Linguistic Typology for NLP researchers: Methods and Resources in the 21st century

Harald Hammarström
harald.hammarstrom@lingfil.uu.se

21 Jan 2026 Yerevan

Today: Languages of the world

§1: A Definition of (Human) Language

A commmunication system with

- conventionalized
- form-meaning pairs
- capable of expressing the entire communicative needs of a human society
- learnable by humans

Based loosely on Hockett's "Design Features of Human Language" (Hockett 1960)

§1: Forms of Language So Far Attested

- With native (L1) speakers:
  - Spoken languages: form is acoustic and there is vowel/consonant distinction
  - Signed languages: form is constellations of the human body
- Without native (L1) speakers:
  - Whistled languages: form is acoustic but there is no vowel/consonant distinction and the signal is a free airstream formed by the lips
  - Drummed: form is acoustic but there is no vowel/consonant distinction and the signal is produced by means of a drum
  - Written languages: form is symbolic
  - …

§1: Canonical vs Mirrored Languages

- It turns out that
  - All known whistled languages (cf. Gartner and Streiter 2006, Meyer 2015, Thierry 2002)
  - All known drummed languages (cf. Stern 1957)
  - All written languages (once actually used by a human society)
  - A handful "sign languages" in Aboriginal Australia (Kendon 1988)
- are representations of spoken languages at some level (phoneme/syllable/morpheme/word)
- in other words, you translate from/to a mirrored language by mapping the phoneme/syllable/morpheme/word from/to the canonical language

§2: Discretizing Human Languages

- 8 billion humans on the planet, each with their own (idio)lect
- Can they be discretized into **languages** in a **scientific way**?
  - **Abstand language**: The decision of whether two varieties are the same language only depends on how different their form-meaning pairs are, i.e., if they are different enough that there is no intelligibility = MI-language
  - (**Ausbau language**: The decision of whether two varieties are the same language depends on how different their form-meaning pairs are and sociopolitical criteria such as how speakers self-identify, the existence of a literary standard, political alignment, … = socio-political language)
- Does such a scientific discretization correspond to common usage of the term "language" (or its closest equivalent in other languages than English)?
  - Not necessarily: cf. fish/mammal or berry/fruit in biology

§2: How Many Languages Are There In The World?

- FAQ to the Linguistic Society of America (LSA)
- Disregarding who has an army and a navy
- LSA says this is impossible to answer (Anderson 2005)
- Even on purely linguistic grounds, because
  - #1 The criterion of yes/no mutual intelligibility leads to contradictions in dialect chains
  - #2 Mutual intelligibility is a matter of degree anyway
  - #3 Our practical knowledge of the facts is not trustworthy

"a matter of opinion rather than science"

§2: #1 Yes/No Mutual Intelligibility Cannot Be Used?

- "… the intelligibility criterion actually leads to contradictory results, namely when we have a dialect chain
- … If one takes a simplified dialect chain *A* - *B* - *C*, where
  - *A* and *B* are mutually intelligible,
  - as are *B* and *C*,
  - but *A* and *C* are mutually unintelligible,
- then one arrives at the contradictory result that *A* and *B* are dialects of the same language, *B* and *C* are dialects of the same language, but *A* and *C* are different languages.
- … There is in fact no way of resolving this contradiction if we maintain the traditional strict difference between language and dialects
- … In this sense, it is impossible to answer the question how many languages are spoken in the world" (Comrie 1987:3)

Hammarström 2008

§2: Almost Every Intro Says Something Tantamount

- "*Such situation are referred to by linguists as 'dialect chains', and they result in sometimes arbitrary decisions being made as to how many languages are involved.*" (Lynch and Crowley 2001:2)
- "*A common situation is a string of similar varieties, in which the speakers of variety A understand those of C, and so on, but the speakers of A do not understand the variety at the other end of the continuum, or even those part way along. Even if we can define 'understand', where is the divide between language and dialect in this situation?*" (Heine and Nurse 2000:2)
- "*If a chain of dialects ABCDE exists such that mutual intelligibility decreases with distance, A and E may be mutually unintelligible, but so long as intermediate dialect gradations exist it is impossible to draw a language boundary anywhere within the chain.*" (Blust 2001:250)

§2: MI and Defining Languages

How can we get it both consistent and intuitive?

