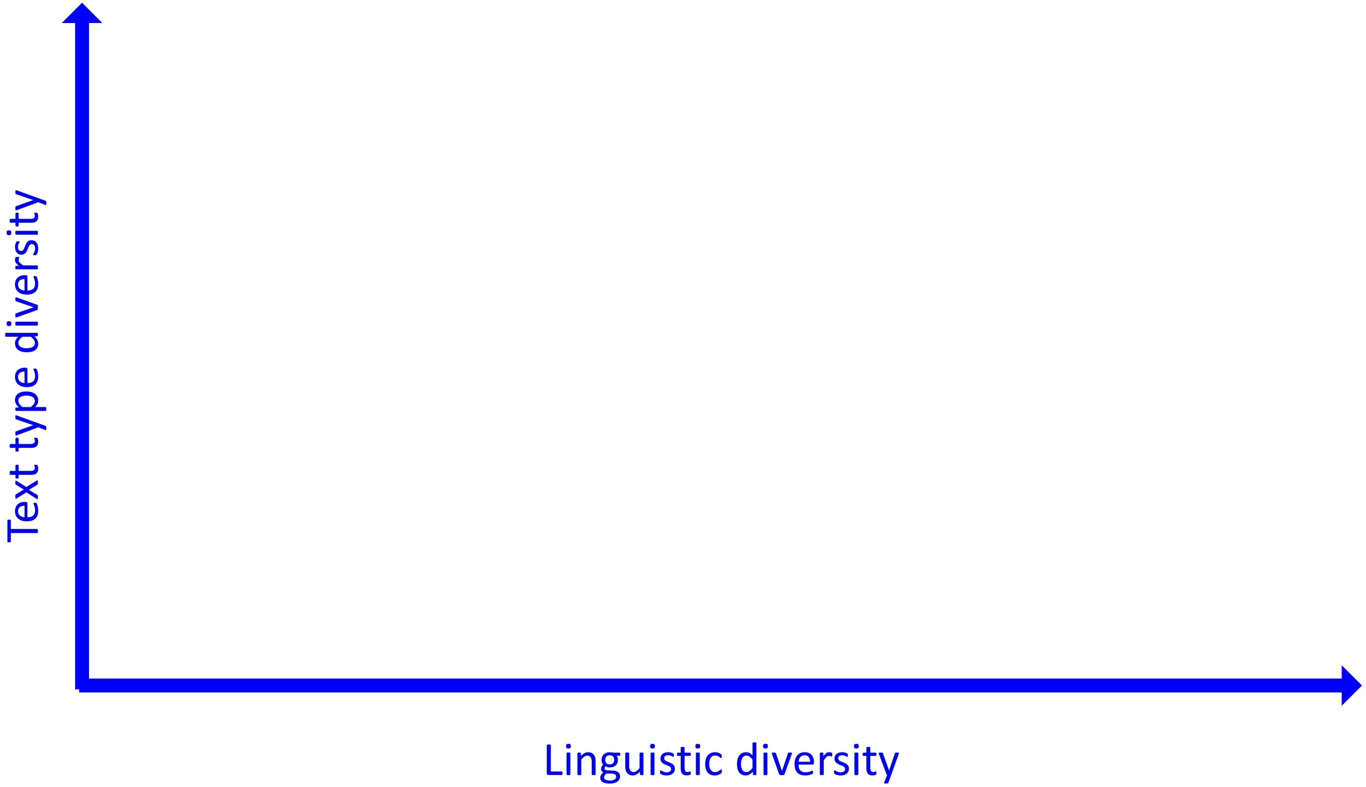
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

Typology in the age of corpora:
Applications and challenges

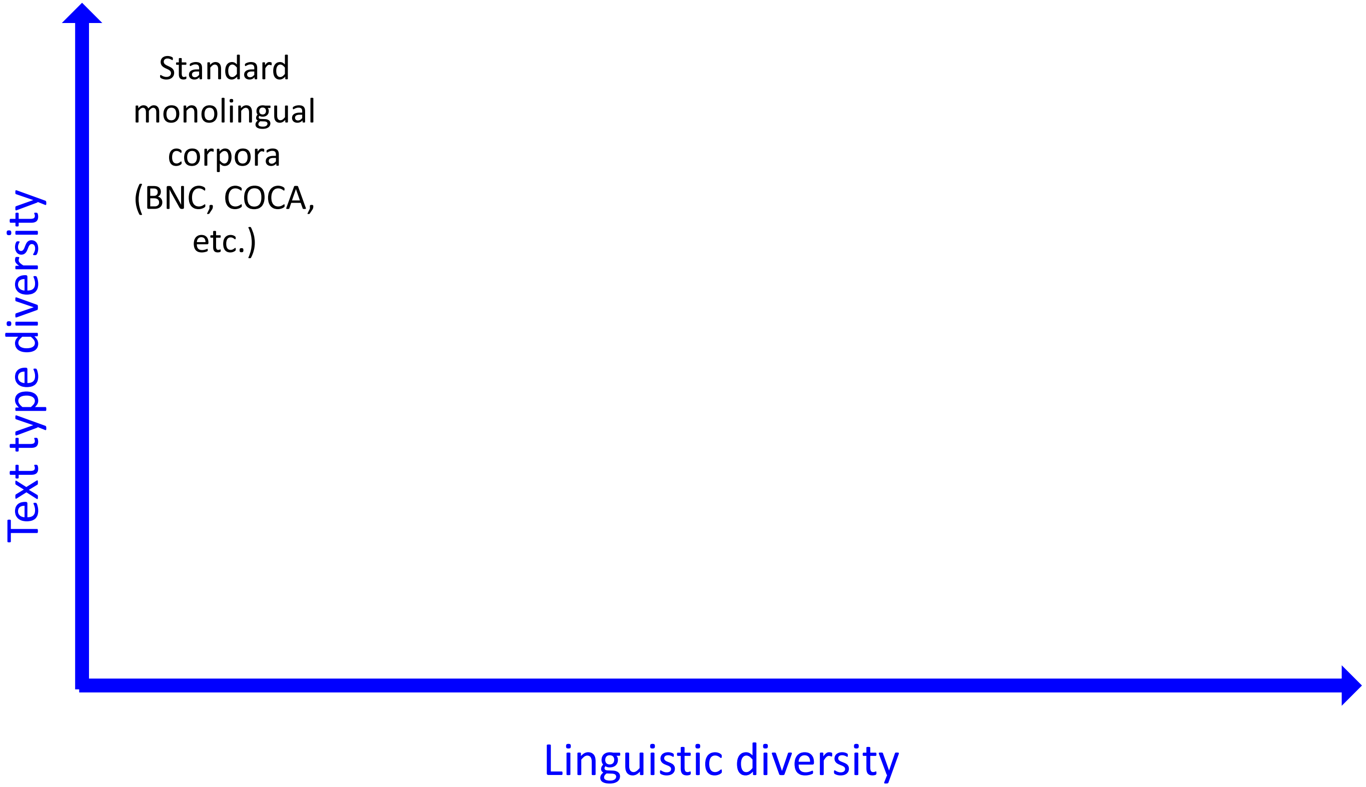
NATALIA LEVSHINA

What kind of corpora are there?

