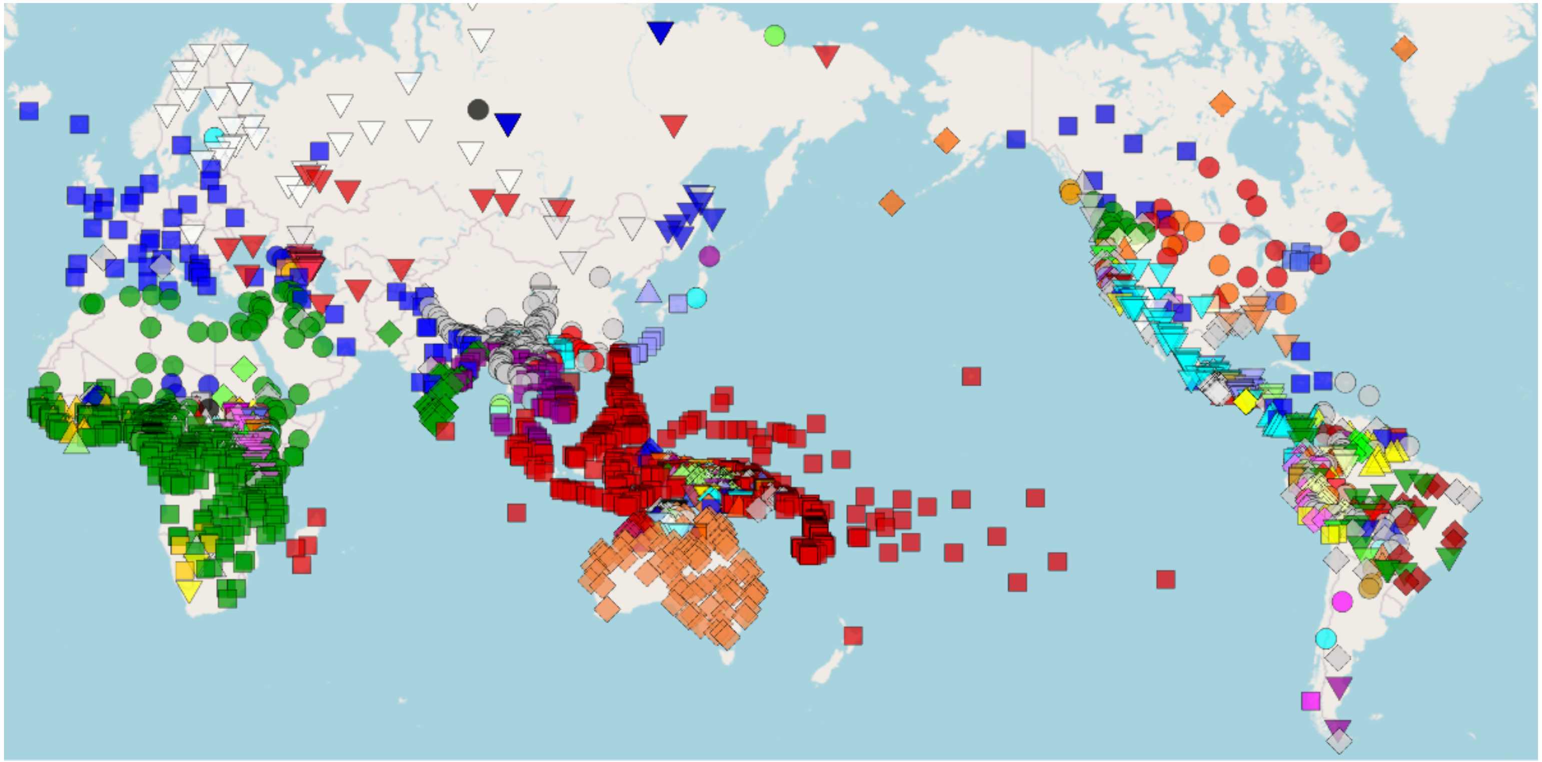
Diversity of the world's languages



Roger Levy

9.19: Computational Psycholinguistics

11 December 2023

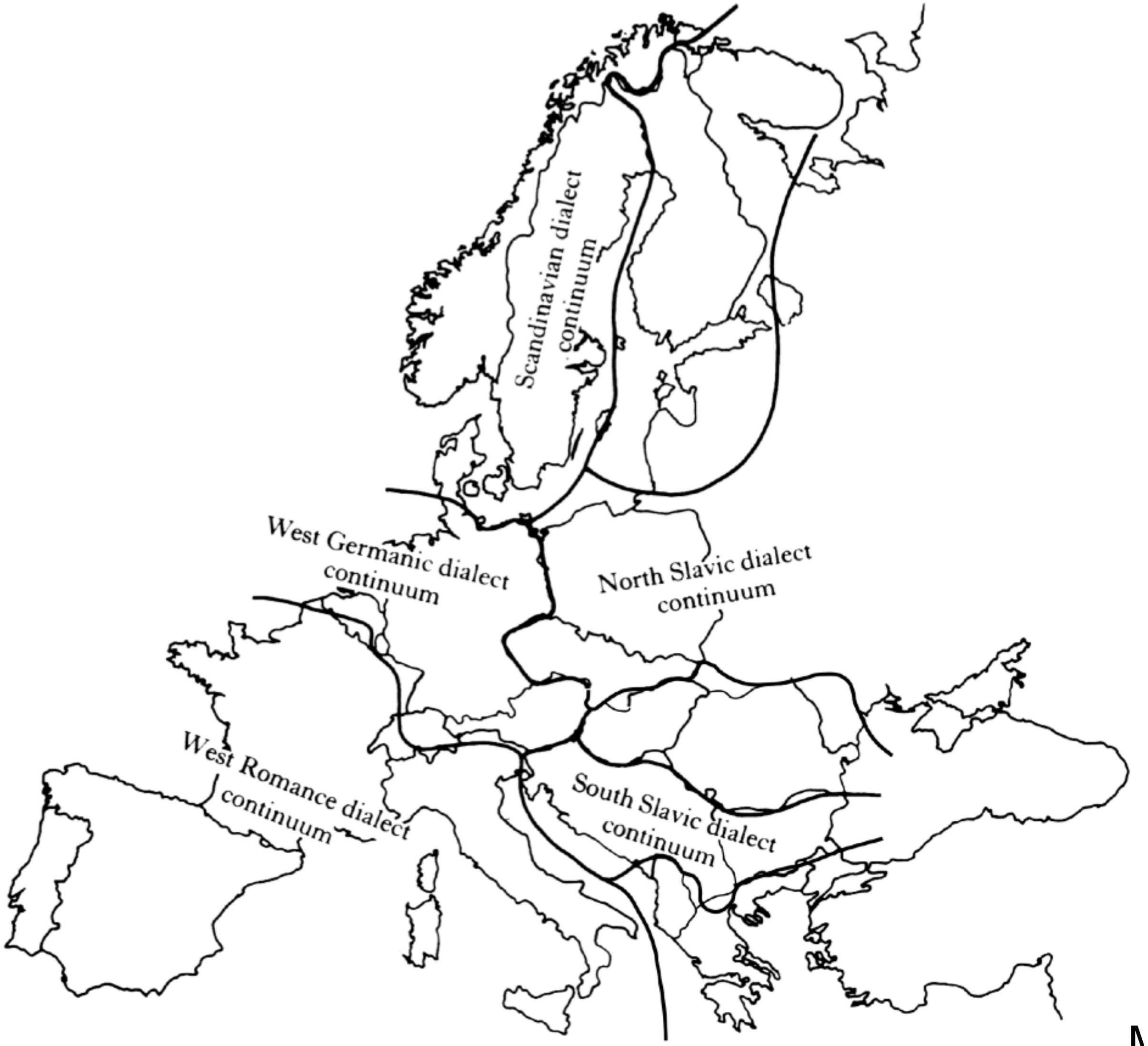
What constitutes a language?

- Chambers & Trudgill (1998) ask instead: what's a *dialect*?

  We...accept the notion that all speakers are speakers of at least one dialect – that standard English, for example, is just as much a dialect as any other form of English – and that it does not make any kind of sense to suppose that any one dialect is in any way linguistically superior to any other.

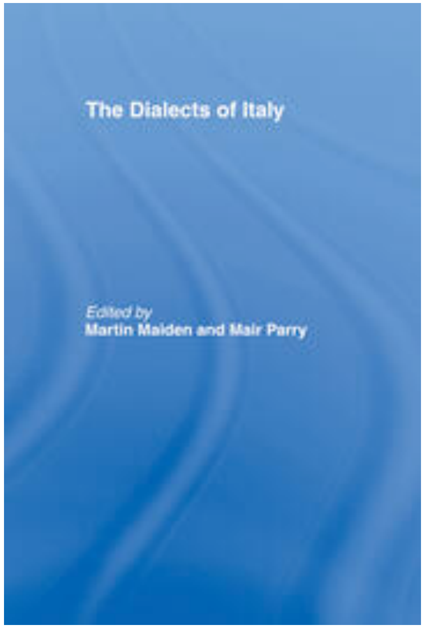
- Candidate for definition: **a language is a collection of mutually intelligible dialects**
- But, there are two potential problems:
  - Political: Mutually intelligible "dialects" may be conventionally viewed as different "languages" (e.g., Norwegian, Swedish, Danish)
  - Scientific: Intelligibility is not a categorical property, and mutual intelligibility is not necessarily a transitive or even symmetric relationship

European dialect continua


Scandinavian dialect continuum
West Germanic dialect continuum
North Slavic dialect continuum
West Romance dialect continuum
South Slavic dialect continuum

Map 1-1. European dialect continua

(Chambers & Trudgill, 1998)



