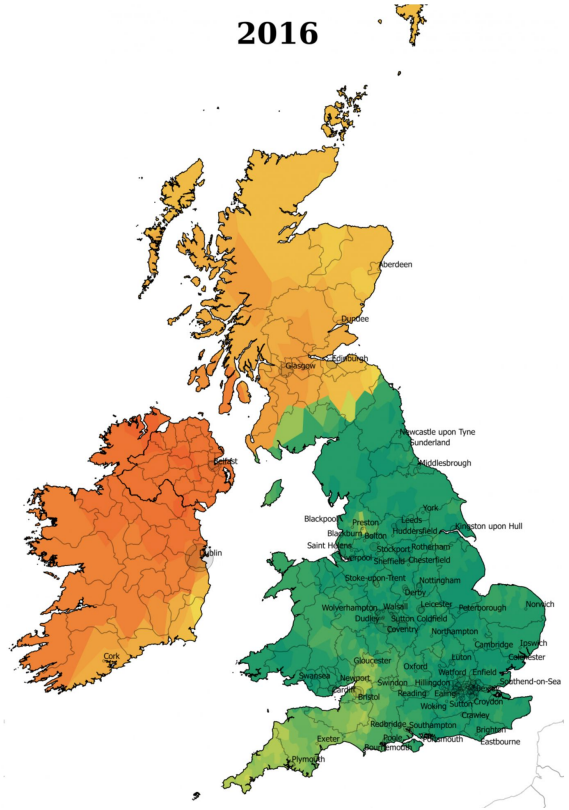
Language variation

23

Phonetic and phonological variation

Do you pronounce the “r” in “arm” ?



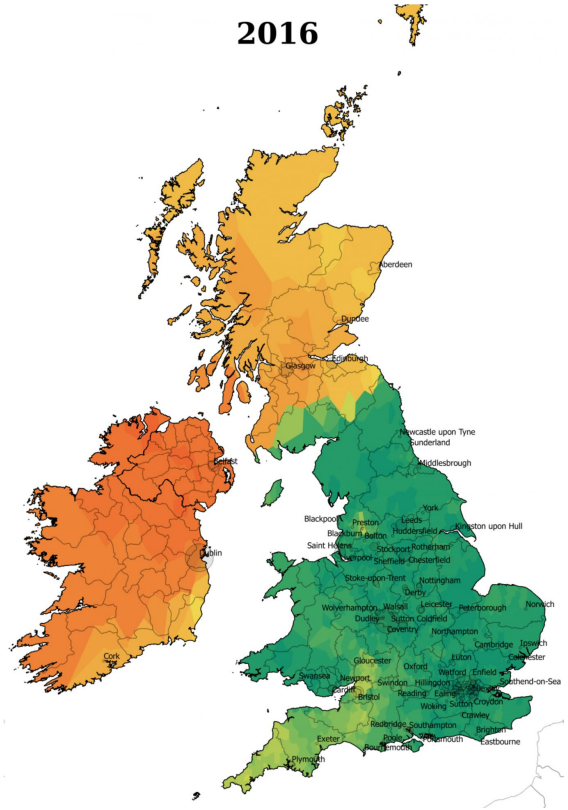


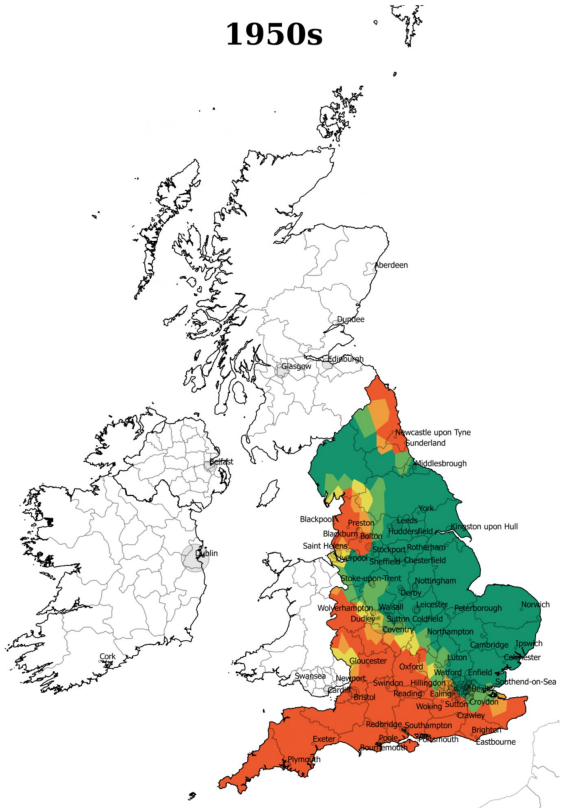
24

Phonetic and phonological variation

Do you pronounce the “r” in “arm” ?

- 0-5%
- 5-10%
- 10-15%
- 15-20%
- 20-25%
- 25-30%
- 30-35%
- 35-40%
- 40-45%
- 45-50%
- 50-55%
- 55-60%
- 60-65%
- 65-70%
- 70-75%
- 75-80%
- 80-85%
- 85-90%
- 90-95%
- 95-100%





25

Spelling “variation”

anagement maagement maanagement
maangement magagement magement
mamagement mamangement manaagement manaement
managaement manageement manageemnt managegment
managemaent managemant managememt managemen managemenet
managementt managemet managemetn managemnent managemnet
managemnt managemrnt managemt managenent managenment managent
managerment managhement managmeent managmement managment managnment
manament manamgement mananement manangment manasgement
manegement manegment mangaement mangagement mangagment
mangament mangement manggement mangment
mangmt menagement mgmt mgnt
mnagement mngmnt mngmt

Sociolinguistic variation

Interpreting tweets produced by Chicago gang members

Tweet	Label	Youth Interpretation
If We see a opp Fuck it We Gne smoke em 😈	Aggression (Threat)	he mean like if he see opp he go kill him opp mean like the people he dont like
Dnt get caught on Dat 800 block lame ass Lil niggas Betta take Dat Shyt on stony spot	Aggression (Insult)	he saying them lil nigga better not get caught on the 800 block or they go kill them so he tell them if they wanna live they better stay on stony
Young niggas still getting shot babies still dying 🙏	Loss	he mean like teen keep die and babys and kid keep die

Blevins et al. 2016

Sociolinguistic variation

T'as vu il l'a bien cherché wsh #AperoChezRicard

> +10000, shah!