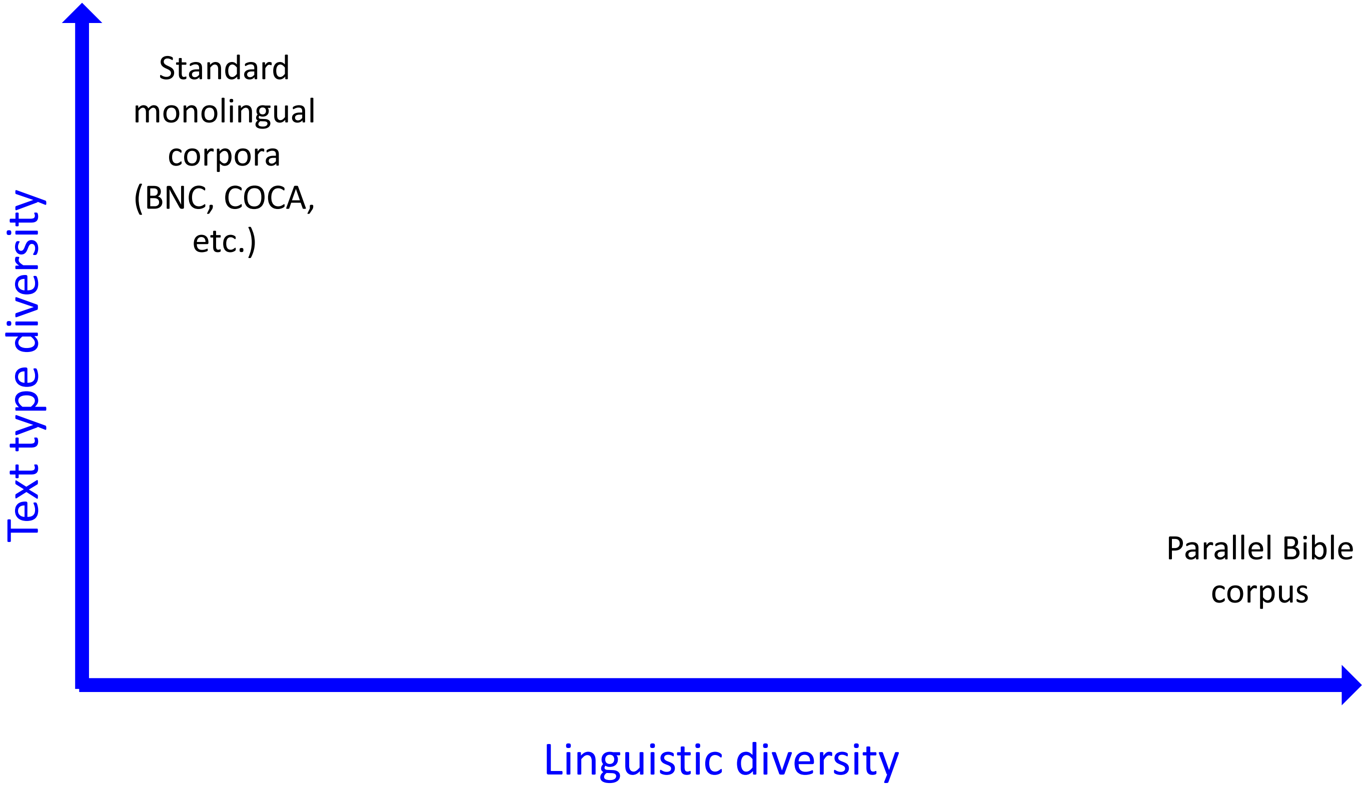
Simple 2D typology of corpora



Simple 2D typology of corpora



Simple 2D typology of corpora



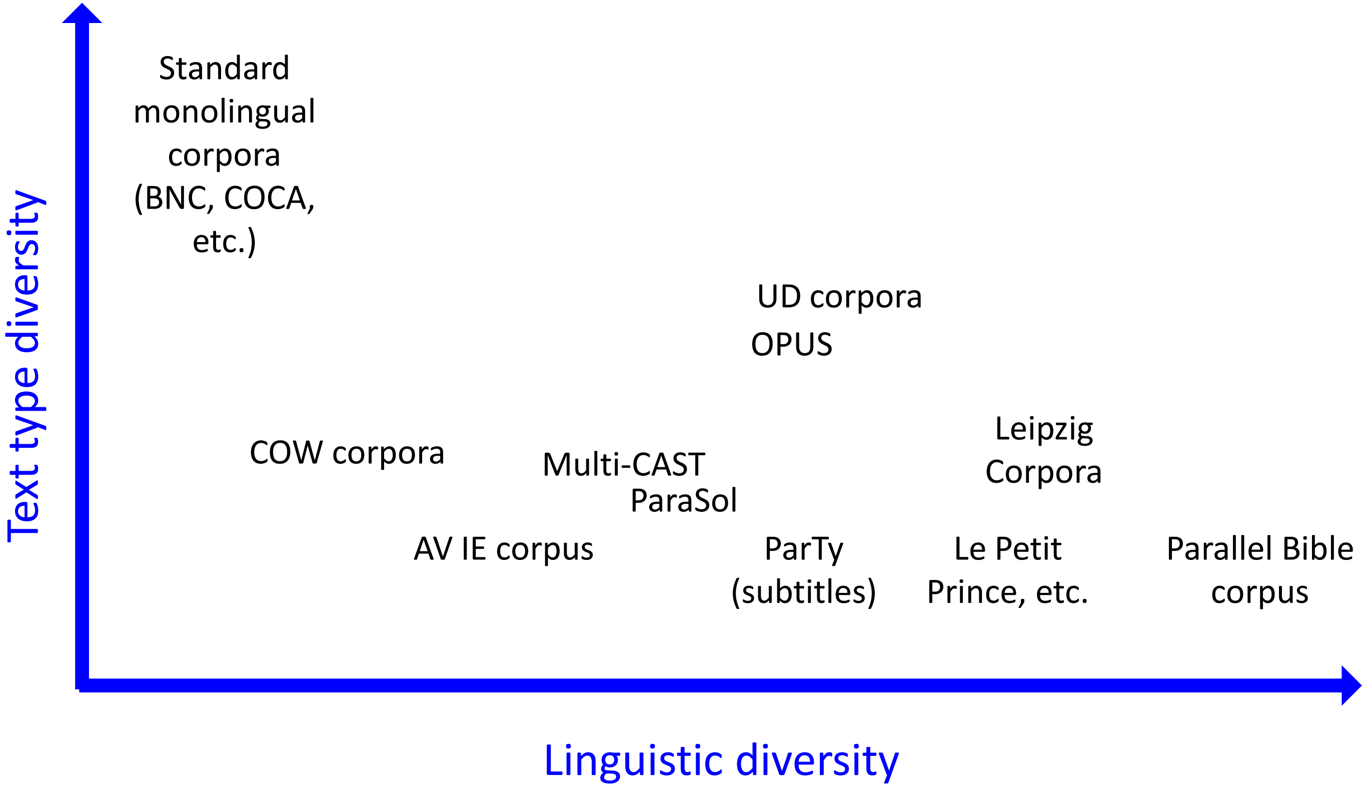
Standard monolingual corpora (BNC, COCA, etc.)

Parallel Bible corpus

Text type diversity

Linguistic diversity

Simple 2D typology of corpora



Text type diversity

Standard monolingual corpora (BNC, COCA, etc.)

UD corpora

OPUS

COW corpora

Multi-CAST

ParaSol

Leipzig Corpora

AV IE corpus

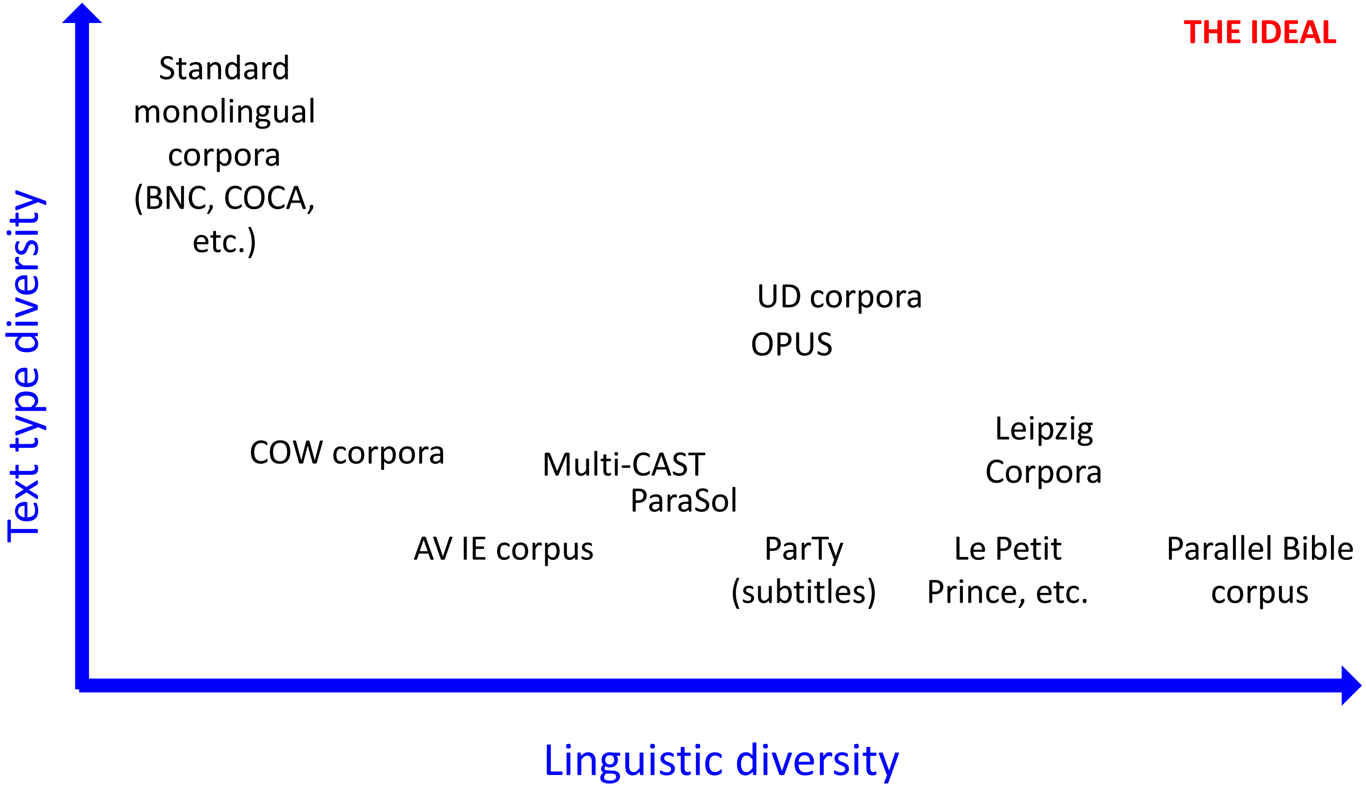
ParTy (subtitles)

Le Petit Prince, etc.

Parallel Bible corpus

Linguistic diversity

Simple 2D typology of corpora



THE IDEAL

Standard monolingual corpora (BNC, COCA, etc.)

UD corpora

OPUS

Leipzig Corpora

COW corpora

Multi-CAST

ParaSol

AV IE corpus

ParTy (subtitles)

Le Petit Prince, etc.

Parallel Bible corpus

Text type diversity

Linguistic diversity

How can corpora help us compare languages?

- Classification of languages based on aggregate indices derived from corpora
- Comparison of semantic and pragmatic functions of related constructions
- Testing and explanation of cross-linguistic generalizations

How can corpora help us compare languages?

- Classification of languages based on aggregate indices derived from corpora
- Comparison of semantic and pragmatic functions of related constructions
- Testing and explanation of cross-linguistic generalizations

Indices in previous research

- Analyticity/syntheticity indices (e.g. Greenberg 1960, Szmrecsanyi 2009)
- Kolmogorov complexity (e.g. Juola 1998)
- Head-dependent order (e.g. Liu 2010)
- and many others…

Word order entropy

Data

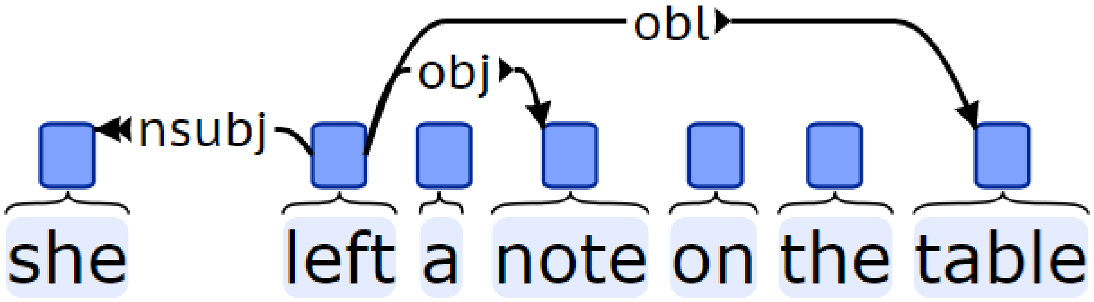
- The Universal Dependencies corpora
- The frequencies of so-called heads + dependent elements in different order:
  - head + dependent
  - dependent + head

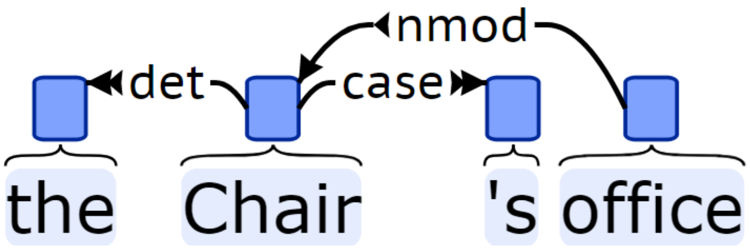
http://universaldependencies.org/, Nivre et al. (2017)

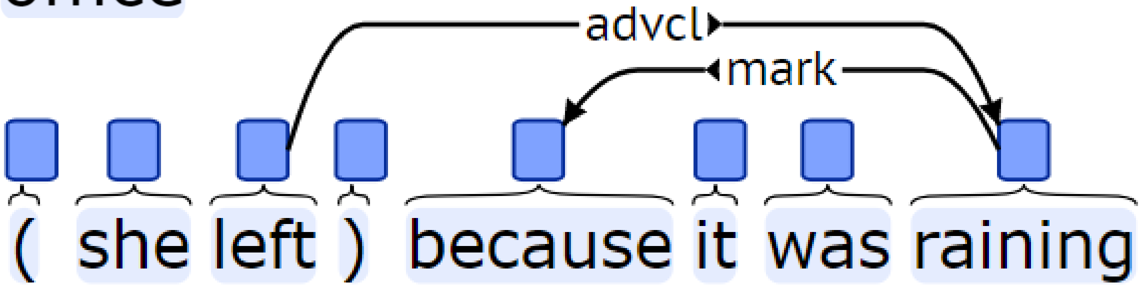
Dependencies

- Nsubj_Noun + Verb
- Nsubj_Pron + Verb
- Obj_Noun + Verb
- Obj_Pron + Verb
- Obl_Noun + Verb
- Obl_Pron + Verb
- Nmod_Noun + Noun
- Nmod_Pron + Noun
- Nummod + Noun
- Amod + Noun
- Advmod + Verb
- Advmod + Adj
- Det + Noun
- Case + Noun
- Aux + Verb
- Cop + NomPred
- Csubj + Main
- Ccomp + Main
- Acl + Noun
- Advcl + Main
- Subordinator + Ccomp
- Subordinator + AdvCl