The Dialects of Italy

Edited by
Martin Maiden and Mair Parry

Max Weinreich (1894–1969):
"A language is a dialect with an army and navy"

Investigating intelligibility & dialect "distance"

- Example (Gooskens et al., 2018): do a cloze test in European languages with speakers of different languages

Investigating intelligibility & dialect "distance"

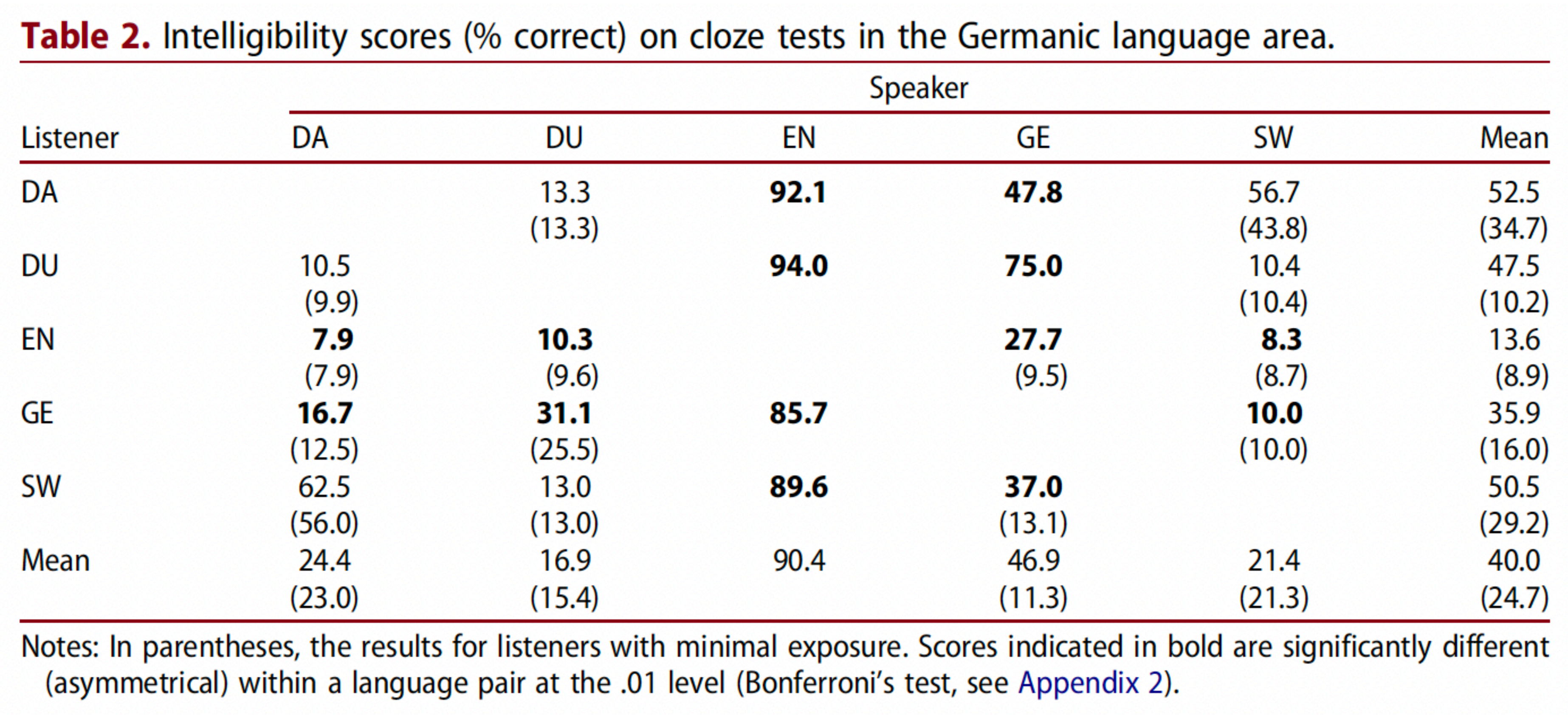


Table 2. Intelligibility scores (% correct) on cloze tests in the Germanic language area.

	Speaker					
Listener	DA	DU	EN	GE	SW	Mean
DA		13.3 (13.3)	**92.1**	**47.8**	56.7 (43.8)	52.5 (34.7)
DU	10.5 (9.9)		**94.0**	**75.0**	10.4 (10.4)	47.5 (10.2)
EN	**7.9** (7.9)	**10.3** (9.6)		**27.7** (9.5)	**8.3** (8.7)	13.6 (8.9)
GE	**16.7** (12.5)	**31.1** (25.5)	**85.7**		**10.0** (10.0)	35.9 (16.0)
SW	62.5 (56.0)	13.0 (13.0)	**89.6**	**37.0** (13.1)		50.5 (29.2)
Mean	24.4 (23.0)	16.9 (15.4)	90.4	46.9 (11.3)	21.4 (21.3)	40.0 (24.7)

Notes: In parentheses, the results for listeners with minimal exposure. Scores indicated in bold are significantly different (asymmetrical) within a language pair at the .01 level (Bonferroni's test, see Appendix 2).