> tabuz, lavé rien fé

> ki ca ? le mec ou son chien ?

> Wtf is wrong with him ? #PETA4EVER

> ki ca ? le chien ?

> looool

Sociolinguistic variation

T'as vu il l'a bien cherché wsh #AperoChezRicard
> +10000, shah!
> tabuz, lavé rien fé
> ki ca ? le mec ou son chien ?
> Wtf is wrong with him ? #PETA4EVER
> ki ca ? le chien ?
> looool

BING translation:
You saw coming it #AperoChezRicard wsh
> +10000, shah!
> tabuz, washed anything fe
> Ki ca? the guy or his dog?
> WTF is wrong with him? #PETA4EVER
> Ki ca? the dog?
> looool

29

Diachronic variation

Li reis Marsilie esteit en Sarraguce.
Alez en est en un verger suz l'umbre;
Sur un perrun de marbre bloi se culchet,
Envirun lui plus de vint milie humes.
Il en apelet e ses dux e ses cuntes:
« Oëz, seignurs, quel pecchet nus encumbret :
Li emper[er]es Carles de France dulce
En cest païs nos est venuz cunfundre.
Jo nen ai ost qui bataille li dunne,
Ne n'ai tel gent ki la sue derumpet.
Cunseilez mei cume mi savie hume,
Si m(e) guarisez e de mort et de hunte. »
N'i ad paien ki un sul mot respundet,
Fors Blancandrins de Castel de Valfunde.

Chanson de Roland (XIth century)

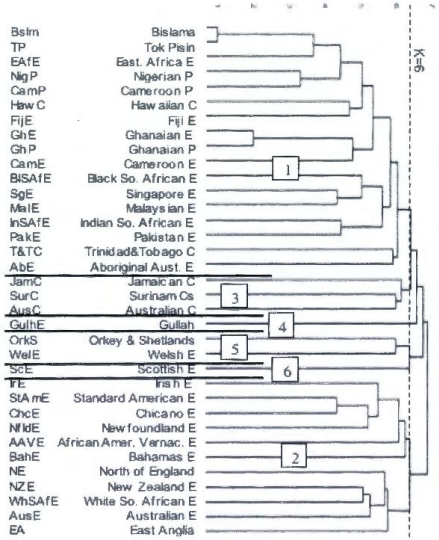
Hwæt! Wé Gárdena in géardagum
þéodcyninga þrym gefrúnon·
hú ðá æþelingas ellen fremedon.
Oft Scyld Scéfing sceaþena þréatum
monegum maégþum meodosetla oftéah·
egsode Eorle syððan aérest wearð
féasceaft funden hé þæs frófre gebád·
wéox under wolcnum· weorðmyndum þáh
oð þæt him aéghwylc þára ymbsittendra
ofer hronráde hýran scolde,
gomban gyldan· þæt wæs gód cyning.

Beowulf (Xth century)

Consequences for speech technologies

- Google Speech API
  - 13 versions of English
    - English (Australia) en-AU
    - English (Canada) en-CA
    - English (Ghana) en-GH
    - English (Great Britain) en-GB
    - English (India) en-IN
    - English (Ireland) en-IE
    - English (Kenya) en-KE
    - English (New Zealand) en-NZ
    - English (Nigeria) en-NG
    - English (Philippines) en-PH
    - English (South Africa) en-ZA
    - English (Tanzania) en-TZ
    - English (United States) en-US
  - 20 versions of Spanish!
  - etc