- **If two varieties are the same language then they are mutually intelligible**

  We keep this!

- ~~**If two varieties are mutually intelligible then they are necessarily of the same language**~~

  Replace by a weaker requirement: Do not unnecessarily multiply the number of languages

§2: Formal MI and Defining Languages

Given:

- A finite set X of speech varieties

  this can be on the level of varieties, dialects, subdialects, idiolects, subidiolects, …

- a binary symmetric strict yes/no relation of mutual intelligibility (henceforth MI)

  It has to be symmetric: $MI(A, B) = MI(B, A)$

Define:

The number of languages in X is the least k such that one can partition X into k blocks such that all members within a block are MI

§2: In Terms of Graphs

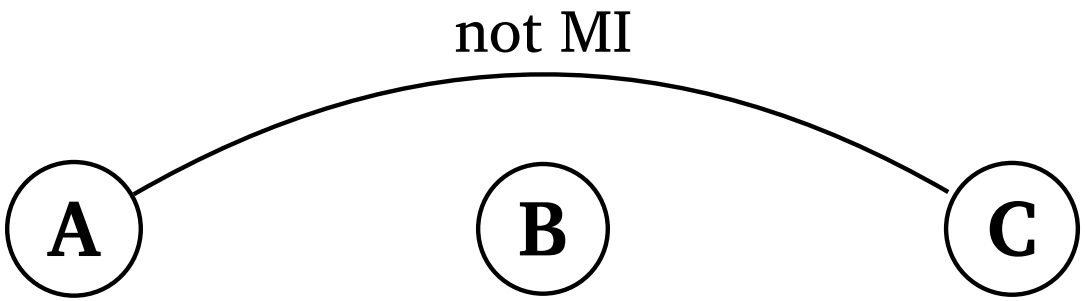
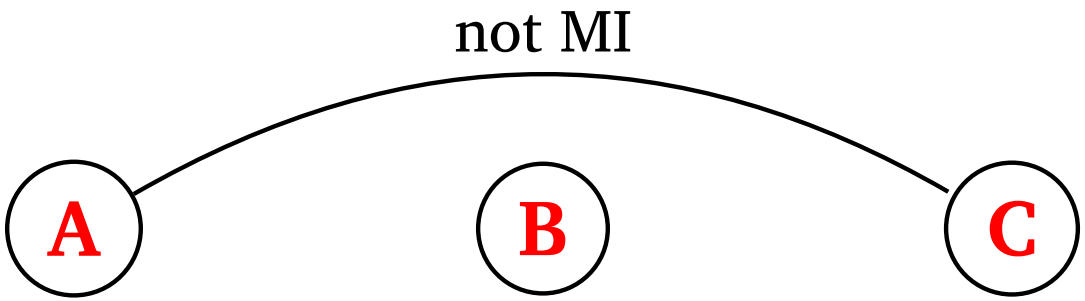


Figure: 1. Graph for (A, B), (B, C) are MI but (A, C) are not MI.

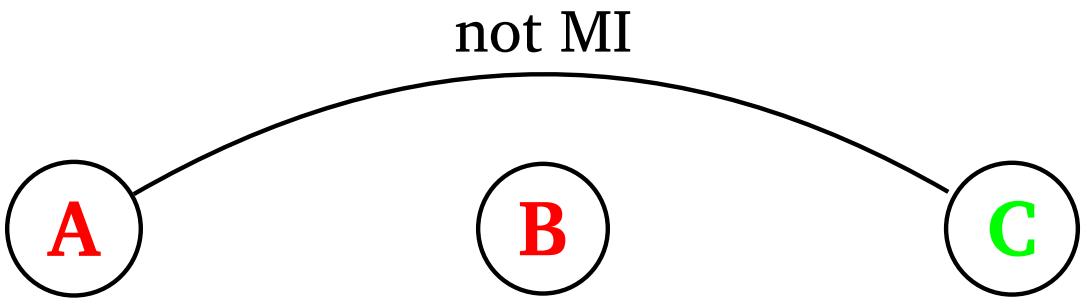
- Represent all varieties as nodes
- Draw an edge between every pair of nodes which are **not** MI
- The number of languages is the **smallest number** k of colours needed to colour all the nodes such that **no pair of nodes that share an edge have the same colour**

§2: A-B-C How Many Languages?