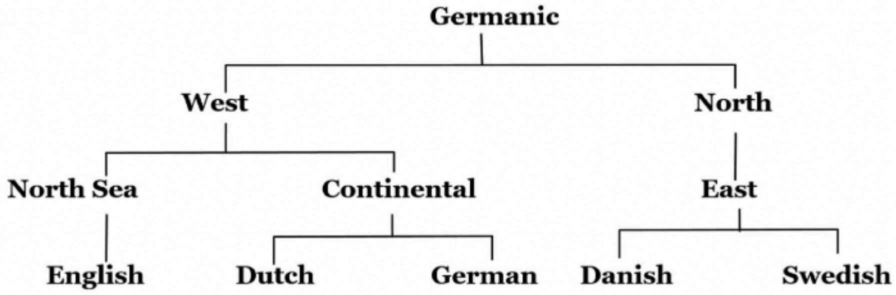


Figure 4. Germanic language tree.

(Gooskens et al., 2018)

Investigating intelligibility & dialect "distance"

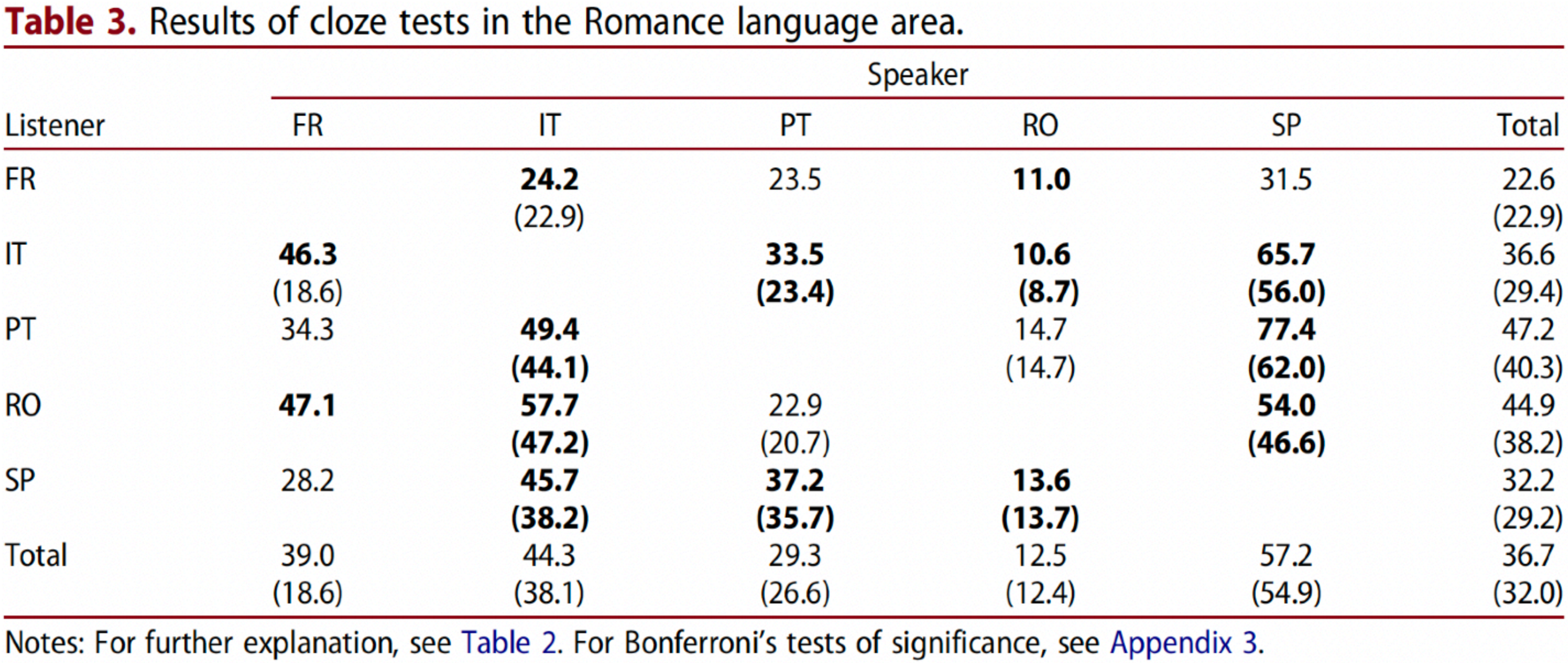


Table 3. Results of cloze tests in the Romance language area.