Trudgill 1999



Nagy et al, 2006

31

Language sparsity

32

Corpora

Corpus = body of text stored in a machine-readable form

Corpora can be annotated, for serving as training, development or test data

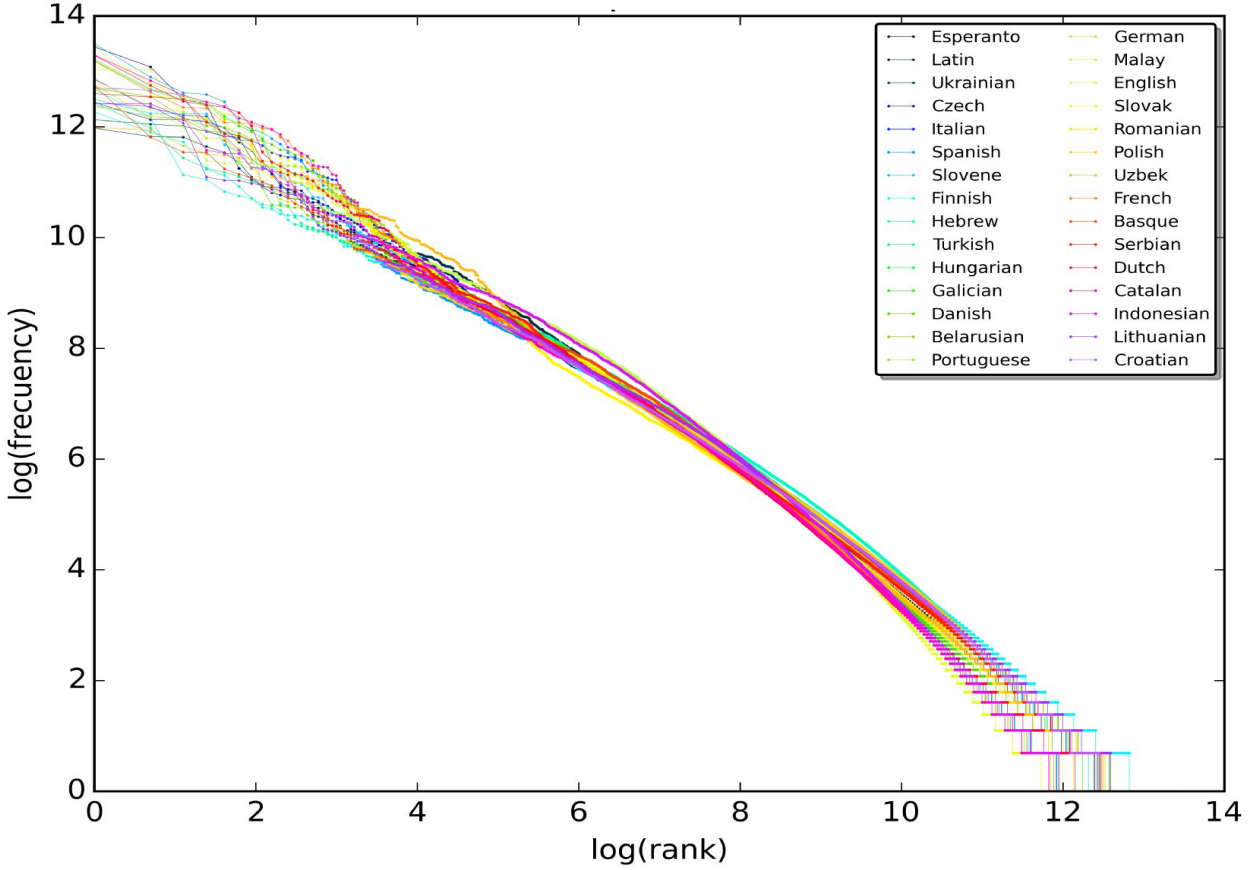
- Morphosyntactically-annotated corpora
- Treebanks (syntactically-annotated)
- Semantically disambiguated corpora
- etc.

33

Zipf's law

A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias

(source: Wikipedia; data: dumps from Oct 2015)

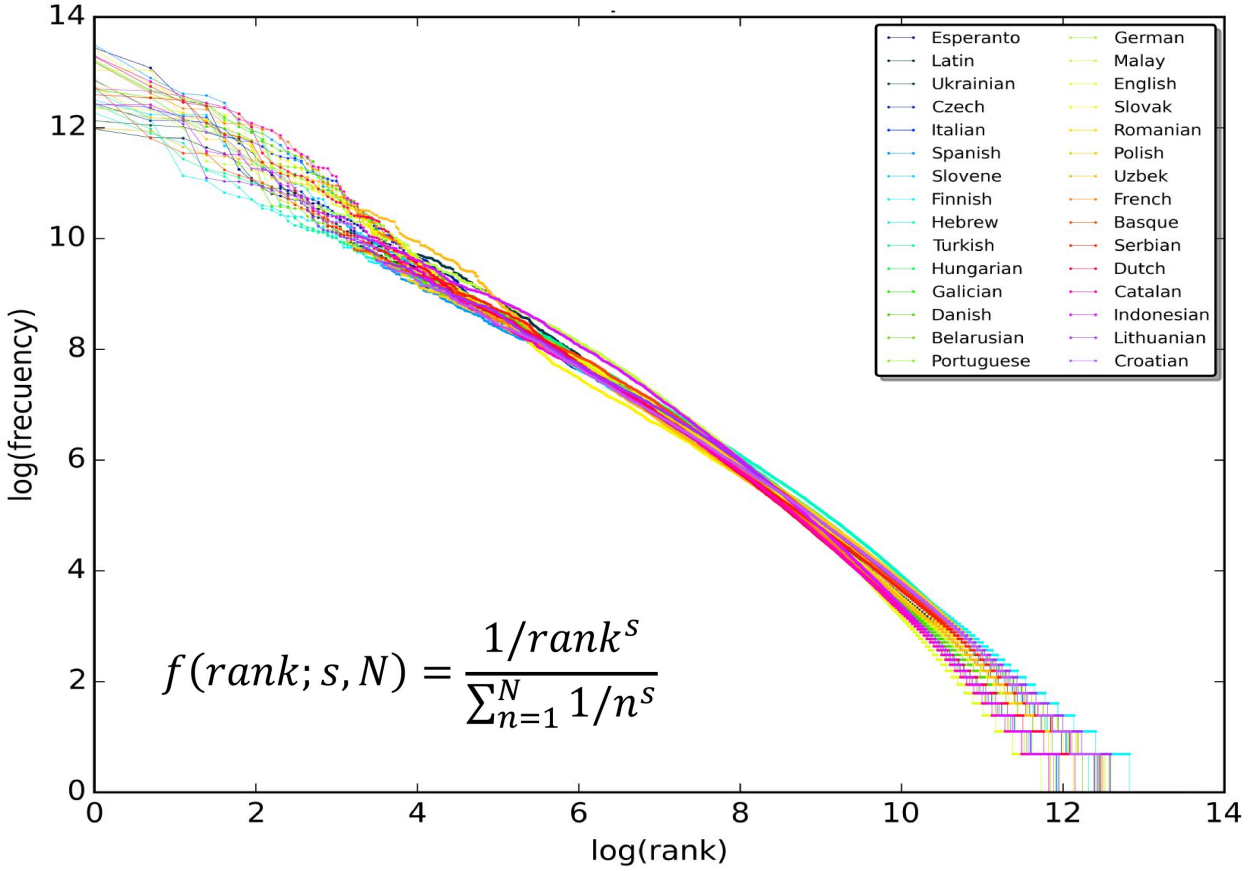


34

Zipf's law

A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias

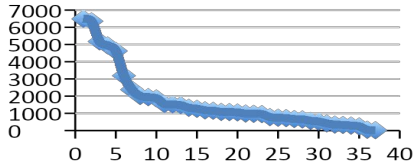
(source: Wikipedia; data: dumps from Oct 2015)