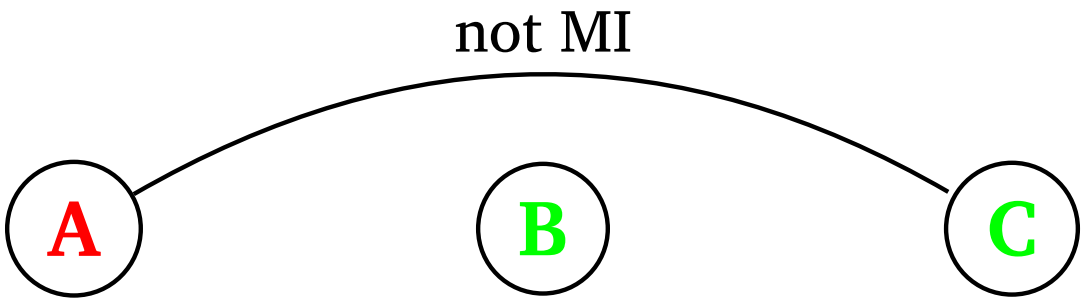
k	Blocks	No nodes that share an edge have the same color
$k = 1$	One possibility: $\{A, B, C\}$	False

§2: A-B-C How Many Languages?



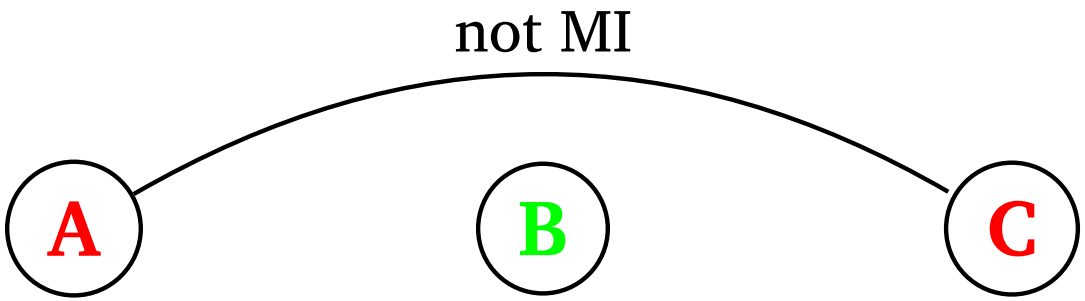
k	Blocks	No nodes that share an edge have the same color
$k = 1$	One possibility: $\{A, B, C\}$	False
$k = 2$	Three possibilities $\{A, B\}, \{C\}$ or	True

§2: A-B-C How Many Languages?



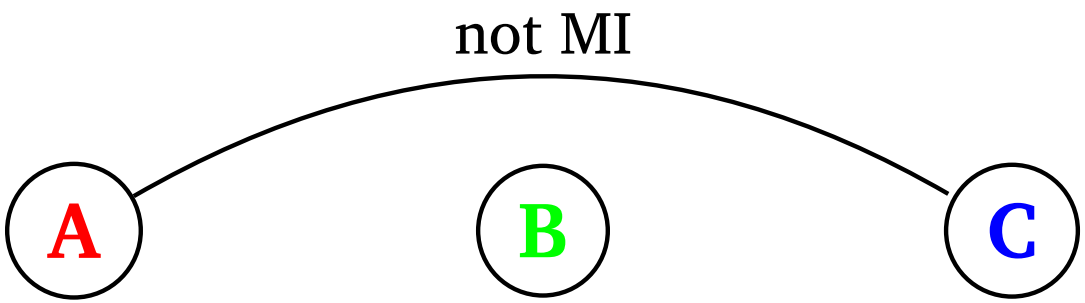
k	Blocks	No nodes that share an edge have the same color
$k = 1$	One possibility: $\{A, B, C\}$	False
$k = 2$	Three possibilities	
	$\{A, B\}, \{C\}$ or	True
	$\{A\}, \{B, C\}$ or	True

§2: A-B-C How Many Languages?