	Speaker					
Listener	FR	IT	PT	RO	SP	Total
FR		**24.2**	23.5	**11.0**	31.5	22.6
		(22.9)				(22.9)
IT	**46.3**		**33.5**	**10.6**	**65.7**	36.6
	(18.6)		**(23.4)**	**(8.7)**	**(56.0)**	(29.4)
PT	34.3	**49.4**		14.7	**77.4**	47.2
		(44.1)		(14.7)	**(62.0)**	(40.3)
RO	**47.1**	**57.7**	22.9		**54.0**	44.9
		(47.2)	(20.7)		**(46.6)**	(38.2)
SP	28.2	**45.7**	**37.2**	**13.6**		32.2
		(38.2)	**(35.7)**	**(13.7)**		(29.2)
Total	39.0	44.3	29.3	12.5	57.2	36.7
	(18.6)	(38.1)	(26.6)	(12.4)	(54.9)	(32.0)

Notes: For further explanation, see Table 2. For Bonferroni's tests of significance, see Appendix 3.

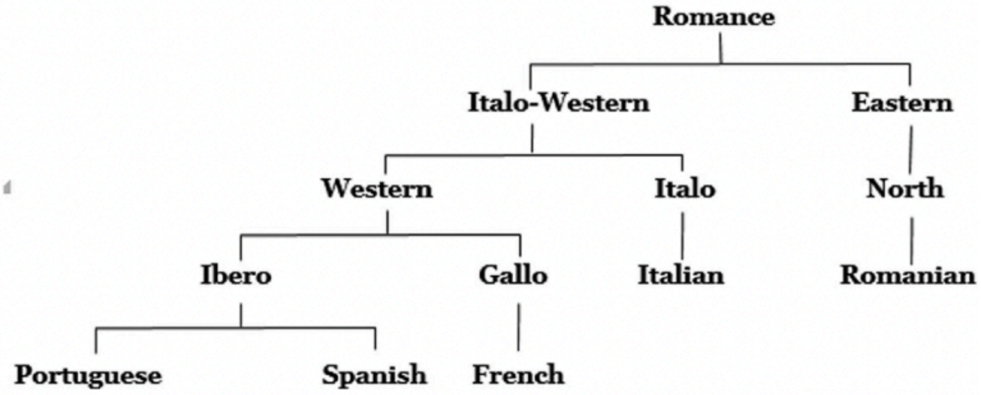


Figure 5. Romance language tree.

(Gooskens et al., 2018)

Investigating intelligibility & dialect "distance"

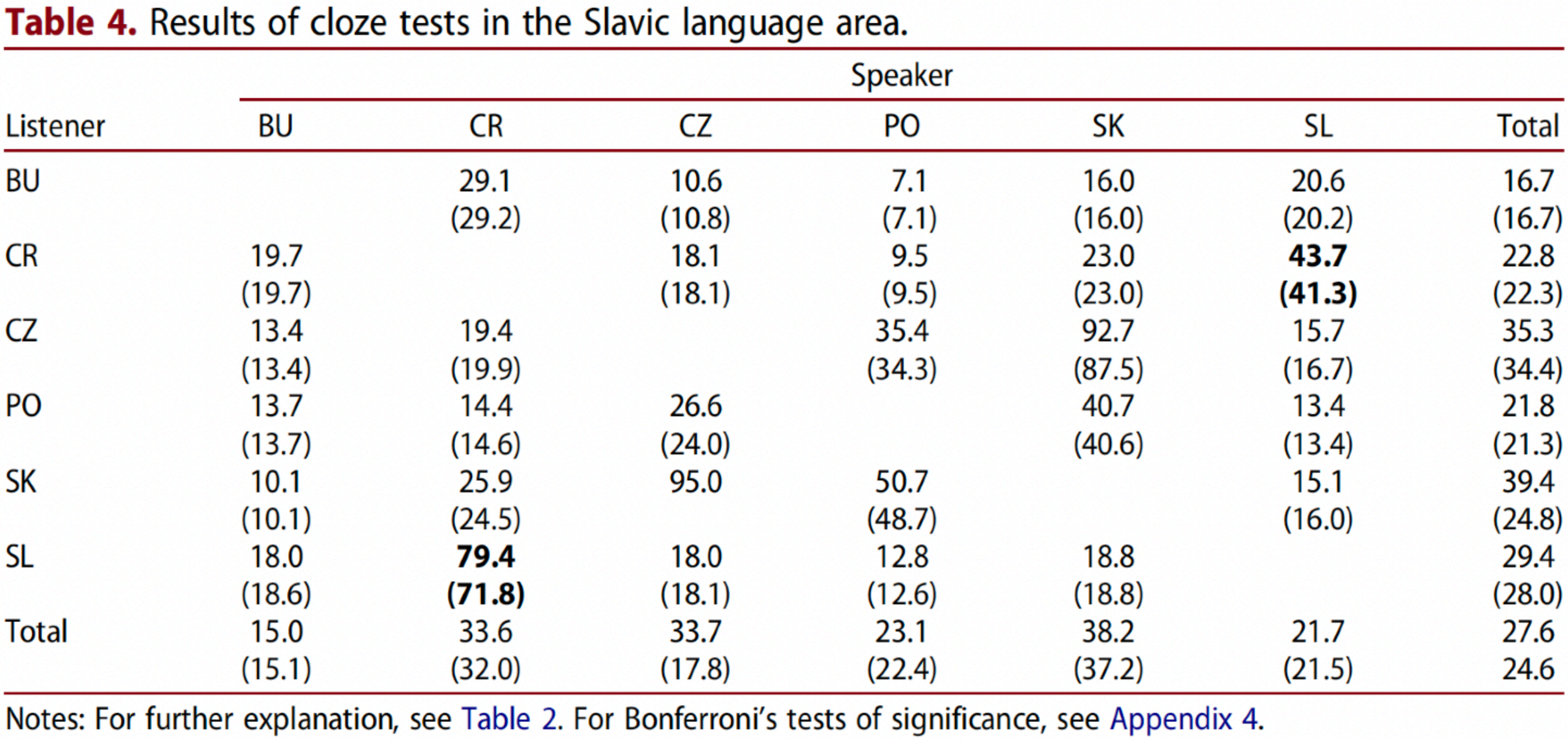
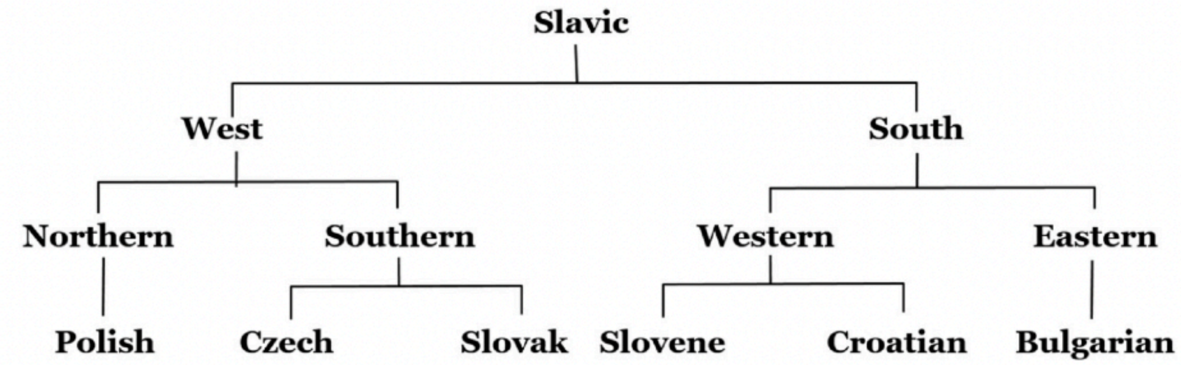