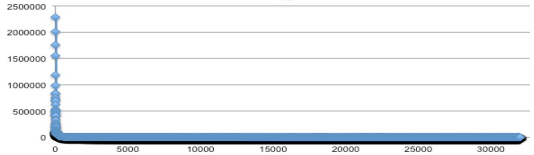


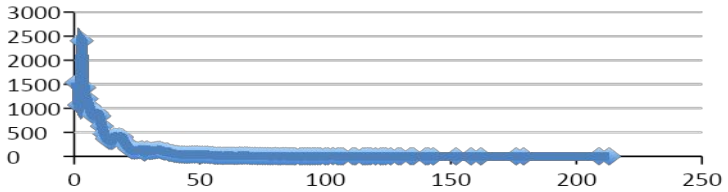
35

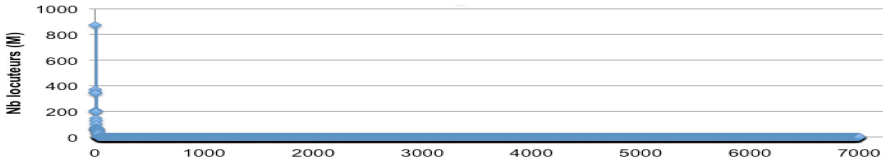
Power law, everywhere

- sounds
- words
- sentences
- languages



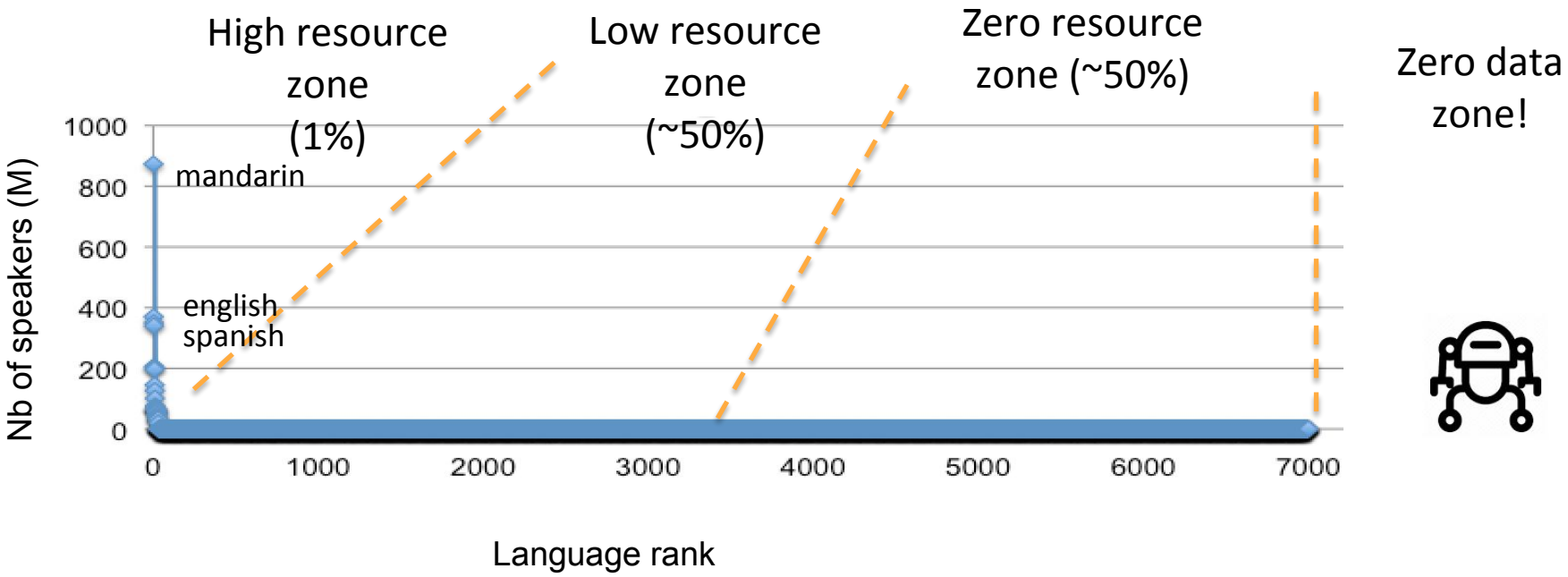




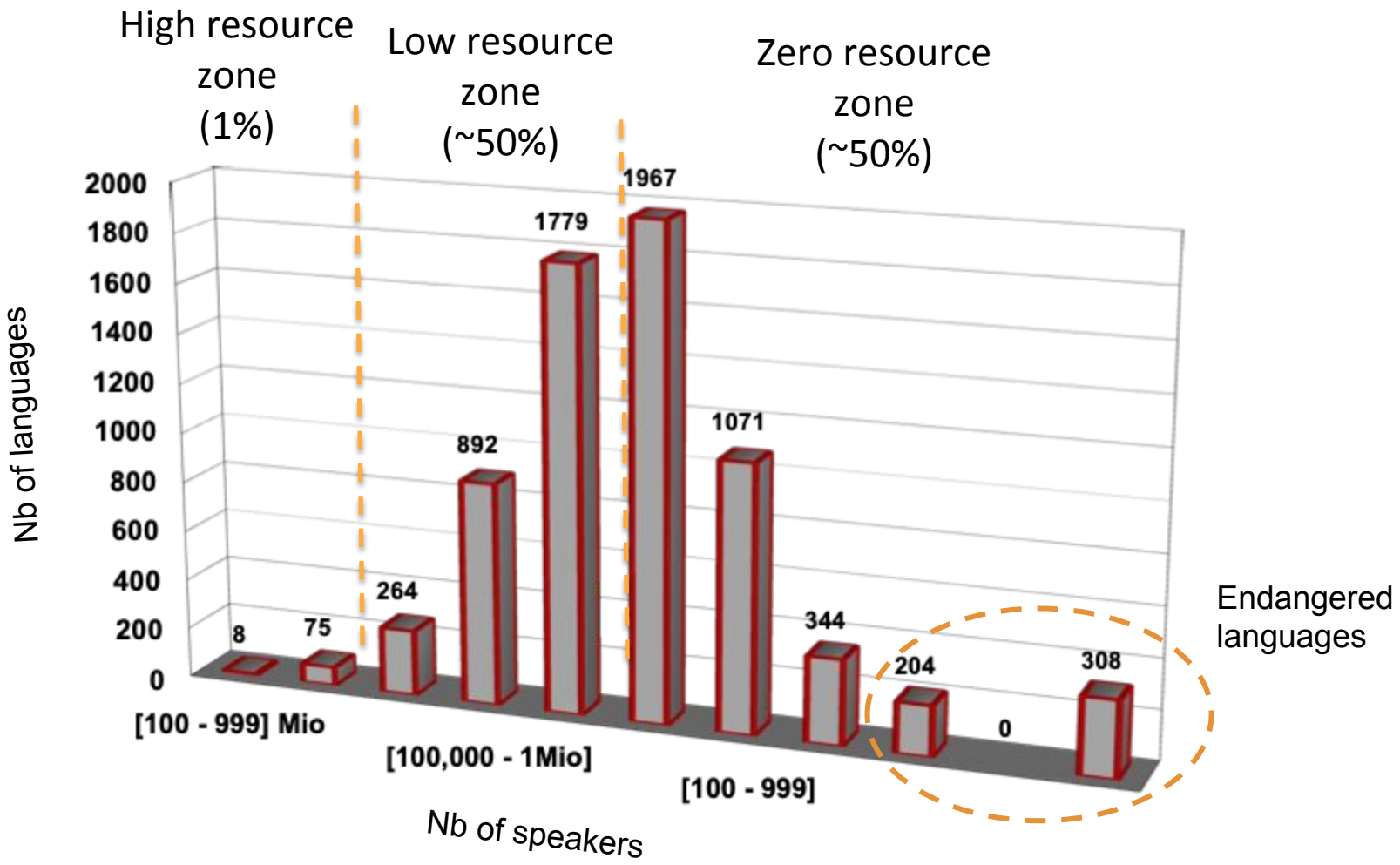


36

High vs low resource languages



37