Examples







http://universaldependencies.org/guidelines.html

Shannon’s entropy

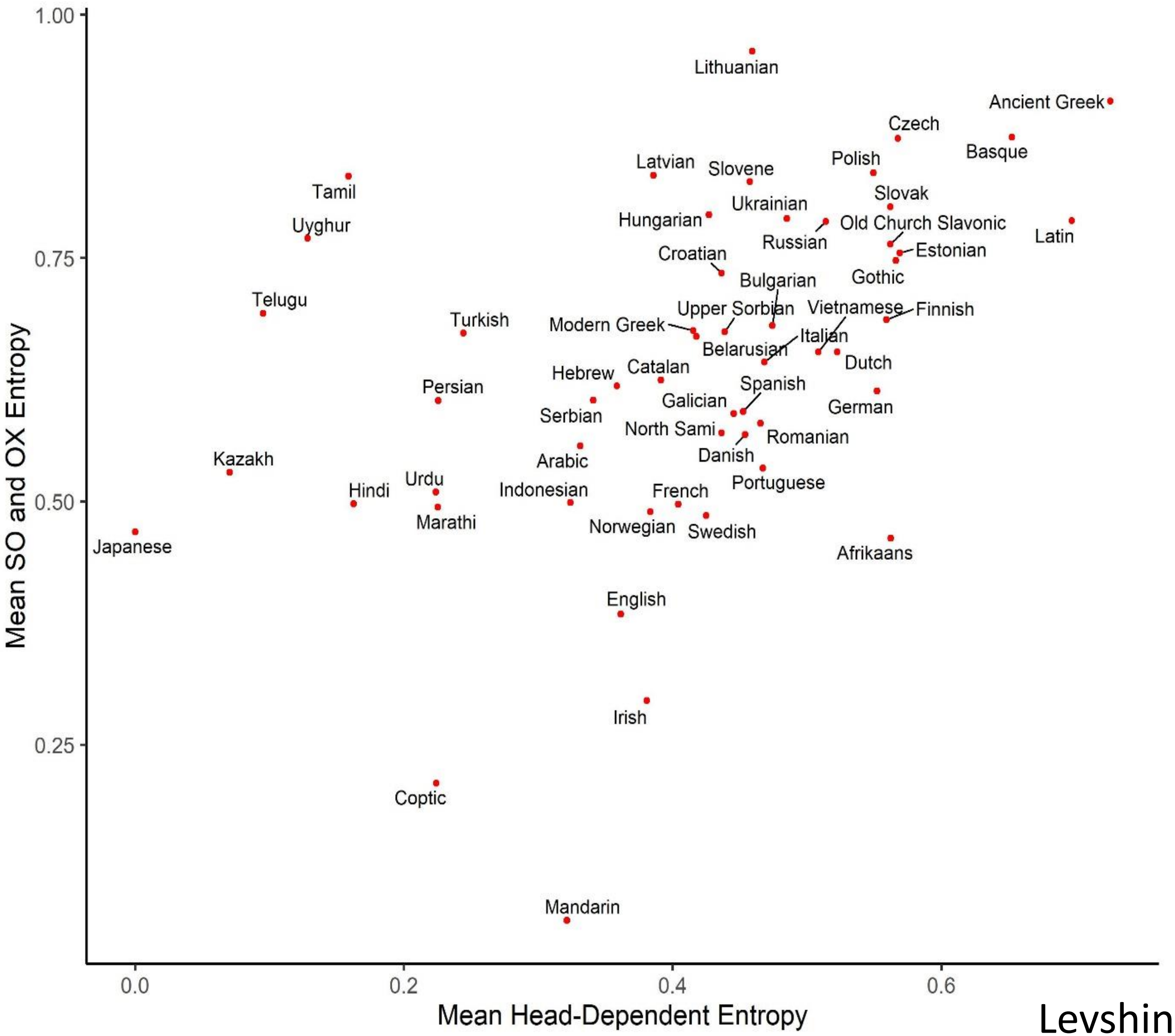
$$H(X) = -\sum_{i=1}^{2} P(x_i)\, log_2\, P(x_i)$$

- If a language has 50% object + verb, 50% verb + object:

  H = 1 (maximal)

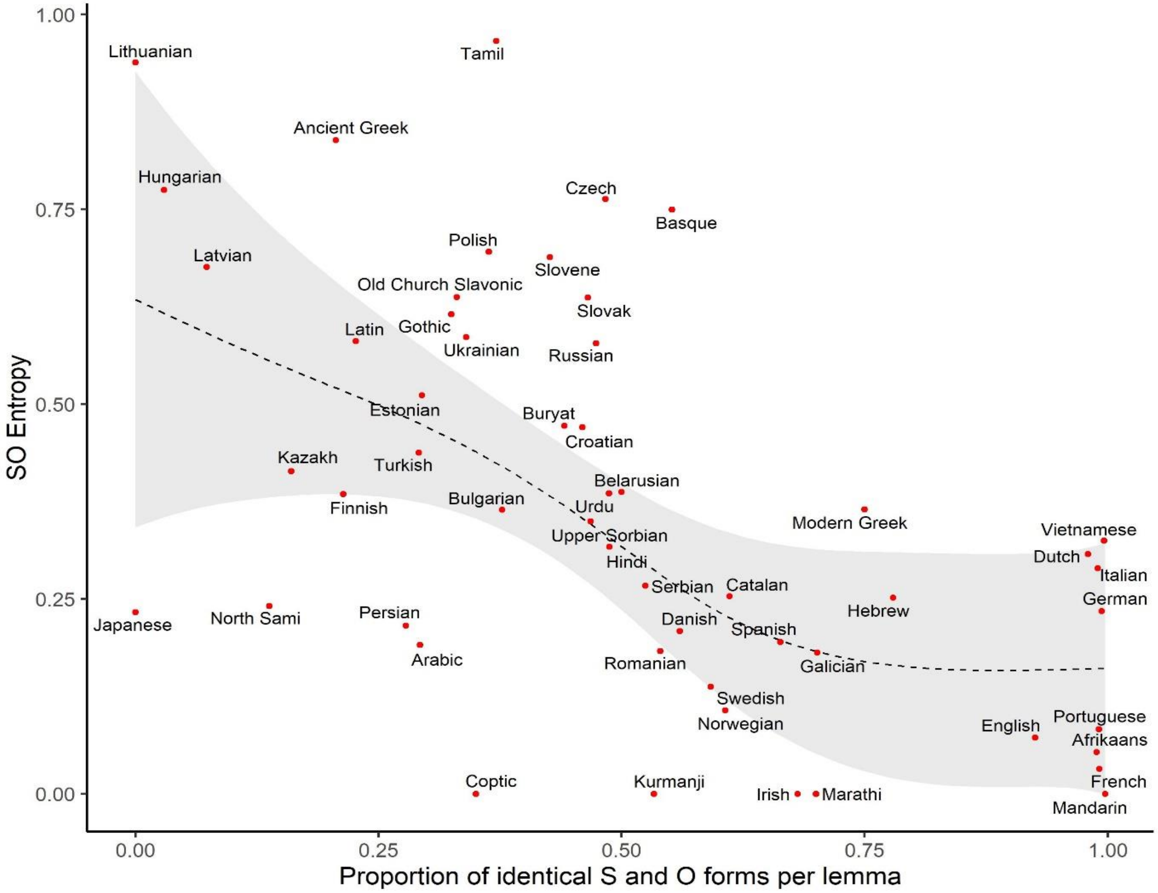
- If a language has 100% object + verb, 0% verb + object, OR if a language has 0% object + verb, 100% verb + object:

  H = 0 (minimal)



Levshina (2019)

SO confusability vs. entropy



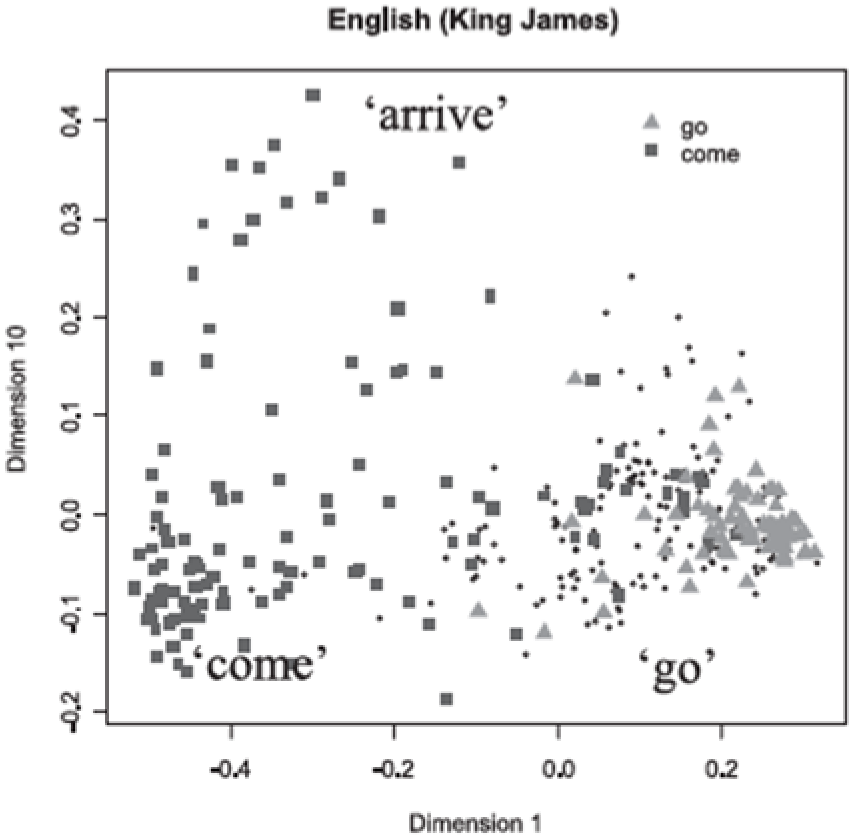
GAM:
Deviance = 83%, adj. R^2 = 0.74

Levshina (2019)

How can corpora help us compare languages?