Table 4. Results of cloze tests in the Slavic language area.

	Speaker						
Listener	BU	CR	CZ	PO	SK	SL	Total
BU		29.1	10.6	7.1	16.0	20.6	16.7
		(29.2)	(10.8)	(7.1)	(16.0)	(20.2)	(16.7)
CR	19.7		18.1	9.5	23.0	**43.7**	22.8
	(19.7)		(18.1)	(9.5)	(23.0)	**(41.3)**	(22.3)
CZ	13.4	19.4		35.4	92.7	15.7	35.3
	(13.4)	(19.9)		(34.3)	(87.5)	(16.7)	(34.4)
PO	13.7	14.4	26.6		40.7	13.4	21.8
	(13.7)	(14.6)	(24.0)		(40.6)	(13.4)	(21.3)
SK	10.1	25.9	95.0	50.7		15.1	39.4
	(10.1)	(24.5)		(48.7)		(16.0)	(24.8)
SL	18.0	**79.4**	18.0	12.8	18.8		29.4
	(18.6)	**(71.8)**	(18.1)	(12.6)	(18.8)		(28.0)
Total	15.0	33.6	33.7	23.1	38.2	21.7	27.6
	(15.1)	(32.0)	(17.8)	(22.4)	(37.2)	(21.5)	24.6

Notes: For further explanation, see Table 2. For Bonferroni's tests of significance, see Appendix 4.



(Gooskens et al., 2018)

Documenting the world's languages

- Some of the key resources in language documentation:
  - **Dictionaries**
  - **Grammars** (descriptions of a language's grammar, written by someone with linguistics training)
  - **Corpora** (collections of naturalistically produced language)
- Organizing the documentation of the world's languages is a *massive data management challenge*
- One well-known, long-standing project: **Ethnologue**
  - But: not an open resource!
- Key ongoing open effort: Cross-Linguistic Linked Data project, including:
  - **Glottolog** (an open Ethnologue replacement)
  - **Grambank** (open inventory of linguistic features)
  - and many more!

Some raw facts

- Ethnologue and Glottolog document over 7,000 languages across the world!
- But 50–90% of the languages in the world are estimated to be likely to disappear by the end of this century.
- The vast majority of languages are spoken by a very small population
- Many of these languages do not necessarily have a written form

How do we identify language relationships?

- We'll cover this now with an in-class handout.