survey by the *Summer Institute of Linguistics* (SIL) from February 1999

38

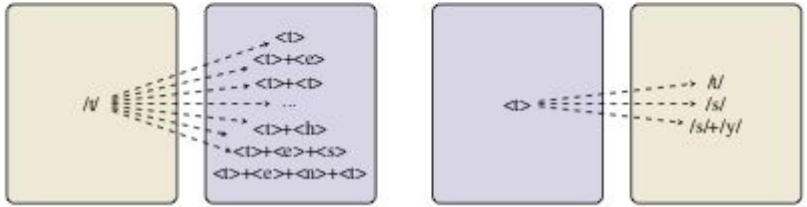
More issues

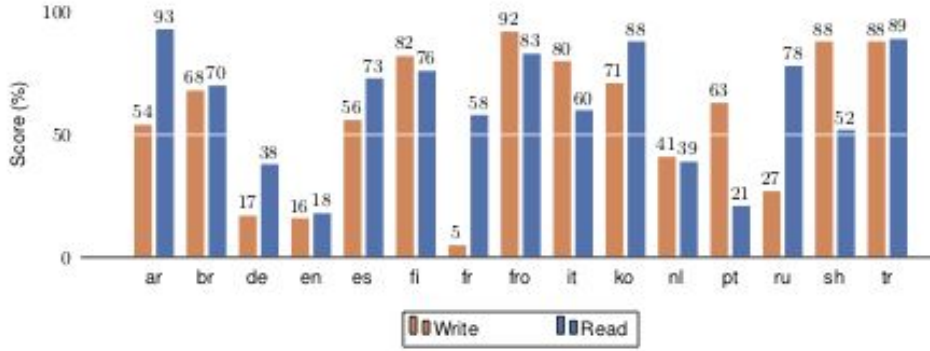
39

Oral vs written language

- Orthographic opacity
  - phonemes!=letters!

Figure 1: Example of unambiguous correspondence during writing and reading tasks in Esperanto.