- Classification of languages based on aggregate indices derived from corpora
- Comparison of semantic and pragmatic functions of related constructions
- Testing and explanation of cross-linguistic generalizations

Corpus-based semantic maps



Motion events (Wälchli & Cysouw 2012)

Token-based MDS maps

1. Collect the data (fictitious example)

	Lang1	Lang2	Lang3	Lang4	Lang5
Situation 1	Bla	Boo	Aha	Ti	Na
Situation 2	Bla	Boo	Aha	Ta	Ne
Situation 3	Bli	Boo	Oho	Ti	Ni

Token-based MDS maps

2. Compute the distances between the situations (rows)

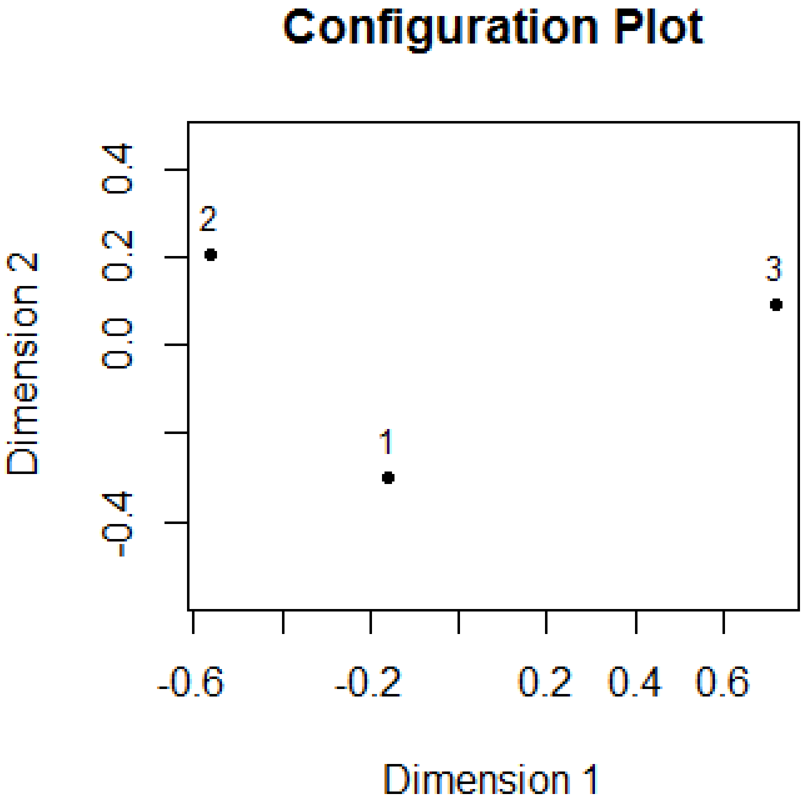
	Lang1	Lang2	Lang3	Lang4	Lang5
Situation 1	Bla	Boo	Aha	Ti	Na
Situation 2	Bla	Boo	Aha	Ta	Ne
Situation 3	Bli	Boo	Oho	Ti	Ni

Overlap 1,2 = 3/5 = 0.6
Overlap 1,3 = 2/5 = 0.4
Overlap 2,3 = 1/5 = 0.2

Distance = 1 – overlap

Token-based MDS maps

3. Perform MDS (package smacof)



Interpretation of MDS distances

- The closer two points (i.e. motion events or causative situations), the more frequently they are expressed by the same constructions across the languages in the doculects.

Analytic causatives

Examples of Analytic Causatives

- Don’t make me cry.
- Let my people go.
- You’re forcing me to be the voice of reason.
- 6 careers that allow to you to travel around the world.

Parallel corpus of film subtitles

AUDREY TAUTOU
MATHIEU KASSOVITZ
Le Fabuleux Destin d'Amélie Poulain
un film de JEAN-PIERRE JEUNET

https://github.com/levshina/ParTy-1.0

Dataset

- Translations in 18 European languages (15 Indo-European and 3 Finno-Ugric languages)
- Automatically aligned
- All ACs extracted manually from each doculect.
- 392 contexts with at least one language having an AC

For more details, see Levshina 2015

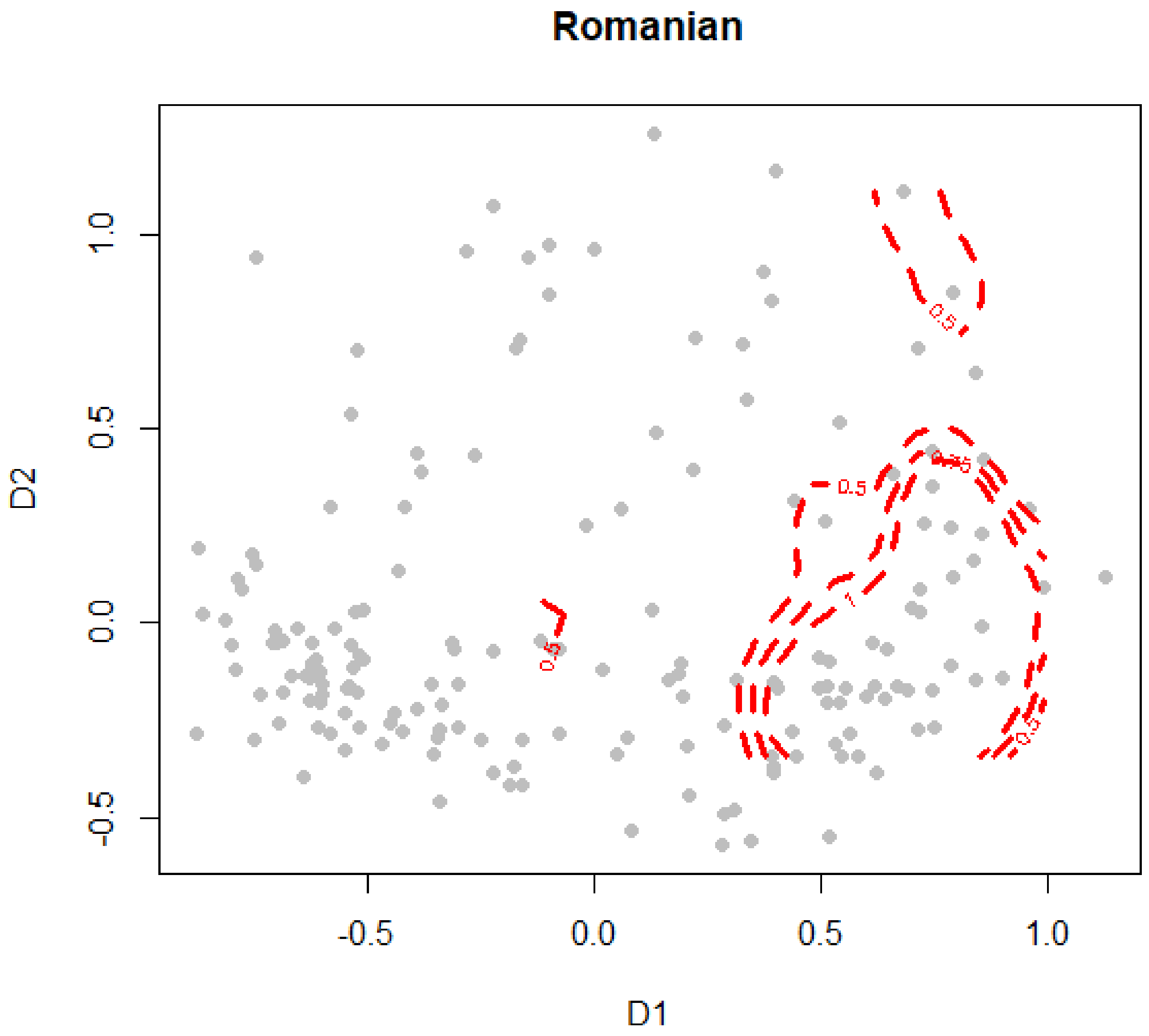
Method

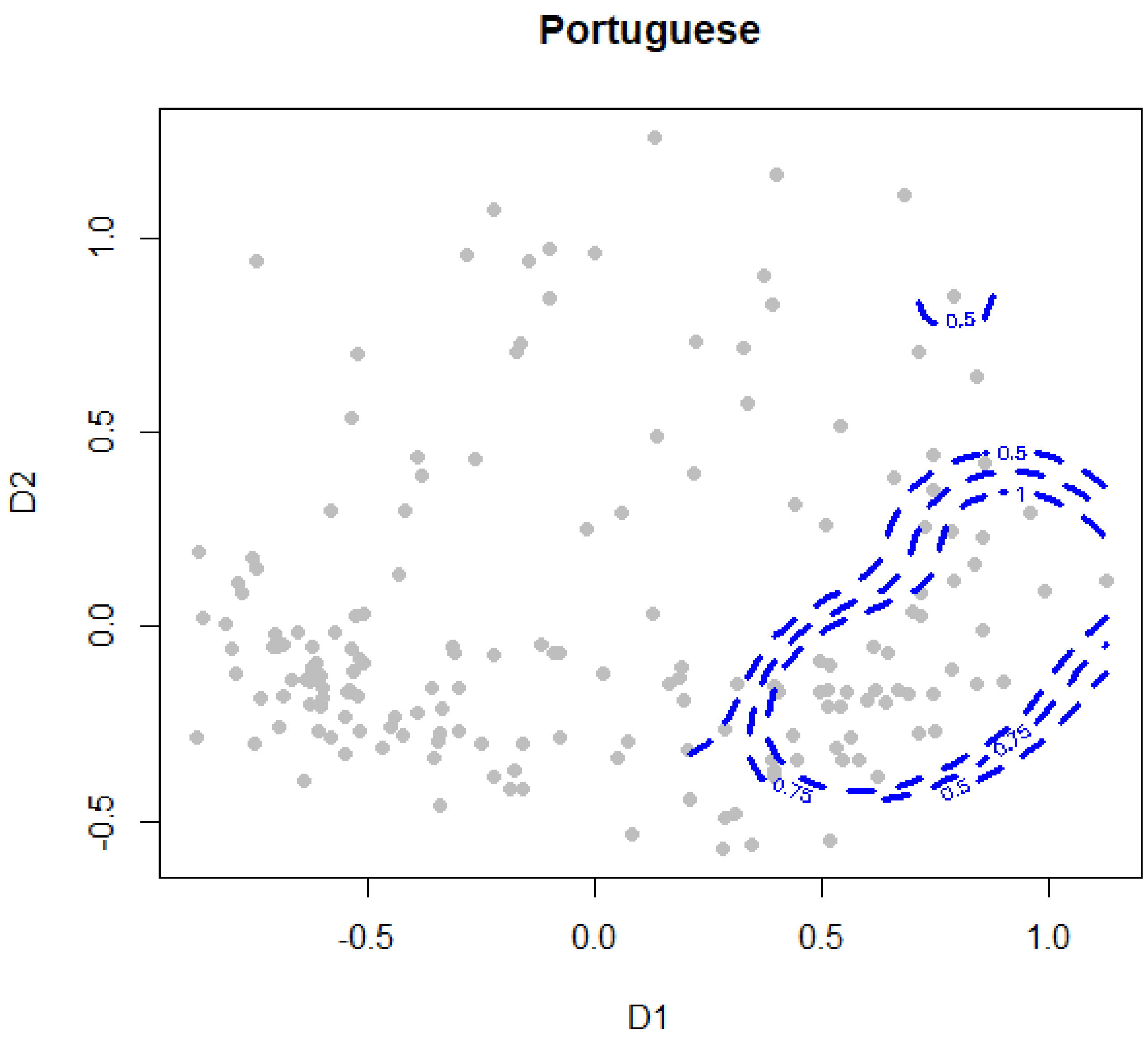
- Multidimensional Scaling with smacof
- An interactive plot with googleVis: http://www.natalialevshina.com/presentations.html