Structured variation in the world's languages

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Japanese:

	beddo	-o	ka-tta
(pro)	bed	-ACC	buy-PAST

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Japanese:

	beddo	-o	ka-tta
(pro)	bed	-ACC	buy-PAST

Oneida (Baker, 1996):

Wa’	-ke	-nakt	-a	-hnínu	-’
FACT	-1sS	-bed	-∅	-buy	-PUNC

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Japanese:

	beddo	-o	ka-tta
(pro)	bed	-ACC	buy-PAST

Oneida (Baker, 1996):

Wa’	-ke	-nakt	-a	-hnínu	-’
FACT	-1sS	-bed	-∅	-buy	-PUNC

- Yet there are strong (universal?) generalizations

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Japanese:

	beddo	-o	ka-tta
(pro)	bed	-ACC	buy-PAST

Oneida (Baker, 1996):

Wa’	-ke	-nakt	-a	-hnínu	-’
FACT	-1sS	-bed	-∅	-buy	-PUNC

- Yet there are strong (universal?) generalizations

Grammatical categories:

N V Adj Prep

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Japanese:

	beddo	-o	ka-tta
(pro)	bed	-ACC	buy-PAST

Oneida (Baker, 1996):

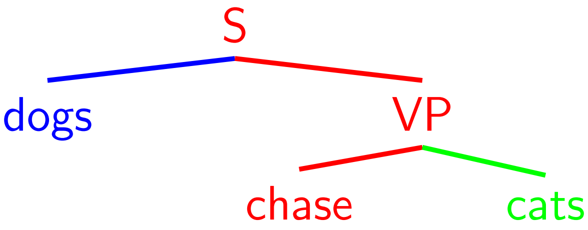
Wa’	-ke	-nakt	-a	-hnínu	-’
FACT	-1sS	-bed	-∅	-buy	-PUNC

- Yet there are strong (universal?) generalizations

Grammatical categories:

N V Adj Prep

Heads & hierarchy:



Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Japanese:

	beddo	-o	ka-tta
(pro)	bed	-ACC	buy-PAST

Oneida (Baker, 1996):

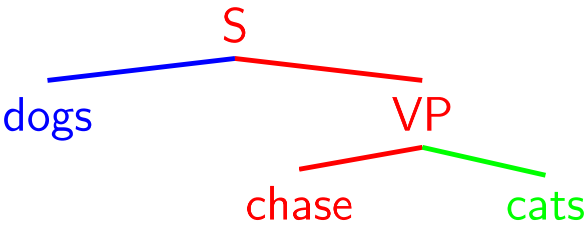
Wa'	-ke	-nakt	-a	-hnínu	-'
FACT	-1sS	-bed	-∅	-buy	-PUNC

- Yet there are strong (universal?) generalizations

Grammatical categories:

N V Adj Prep

Heads & hierarchy:



Idiosyncrasy:

$$[\![\text{kick the bucket}]\!] \neq [\![\text{kick}]\!](\iota(\lambda x.[\![bucket]\!](x)))$$

Structured variation in the world's languages

- Languages vary dramatically across the world in structure

English:

I bought the bed

Japanese:

beddo -o ka-tta
(pro) bed -ACC buy-PAST

Oneida (Baker, 1996):

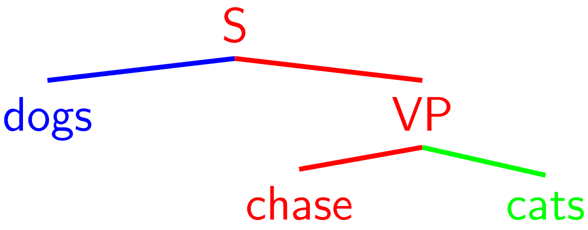
Wa' -ke -nakt -a -hnínu -'
FACT -1sS -bed -∅ -buy -PUNC

- Yet there are strong (universal?) generalizations

Grammatical categories:

N V Adj Prep

Heads & hierarchy:



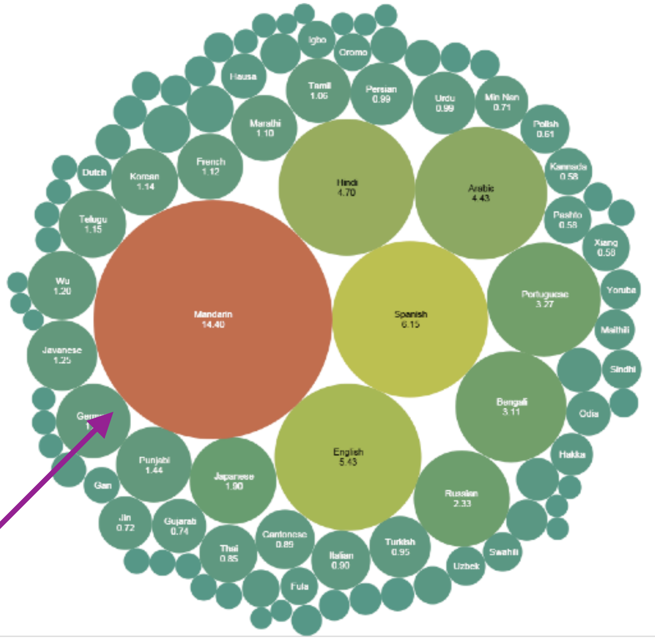
Idiosyncrasy:

$$[\![\text{kick the bucket}]\!] \neq [\![\text{kick}]\!](\iota(\lambda x.[\![bucket]\!](x)))$$