OTEANN: Estimating the Transparency of Orthographies with an Artificial Neural Network Xavier Marjou (2019)

40

Oral vs written language

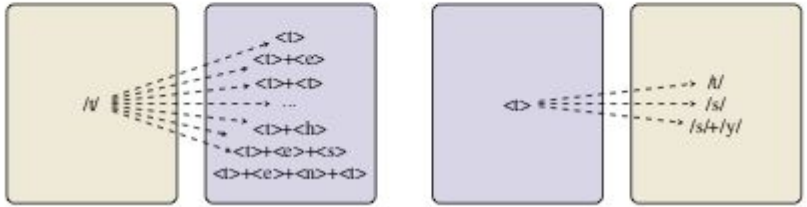
- Orthographic opacity
  - phonemes!=letters!
- (lack of) orthographic standardization

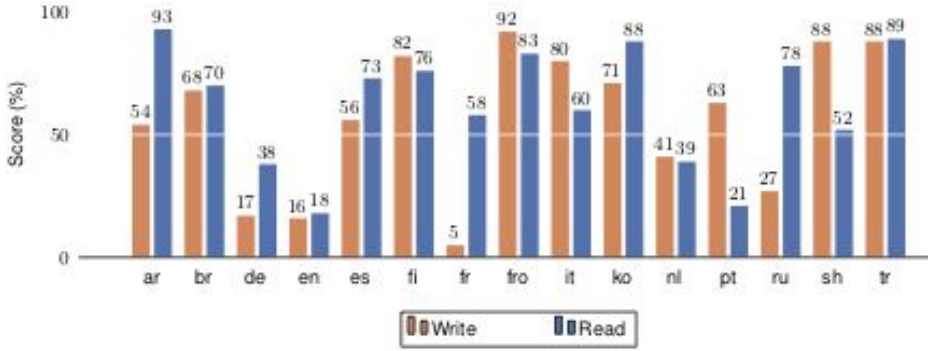
/mabiPulha:S/ *

(‘he does not say it)’

Arabic Orthography	Arabic Transliteration	Frequency
مبيقولهاش	*mbyqwlhAI*	≈ 26,000
ما بيقولهاش	*mA byqwlhAI*	≈ 13,000
مابقلهاش ، مبقولهاش ، مبقلهاش ، ما بقلهاش ، مابيقولهاش	*mAbqlhAI, mbqwlhAI, mbqlhAI, mA bqlhAI, mAbyqwlhAI*	≤ 10,000
مابقولهاش ، ما بقولهاش ، مبيقلهاش ، ما بيقلهاش	*mAbqwlhAI, mA bqwlhAI, mbyqlhAI, mA byqlhAI*	≤ 1,000
مبثلهاش ، مابيثولهاش ، ما بيثولهاش ، مابيؤلهاش	*mbθlhAI, mAbyθwlhAI, mA byθwlhAI, mAbyθlhAI*	≤ 100
ما بيؤلهاش ، مابثلهاش ، مبيثولهاش ، ما بيثلهاش ، مابثولهاش ، ما بثلهاش ، ما بؤلهاش ، مبثولهاش ، مبيؤلهاش ، مابؤلهاش ، مبؤلهاش	*mA byθlhAI, mAbθlhAI, mby θwlhAI, mA by θlhAI, mAbθwlhAI mA bθlhAI, mbθwlhAI, mAbyθlhAI, mAbθlhAI mbθlhAI*	≤ 10

Figure 1: Example of unambiguous correspondence during writing and reading tasks in Esperanto.





OTEANN: Estimating the Transparency of Orthographies with an Artificial Neural Network Xavier Marjou (2019)

* N. Habash, et al. (2018)i, “Unified guidelines and resources for Arabic dialect orthography,” in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, 2018.

41

Oral vs written language

- Orthographic opacity
  - phonemes!=letters!
- (lack of) orthographic standardization

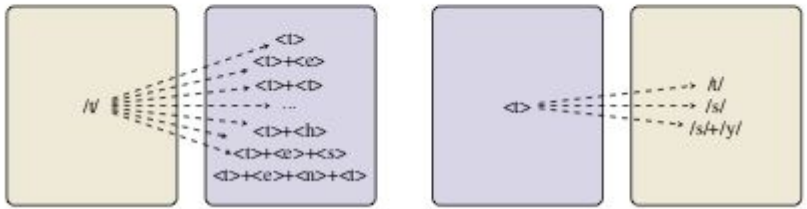
/mabiPulha:S/ *

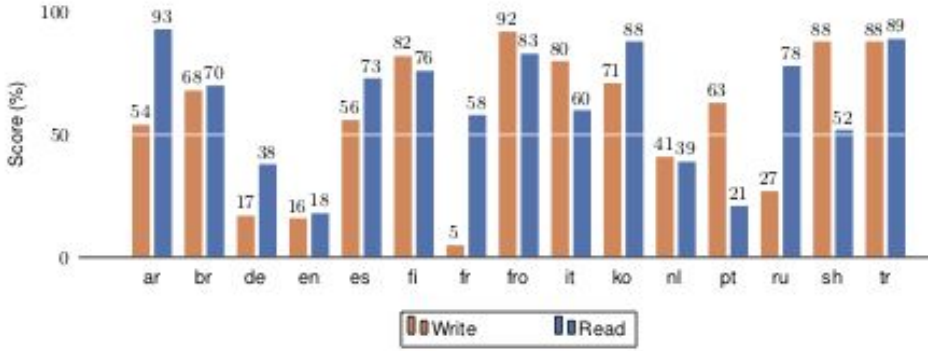
(‘he does not say it)’