For more details, see Levshina 2015

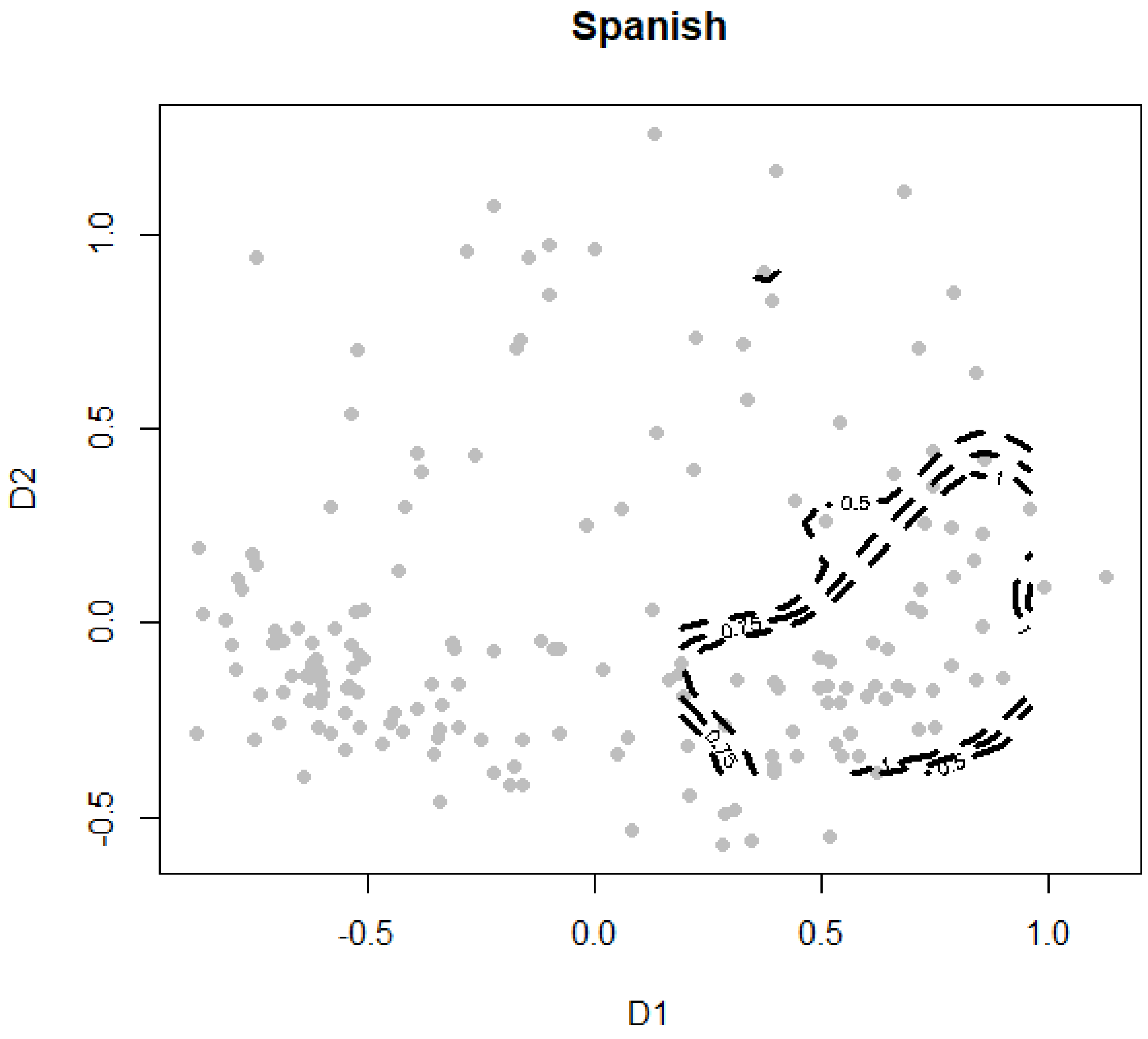
Zooming in on Romance ACs

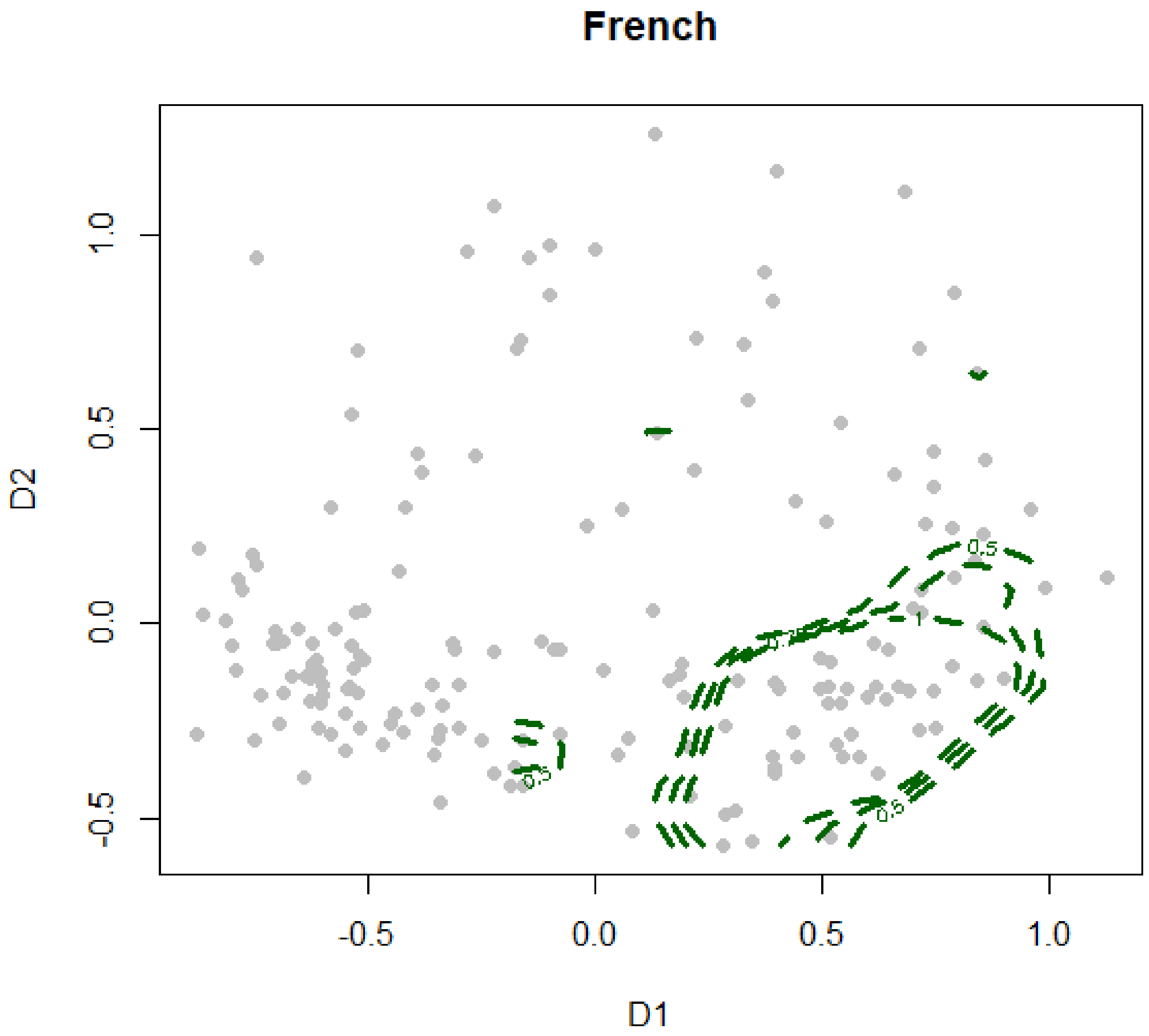
- ita: *fare* + Vinf
- fra: *faire* + Vinf
- spa: *hacer* + (NP) + Vinf
- por: *fazer* + (NP) + Vinf/Vinf_inflected
- rom: *a face* + *să* + Vsubj


Romanian
D2
D1
1.0
0.5
0.0
-0.5
-0.5
0.0
0.5
1.0
0.5
0.75
0.5
1
0.5
0.5

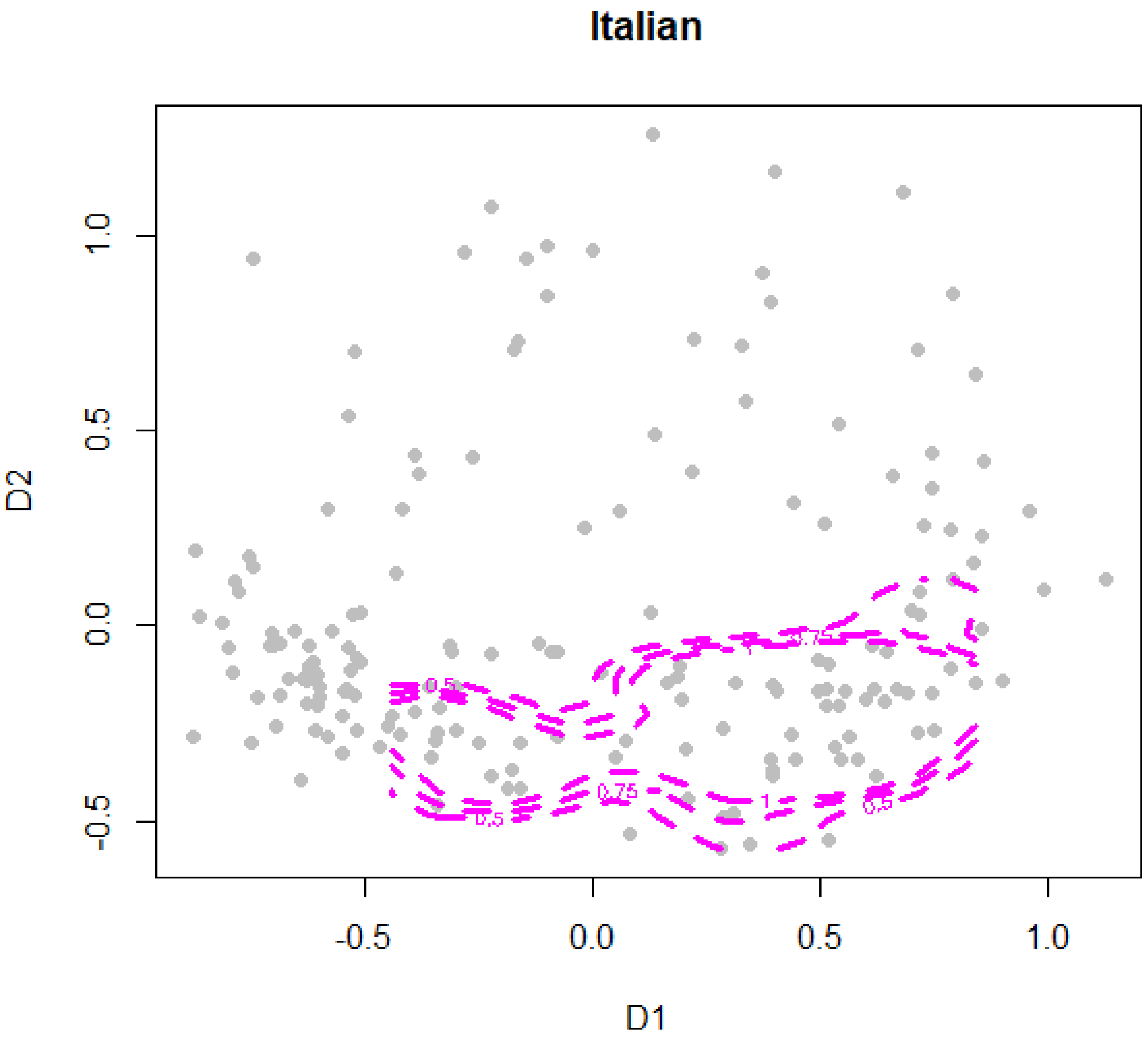


Portuguese
D2
D1
-0.5
0.0
0.5
1.0
-0.5
0.0
0.5
1.0
0.5
0.5
1
0.75
0.75
0.5


Spanish
D1
D2
-0.5
0.0
0.5
1.0
-0.5
0.0
0.5
1.0
0.5
1
0.75
0.75
1
0.5



French


Italian
D2
1.0
0.5
0.0
-0.5
-0.5
0.0
0.5
1.0
D1
0.5
0.75
1
0.75
1
0.5
0.5

Examples

- French, Amélie

  Amandine Poulain aime: (…) Faire briller le parquet avec des patins…

  Amandine Poulain likes: (…) polishing the parquet with slippers…

- Italian, Avatar

Stronzate,	*fammi*	*vedere!*
Bullshit	make.me	see

Bullshit, let me see that!