- **GOAL:** develop theories of language understanding, production, and acquisition that can account for

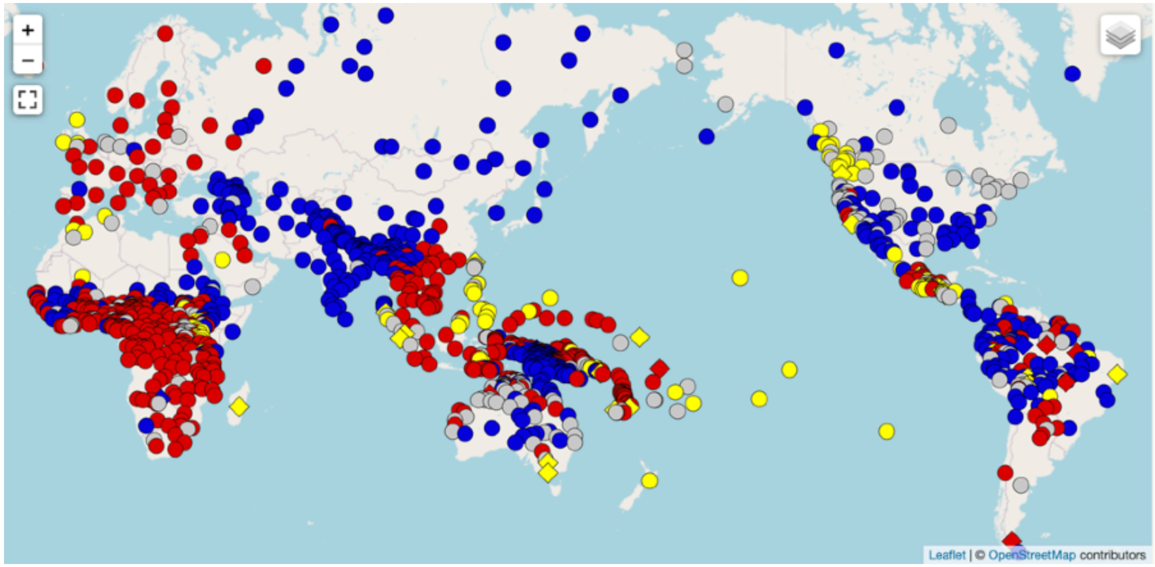
Linguistic diversity across the world

- There are 6000-7000 languages in the world
- The 100 languages with the most native speakers comprise only 85% of the world's population



Mandarin; 14.4% of the world's native speakers

●	SOV	564
●	SVO	488
●	VSO	95
◆	VOS	25
◆	OVS	11
◆	OSV	4
●	No dominant order	189



(# speakers from Nationalencyklopedin 2010; visualization due to Wikipedia user Jroehl; map from Dryer 2013, wals.info)

A bit about language typology

A bit about language typology

- Languages are systematic in different ways

A bit about language typology

- Languages are systematic in different ways
- Those differences are called **features**

A bit about language typology

- Languages are systematic in different ways
- Those differences are called **features**
- For example, Japanese and English vary in many features that these sentences exemplify...what are they?

A bit about language typology

- Languages are systematic in different ways
- Those differences are called **features**
- For example, Japanese and English vary in many features that these sentences exemplify...what are they?

English:
I bought the bed

A bit about language typology

- Languages are systematic in different ways
- Those differences are called **features**
- For example, Japanese and English vary in many features that these sentences exemplify...what are they?

English:	Japanese:			
I bought the bed		beddo	-o	ka-tta
	(pro)	bed	-ACC	buy-PAST

A bit about language typology

- Languages are systematic in different ways
- Those differences are called **features**
- For example, Japanese and English vary in many features that these sentences exemplify...what are they?

English:	Japanese:			
I bought the bed		beddo	-o	ka-tta
	(pro)	bed	-ACC	buy-PAST

- Large databases based on **grammars** of the world's languages have collated these features, and there turn out to be many interesting correlations.

A bit about language typology

- Languages are systematic in different ways
- Those differences are called **features**
- For example, Japanese and English vary in many features that these sentences exemplify...what are they?

English:	Japanese:			
I bought the bed		beddo	-o	ka-tta
	(pro)	bed	-ACC	buy-PAST

- Large databases based on **grammars** of the world's languages have collated these features, and there turn out to be many interesting correlations.
- One influential resource: the World Atlas of Language Structures (WALS; wals.info)