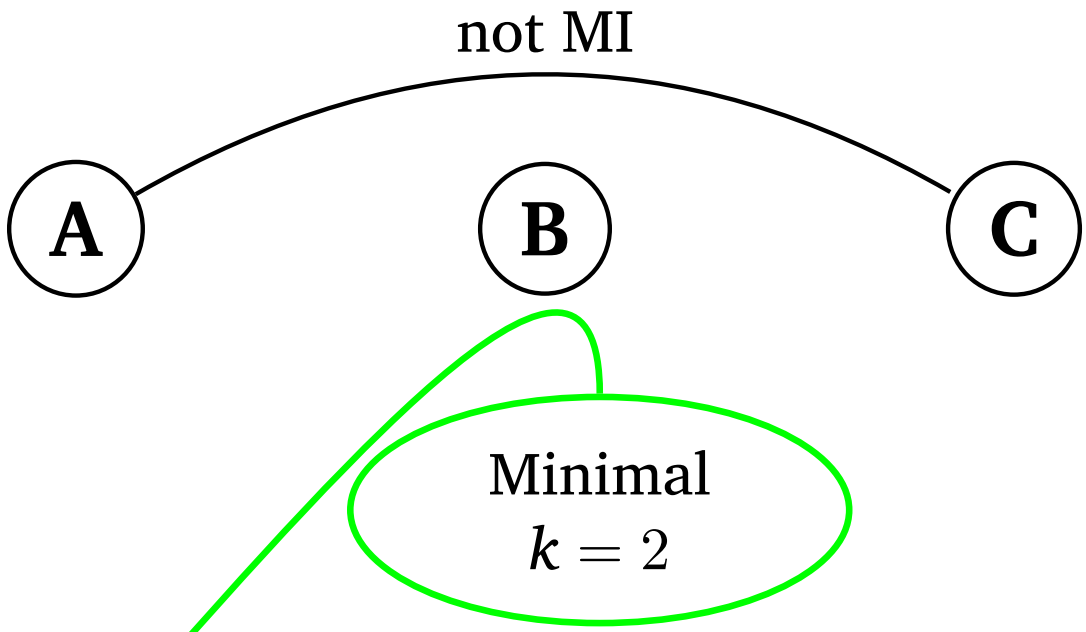
k	Blocks	No nodes that share an edge have the same color
$k = 1$	One possibility: $\{A, B, C\}$	False
$k = 2$	Three possibilities	
	$\{A, B\}, \{C\}$ or	True
	$\{A\}, \{B, C\}$ or	True
	$\{A, C\}, \{B\}$	False

§2: A-B-C How Many Languages?



k	Blocks	No nodes that share an edge have the same color
$k = 1$	One possibility: $\{A, B, C\}$	False
$k = 2$	Three possibilities	
	$\{A, B\}, \{C\}$ or	True
	$\{A\}, \{B, C\}$ or	True
	$\{A, C\}, \{B\}$	False
$k = 3$	One possibility: $\{A\}, \{B\}, \{C\}$	True

§2: A-B-C = 2 languages



k	Blocks	No nodes that share an edge have the same color
~~$k = 1$~~	~~One possibility: $\{A, B, C\}$~~	~~False~~
$k = 2$	Three possibilities	
	$\{A, B\}, \{C\}$ or	True
	$\{A\}, \{B, C\}$ or	True
	~~$\{A, C\}, \{B\}$~~	~~False~~
~~$k = 3$~~	~~One possibility: $\{A\}, \{B\}, \{C\}$~~	~~True~~