How can corpora help us compare languages?

- Classification of languages based on aggregate indices derived from corpora
- Comparison of semantic and pragmatic functions of related constructions
- Testing and explanation of cross-linguistic generalizations

What kind of universals?

- Categorical vs. continuous data per language
- Implicational (one-way) vs. correlational (two-way)

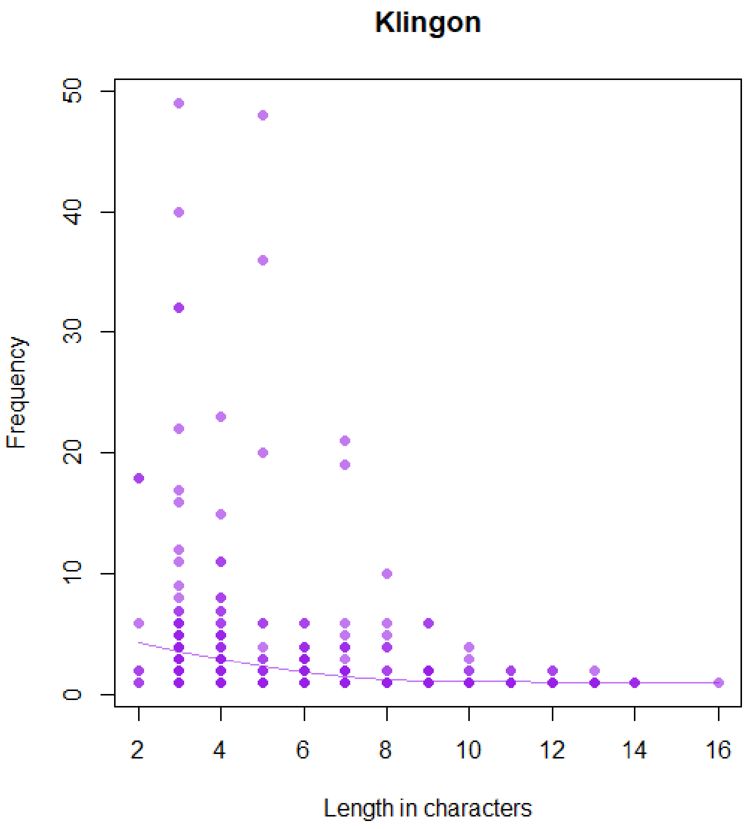
What kind of universals?

- Categorical vs. continuous data per language
- Implicational (one-way) vs. correlational (two-way)

Zipf’s law of abbreviation

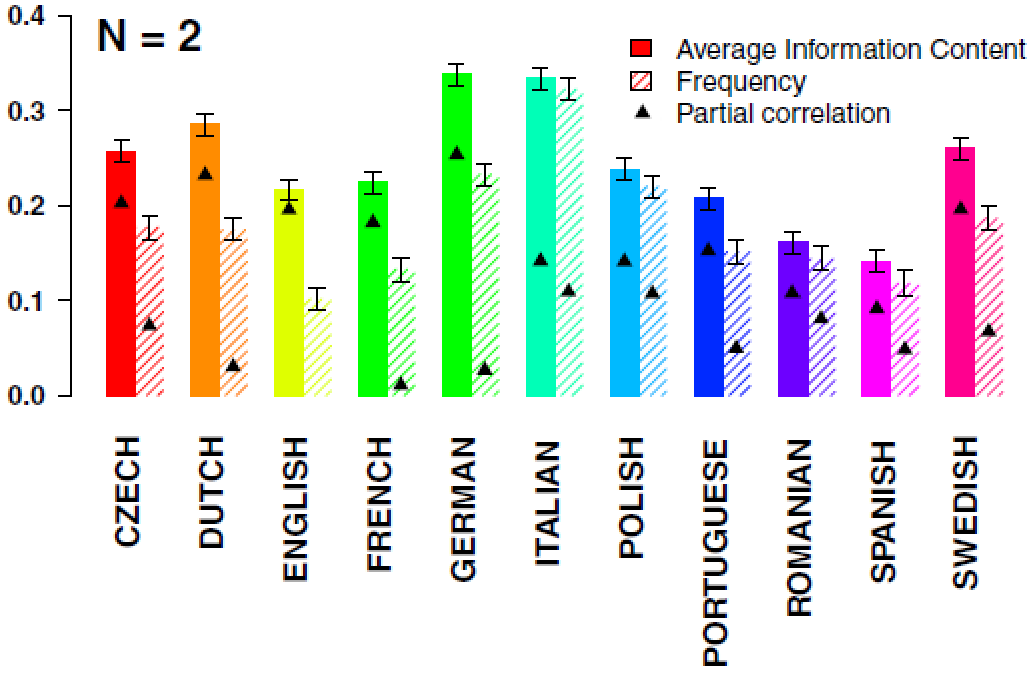
Zipf’s law of abbreviation

- Frequent words tend to be shorter (Zipf 1935)
- Benz & Ferrer-i-Cancho (2016):
    - Almost 1K languages
    - Negative correlations between length and frequency



Based on a text on http://nuqbopbom.blogspot.com/

Conditional probability vs. frequency



Piantadosi et al. 2011

Gibson et al. (2019) about Zipf

- “… Zipf worked before information theory provided a mathematical framework for understanding optimal codes. In an optimal code, the length of a signal will depend on its probability in context, not its overall frequency.”

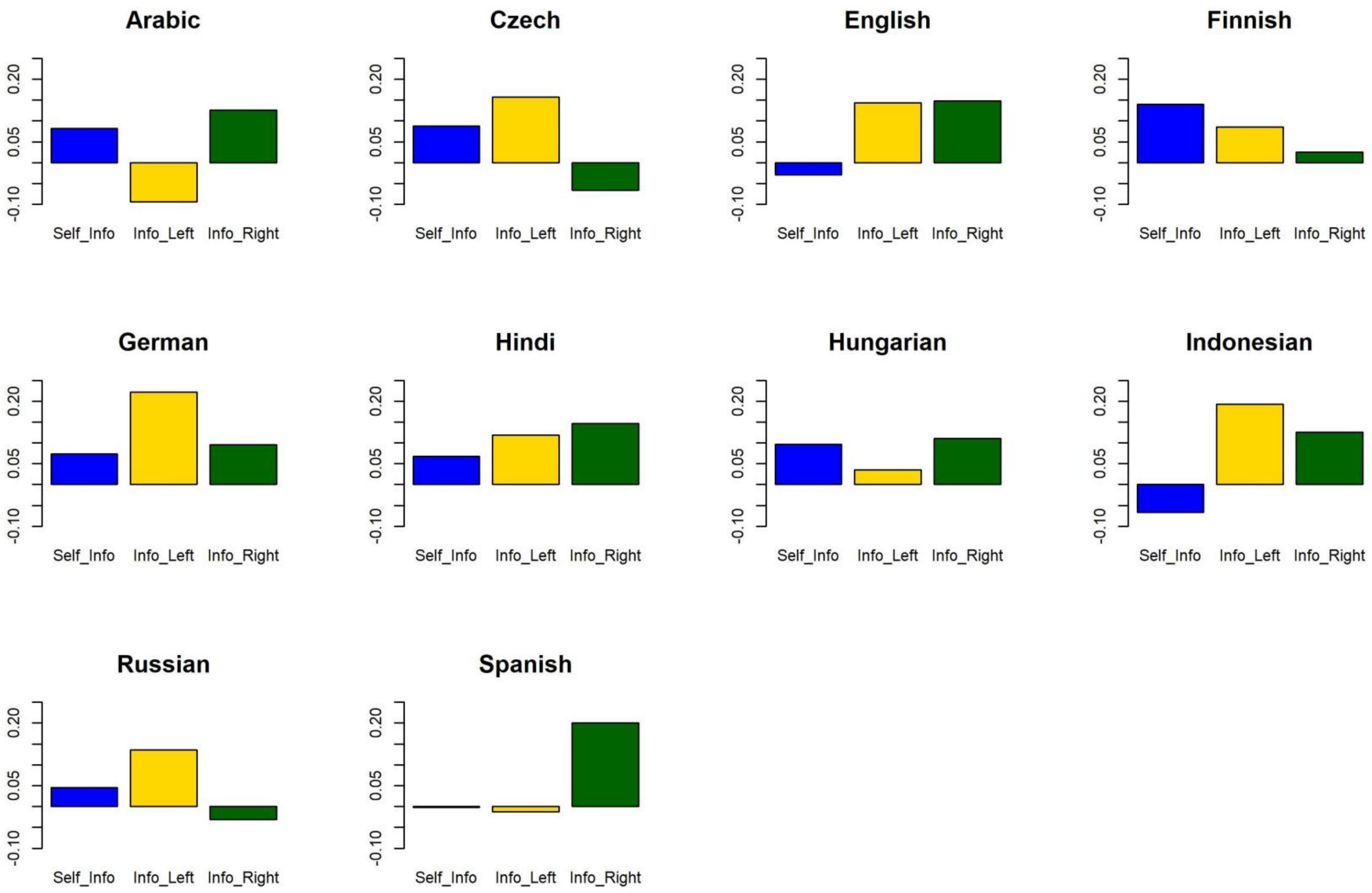
Data

- Leipzig Corpora Collection (Goldhahn et al. 2012), online news/newscrawler
  - http://wortschatz.uni-leipzig.de/en/download/
  - Large, free, typologically and genealogically diverse
- 10 languages: Arabic, Czech, English, Finnish, German, Hindi, Hungarian, Indonesian, Russian, Spanish
- Different corpus sizes (1M tokens, 10M tokens, 30M tokens)
- A random sample of 4,000 tokens with frequency > 20, only alphabetic characters
- Length of words in utf-8 characters
- Frequencies of unigrams (tokens), bigrams (1 word on the left, 1 word on the right)

Variables and method

- Self-information: $I = -\log_2(P_w)$
- Average Information Content given 1 token on the left
- Average Information Content given 1 token on the right
- Partial correlations with length (Kendall’s tau and Spearman’s rho), R package ppcor

Partial Kendall *tau*, 30M corpus samples



Explaining generalizations

- Form-frequency correspondences, for instance:
    - causative alternations in Haspelmath et al. (2014)
    - singulatives and pluratives in Haspelmath & Karjus (2017)

Differential case marking of A and P

Differential case marking of A

e.g. Quiang (Sino-Tibetan, LaPolla & Huang 2003: 79–80):

A. Animate A: unmarked

The:	*qa*	*dʑete.*
3SG	1SG	hit

‘He is hitting me.’