Arabic Orthography	Arabic Transliteration	Frequency
مبيقولهاش	*mbyqwlhAI*	≈ 26,000
ما بيقولهاش	*mA byqwlhAI*	≈ 13,000
مابقلهاش ، مبقولهاش ، مبقلهاش ، ما بقلهاش ، مابيقولهاش	*mAbqlhAI, mbqwlhAI, mbqlhAI, mA bqlhAI, mAbyqwlhAI*	≤ 10,000
مابقولهاش ، ما بقولهاش ، مبيقلهاش ، ما بيقلهاش	*mAbqwlhAI, mA bqwlhAI, mbyqlhAI, mA byqlhAI*	≤ 1,000
مبشهاش ، مابيشولهاش ، ما بيشولهاش ، مابيؤلهاش	*mbyšlhAI, mAbyšwlhAI, mA byšwlhAI, mAbyvšlhAI*	≤ 100
ما بيؤلهاش ، مابشهاش ، مبيشولهاش ، ما بيشهاش ، مابشولهاش ، ما بشهاش ، ما بؤلهاش ، مبشولهاش ، مبيؤلهاش ، مابؤلهاش ، مبؤلهاش	*mA byšlhAI, mAbšlhAI, mby šwlhAI, mA by šlhAI, mAbyšwlhAI, mA bšlhAI, mA bvšlhAI, mbyšlhAI, mAbšlhAI, mbšlhAI*	≤ 10

- ‘Defective’ orthographies
  - Lack of stress; en. “record” (N) /’ɹɛkə͡d/ vs (V) /ɹɪˈkɔɹd/
  - abjad: only transcribe the consonants: arabic, hebrew
- Unwritten languages
  - Roughly 50% of the world languages may not use a written form (Ethnologue, 23th edition)

Figure 1: Example of unambiguous correspondence during writing and reading tasks in Esperanto.





OTEANN: Estimating the Transparency of Orthographies with an Artificial Neural Network Xavier Marjou (2019)

* N. Habash, et al. (2018)i, “Unified guidelines and resources for Arabic dialect orthography,” in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, 2018.

42

Oral vs written language

- Code switching
  - LM-level
    - Between sentences (inter-sentential switching)
      - “***Itula. Mama dah agak dah. Adiknidemamni.*** *Pity you. Your voice also different already.*” *[That’s why. I knew it. You are having a fever]* (Stapa& Khan, 2016).

Oral vs written language

- Code switching
  - LM-level
    - Between sentences (inter-sentential switching)
      - "*Itula. **Mama dah agak dah. Adiknidemamni.** Pity you. Your voice also different already.***"*** *[That's why. I knew it. You are having a fever]* (Stapa& Khan, 2016).
    - Within sentences (intra-sentential switching):
      - *"Sometimes I'll start a sentence in English **y termino en español**." ['and finish it in Spanish.']* (Cakrawarti, 2011)
      - *"This morning I **hantar** my baby **tu dekat** babysitter **tu lah**.*
        ['This morning I took my baby to the babysitter.']

Oral vs written language

- Code switching
  - LM-level
    - Between sentences (inter-sentential switching)
      - "***Itula. Mama dah agak dah. Adiknidemamni.*** *Pity you. Your voice also different already.***"** *[That's why. I knew it. You are having a fever]* (Stapa& Khan, 2016).
    - Within sentences (intra-sentential switching):
      - *"Sometimes I'll start a sentence in English **y termino en español**." ['and finish it in Spanish.']* (Cakrawarti, 2011)
      - *"This morning I **hantar** my baby **tu dekat** babysitter **tu lah**.*
        ['This morning I took my baby to the babysitter.']
  - Lexicon-level
    - Within word
      - *"**But ma**-day-s **a-no a-ya ha-ndi-si ku-mu-on-a**. ["But these days I don't see him much."]* (the prefix "ma" means plural (Winford 2003)

Oral vs written language

- Code switching
  - LM-level
    - Between sentences (inter-sentential switching)
      - "***Itula. Mama dah agak dah. Adiknidemamni.*** *Pity you. Your voice also different already.***"** *[That's why. I knew it. You are having a fever]* (Stapa& Khan, 2016).
    - Within sentences (intra-sentential switching):
      - *"Sometimes I'll start a sentence in English **y termino en español**." ['and finish it in Spanish.']* (Cakrawarti, 2011)
      - *"This morning I **hantar** my baby **tu dekat** babysitter **tu lah**.*
        ['This morning I took my baby to the babysitter.']
  - Lexicon-level
    - Within word
      - *"**But ma**-day-s **a-no a-ya ha-ndi-si ku-mu-on-a**. ["But these days I don't see him much."]* (the prefix "ma" means plural (Winford 2003)
- Lexical borrowings
  - words are borrowed from the source and adapted to the target AM
    - *plot* (en) → *plotter* (fr)
    - *Chrismas* (en) → /kulisumasu/ (jp)
- Syntactic borrowing
  - mixing of the LMs; eg, in spanish nouns used as adjectives as in English (*coche patrulla [patrol car], hora punta [rush hour], fecha límite [deadline]*)

"on-line" processes

Historical processes

46

Oral vs written language

- Variations in units definition
  - phonemes, morphemes, words: tricky definitions
  - eg, with **words**, at least 4 notions must be distinguished:
    - The **prosodic word** (can be separated by pause): eg. *“the dog”* :
    - The typographic word, or **token** (sequence of letters without separators)**:**
      - *“This sentence has 8 tokens : really ?”* (separators: ‘ ’)
      - But in some languages, no separators: “我的漢語說得不太好。” “ฉันฟังไม่เข้าใจ”
    - The morphosyntactic word, or **wordform** (syntactic atomic unit);
      - **amalgams**: en. *won’t =(will) (not)*, fr. *du = (de) (le)*, sp. *dámelo =(da) (me) (lo)*
      - **compounds** fr. “*(au fur et à mesure)”*, en. “*(all of a sudden)”*
    - The **semantic word** (sequence of word forms with non compositional meaning)
      - Eg. fr “*pomme de terre*”, “*red herring*”
      - See also, named entities: *“Los Angeles”, “Apple Inc”*.