§2: Another Example

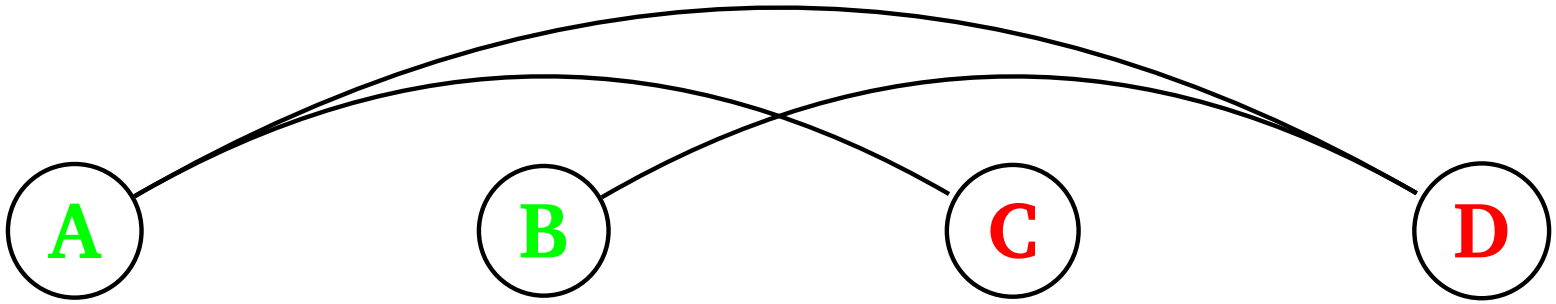
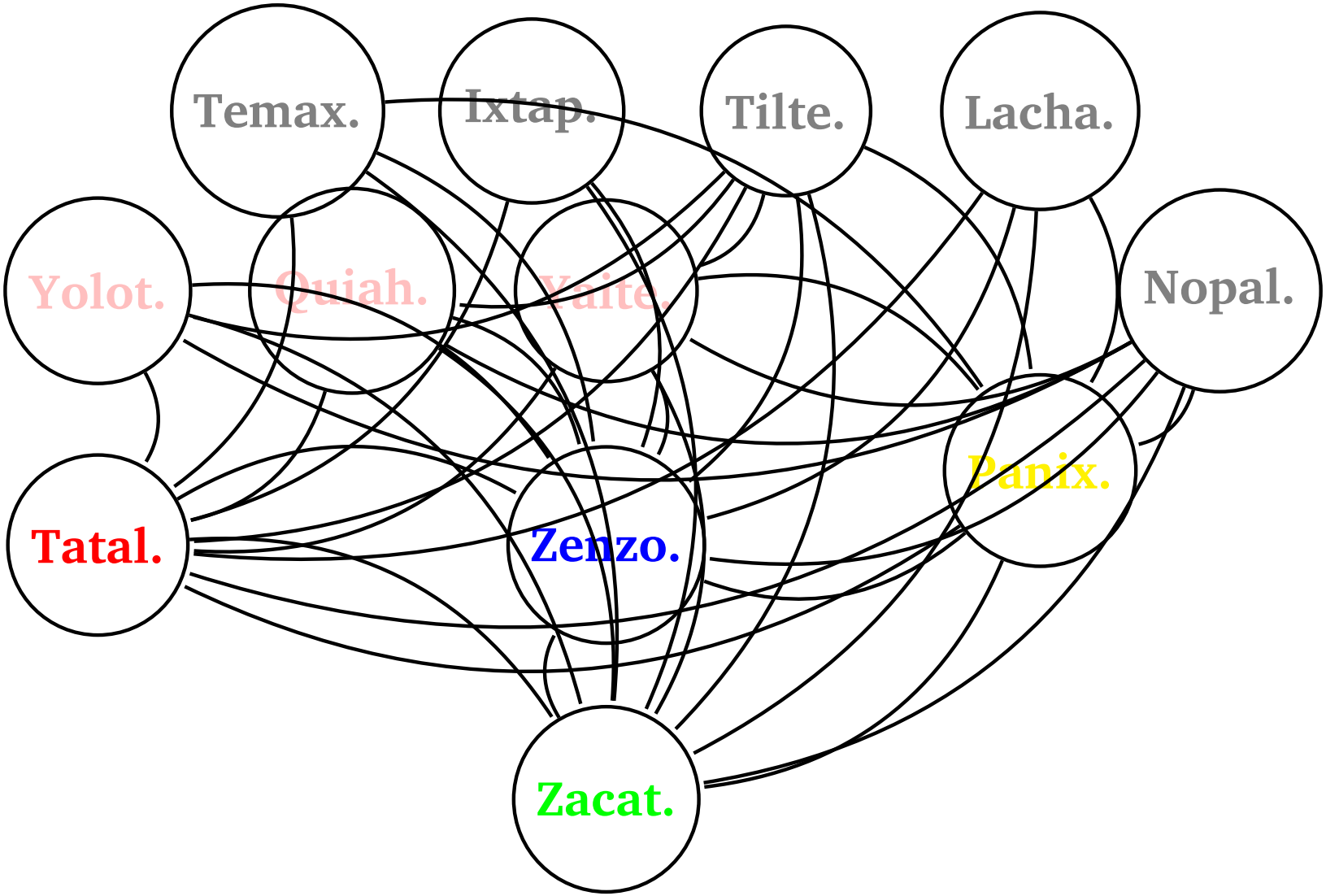