B. Inanimate A: marked

*Moʁu-**wu***	*qa*	*da-tuə-ʐ.*
wind-AGT	1SG	DIR-fall.over-CAUS

‘The wind knocked me over.’

Differential case marking of P

e.g. Spanish

a. Inanimate P: unmarked

Vi	*una*	*mesa.*
saw.1SG	INDEF	table

‘I saw a table.’

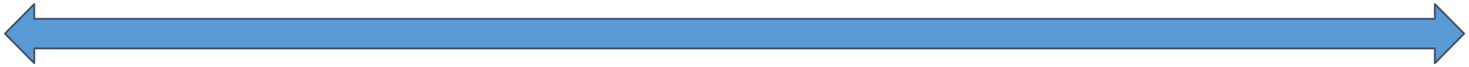
b. Animate P: marked

Vi	***a***	*una*	*mujer.*
saw.1SG	**OBJ**	INDEF	woman

‘I saw a woman.’

Referential scales

- Human > Animal > Inanimate
- 1 and 2 Person > 3 Person
- Pronoun > Noun
- Definite > (Indefinite) Specific > Non-specific
- Given > New



UNMARKED A MARKED A

MARKED P UNMARKED P

(Silverstein 1976, Bossong 1991: 159, Comrie 1986: 94, Croft 2003: 132)

Scale effects: Some issues

- Asymmetry in splits between A and P (Malchukov 2008, de Hoop & Malchukov 2008, Fauconnier & Verstraete 2014), e.g. more evidence of DOM than of DAM, different scales are relevant

Scale effects: Some issues

- Asymmetry in splits between A and P (Malchukov 2008, de Hoop & Malchukov 2008, Fauconnier & Verstraete 2014), e.g. more evidence of DOM than of DAM, different scales are relevant
- Debates about evaluating the cross-linguistic evidence: Cf. Filimonova (2005), Bickel et al. (2015) vs. Schmidtke-Bode & Levshina (2018)

Languages

- 5 typologically diverse languages: English, Lao (Tai-Kadai), N|uu/N‖ng (Tuu), Russian and Ruuli (Bantu).

- It is not important whether the languages have DAM/DOM or not. Since the scale effects are claimed to be universal, we assume that the associations between the roles and referential features are very similar across the languages.

Dialogical corpora

- English: Santa Barbara Corpus of Spoken American English (Du Bois et al. 2005), 8 conversations, 201 transitives
- Russian: 4 conversations from Zemskaja’s collection (1978), 202 transitives
- Lao: 5 conversations from Enfield (2007), 101 transitives
- Ruuli: 5 conversations from A. Witzlack-Makarevich et al. (2017–) corpus, 222 transitives
- N‖ng: 5 conversations from Güldemann et al. (2012), 225 transitives

Levshina & Witzlack-Makarevish, In prep.

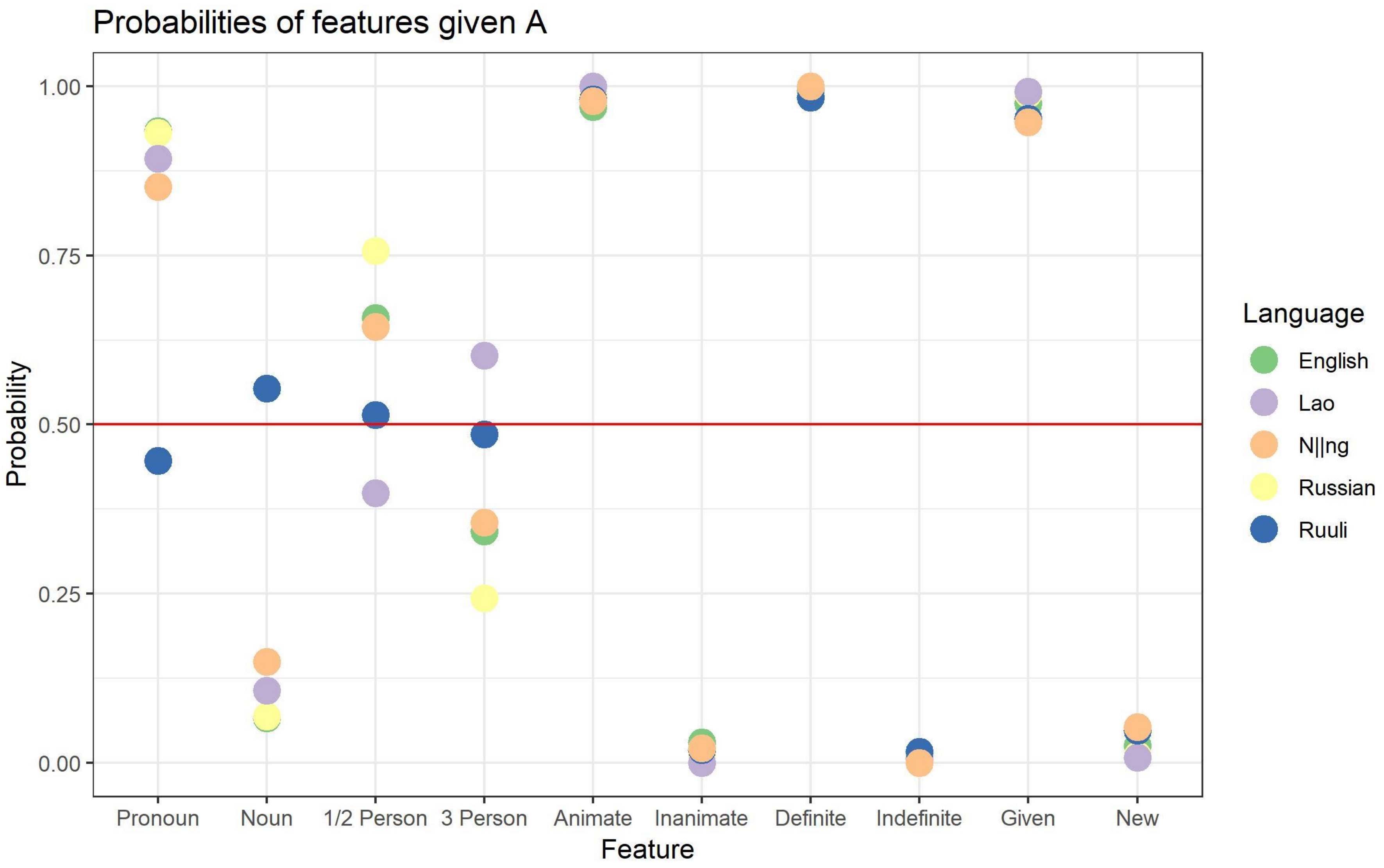
Question

- Which probabilities are relevant for emergence of differential case marking?
  - P (Feature|Role) – markedness, typicality
  - P (Role|Feature) – efficiency, economy

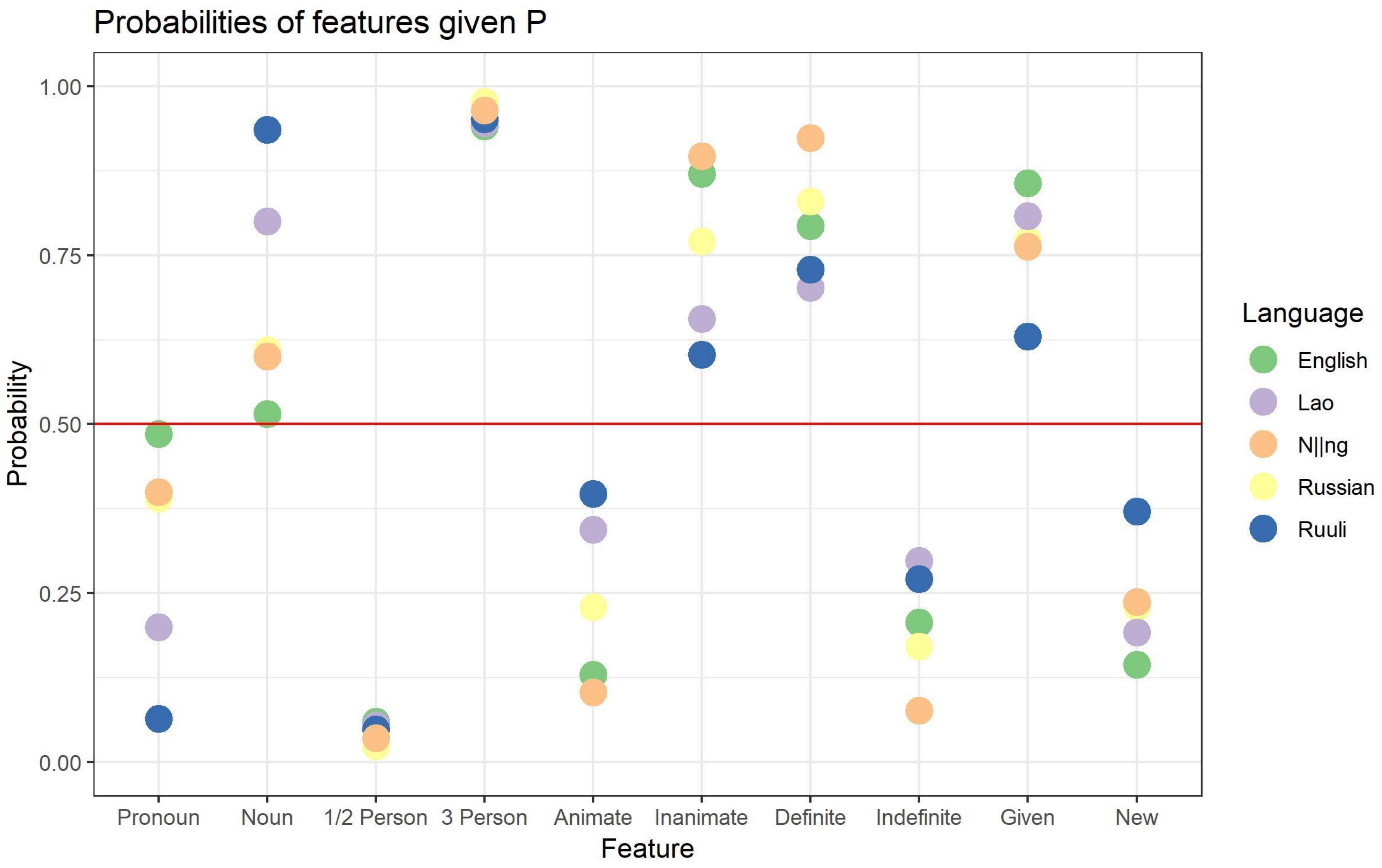
P (Feature|Role)

Feature (animate, pronoun, etc.)