Oral vs written language

- Variations in units definition
  - phonemes, morphemes, words: tricky definitions
  - eg, with **words**, at least 4 notions must be distinguished:
    - The **prosodic word** (can be separated by pause): eg. *“the dog”* :
    - The typographic word, or **token** (sequence of letters without separators)**:**
      - *“This sentence has 8 tokens: really?”* (separators: ‘ ’)
      - But in some languages, no separators: “我的漢語說得不太好。” “ฉันฟังไม่เข้าใจ”
    - The morphosyntactic word, or **wordform** (syntactic atomic unit);
      - **amalgams**: en. *won’t =(will) (not)*, fr. *du = (de) (le)*, sp. *dámelo =(da) (me) (lo)*
      - **compounds** fr. “*(au fur et à mesure)”*, en. “*(all of a sudden)”*
    - The **semantic word** (sequence of word forms with non compositional meaning)
      - Eg. fr “*pomme de terre*”, “*red herring*”
      - See also, named entities: *“Los Angeles”, “Apple Inc”*.
  - eg, with **sentences**:
    - Typography is sometimes misleading: *“Best. Movie. Ever.”*
    - Nested structures: *“‘Give me the box,’ John said.”*

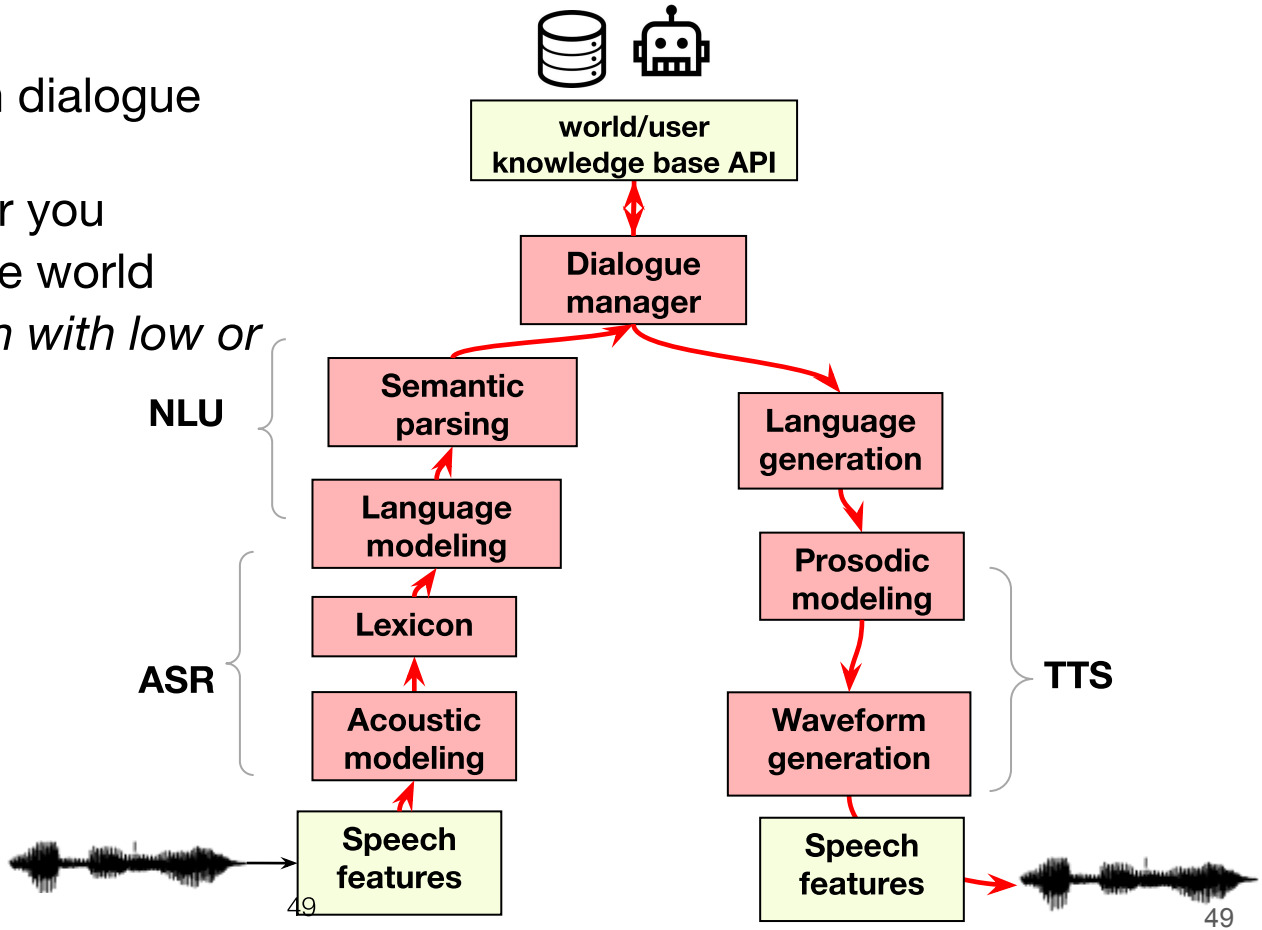
→ typography is often used as a proxy for linguistic units; but it varies across languages and some times not consistent or not available (unwritten languages)

48

So, how to do this?

Overall objective

- Construct a whole spoken dialogue system
- That does useful things for you
- For all the languages of the world
- *Across variations and even with low or no resources!*



49

Thanks

Special Thanks to Benoit Sagot for slides from the MVA course ‘algorithms for speech and language processing’ on diversity, variability and sparsity