Role (A or P)



Levshina & Witzlack-Makarevish, In prep.

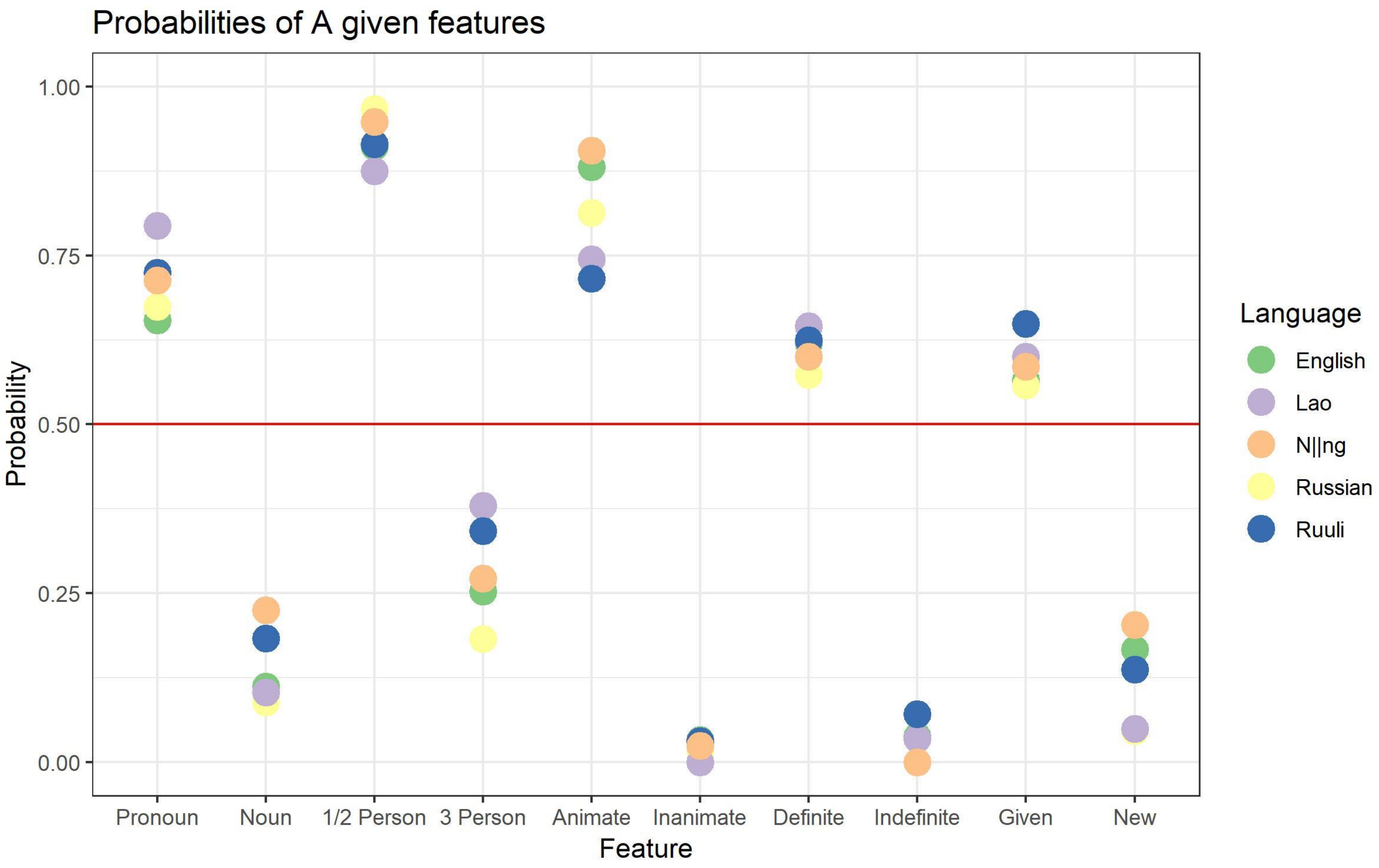


Levshina & Witzlack-Makarevish, In prep.

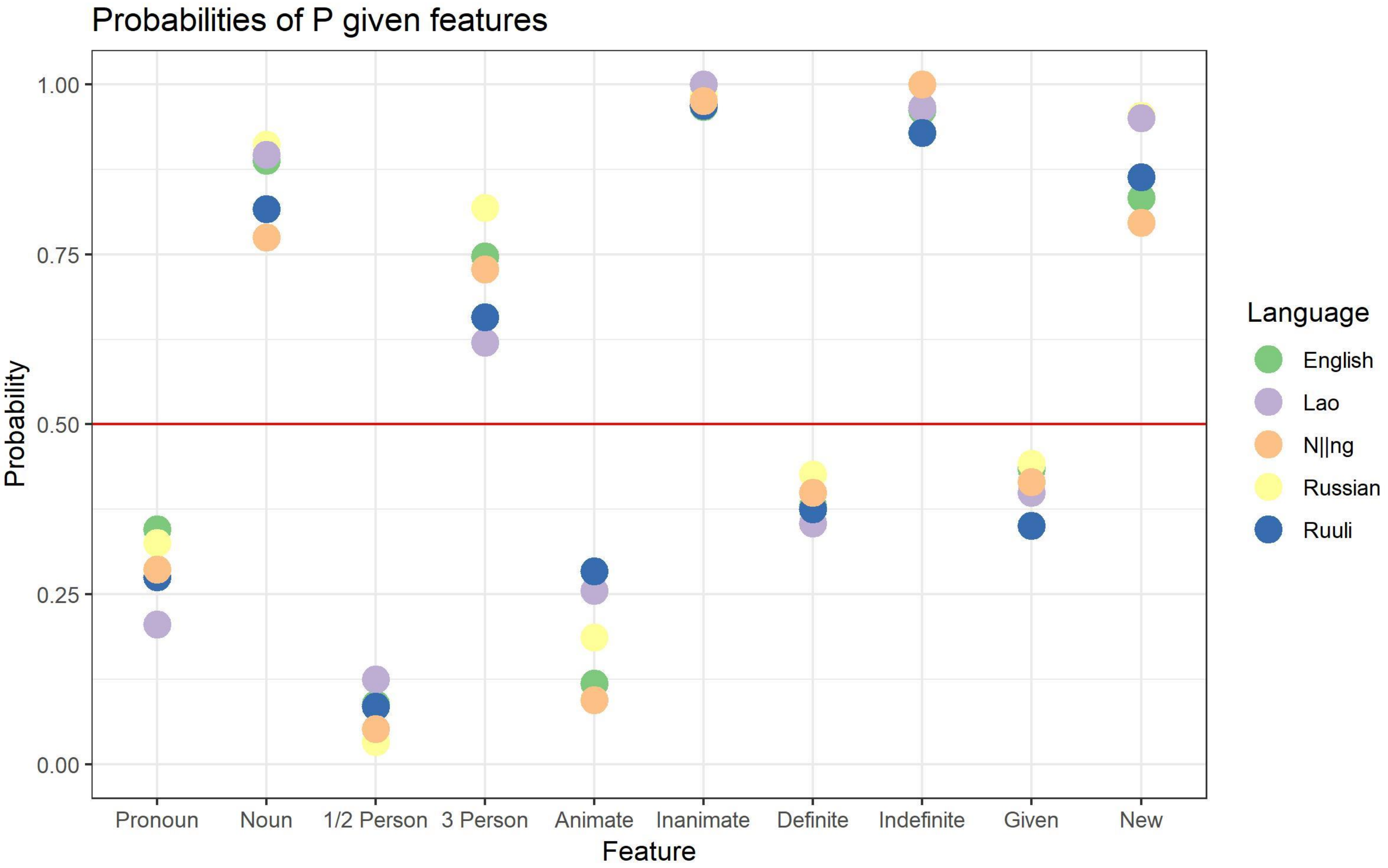
P (Role|Feature)

Role (A or P)

Feature (animate, pronoun, etc.)



Levshina & Witzlack-Makarevish, In prep.



Levshina & Witzlack-Makarevish, In prep.

Interpretation

- No need to use formal marking if a nominal with particular properties is typically an A or a P; the marking is useful when the nominal is rarely used as an A or P → efficient communication.

Interpretation

- No need to use formal marking if a nominal with particular properties is typically an A or a P; the marking is useful when the nominal is rarely used as an A or P → efficient communication.

- Cf. Haspelmath (2017):
  - non-alienable possession constructions ("my arm, sister, etc.") tend to be shorter than alienable possession constructions ("my garden, knife, etc.")
  - arm, sister, etc. are more frequently used in the possessive constructions than garden, knife, etc.
    = P(Possessed|arm) > P (Possessed|garden)

Conclusions

Advantages of using corpora

- Corpora make new directions of research possible (e.g. degrees of variability, lexical variation, fine-grained semantic distinctions).

- They allow us to reverse-engineer cross-linguistic generalizations.

- They make us think how to express hypotheses in a testable and quantifiable way.

ha

Advantages of using corpora

- Corpora make new directions of research possible (e.g. degrees of variability, lexical variation, fine-grained semantic distinctions).

- They allow us to reverse-engineer cross-linguistic generalizations.

- They make us think how to express hypotheses in a testable and quantifiable way.

Advantages of using corpora

- Corpora make new directions of research possible (e.g. degrees of variability, lexical variation, fine-grained semantic distinctions).

- They allow us to reverse-engineer cross-linguistic generalizations.

- They make us think how to express hypotheses in a testable and quantifiable way.

Challenges of using corpora

- A lot of work
- Bias towards major and Indo-European languages
- Bias towards written texts
- Theoretical and practical issues of cross-linguistic comparability (tokenization, POS annotation, syntactic parsing)
- Keeping in mind that we are dealing with **doculects**, not with **languages** per se (but what are the latter?)

Challenges of using corpora

- A lot of work
- Bias towards major and Indo-European languages
- Bias towards written texts
- Theoretical and practical issues of cross-linguistic comparability (tokenization, POS annotation, syntactic parsing)
- Keeping in mind that we are dealing with **doculects**, not with **languages** per se (but what are the latter?)

Challenges of using corpora

- A lot of work
- Bias towards major and Indo-European languages
- Bias towards written texts
- Theoretical and practical issues of cross-linguistic comparability (tokenization, POS annotation, syntactic parsing)
- Keeping in mind that we are dealing with **doculects**, not with **languages** per se (but what are the latter?)

Challenges of using corpora

- A lot of work
- Bias towards major and Indo-European languages
- Bias towards written texts
- Theoretical and practical issues of cross-linguistic comparability (tokenization, POS annotation, syntactic parsing)
- Keeping in mind that we are dealing with **doculects**, not with **languages** per se (but what are the latter?)

Challenges of using corpora

- A lot of work
- Bias towards major and Indo-European languages
- Bias towards written texts
- Theoretical and practical issues of cross-linguistic comparability (tokenization, POS annotation, syntactic parsing)
- Keeping in mind that we are dealing with **doculects**, not with **languages** per se (but what are the latter?)

Thank you for your attention!