Figure: 2. Graph for (A, B), (B, C), (C, D) are MI but no other pairs are MI.

- There is only one 2-colouring: A, B green and C, D red
  $<=$ *One colour is not sufficient. If one tries 2 colours: A must have some colour, then C and D must have a different colour. Nothing prevents C and D from having the same colour, so they get the same colour. Now, only B remains which cannot be coloured by the colour of C and D, but it can have A:s colour.*

§2: Real Example: Chatino (Egland et al. 1983:10-11)



The minimal number of colours is 6

§2: Some Properties of the MI Definition

- There is always only one minimal number k of languages
- The exact constellation is **not** unique: there may be several different constellations which all have a minimal number of colours
- One can prove mathematically that:
  - All those who speak the same language speak varieties which are mutually intelligible.
  - There are no "superfluous" languages, i.e., a person speaking exactly one variety of each of the languages can communicate with everyone, whereas someone speaking less than k varieties cannot communicate with everyone.
- To find the number of languages in an arbitrary graph is NP-hard (Garey and Johnson 1979:191, in the worst case you have to try out all partitions, or approximate)

§2: #2 Mutual Intelligibility in Natural Languages

Traditional View (Beijering et al. 2008, Casad 1974, Gooskens 2013, Milliken and Milliken 1996, Voegelin and Harris 1951 etc)

- Intelligibility is a **gradient** property ranging from 0% to 100%
- Any claim of it being yes/no property corresponds to an arbitrary threshold, e.g.
  - 70% lexical similarity
  - Score 87% in a sentence repetition test
  - Score X% in a text comprehension test
  - …

§2: Example

For a variety of reasons, it is simply impossible to prepare anything like a 'definitive' index of African language and dialect names. … However, there are countless instances in which 'mutual intelligibility' is a **border-line proposition**; *members of two groups may be able to communicate if they speak carefully, or after a few hours or days or weeks of experience. In such cases, even if the facts are known and defined, there is* **no accepted criterion** *for deciding* **how much difference** *justifies defining the speech of different groups as different languages rather than dialects of a single language. (Welmers 1971:761)*

§2: A Similar Issue in Biology

Producing a fertile offspring is not a yes/no property

- Some individuals can breed a fertile offspring
- Some individuals can never breed a fertile offspring
- But for some, a non-trivial percentage, e.g. 70% of the offspring are fertile

§2: BUT: Anecdote

I have some level of proficiency in a foreign language and I am having a conversation with somebody. He/she says something which I partly understand, and I ask him/her about the part(s) I didn't understand:

Yes: I understand the his/her explanation of the part(s) I didn't understand, and therefore I (now) understand the original utterance.

No: His/her explanation of the part(s) I didn't understand contain further parts I don't understand, and I my request for explanation thus doesn't make headway

§2: A Simplistic Formal Language

a. A mapping of [single] words/morphemes to meaning:

$$L : W \rightarrow \mathcal{U}$$

b. A mapping of sentences $\langle w_1, \ldots, w_n \rangle$, $w_i \in W$ to meaning, which satisfies substitutability:

$$L : \langle w_1, \ldots, w_n \rangle \rightarrow \mathcal{U} = f(L(w_1), \ldots, L(w_n))$$

c. For every morpheme $w \in W$, there is a sentence composed of the other words $\subseteq W \setminus \{w\}$ which describes w:

$$\forall w \in W \; \exists \langle s \rangle \in [W \setminus \{w\}] \; L(\langle s \rangle) = L(w)$$

$=$ *Monolingual Dictionary Property (MDP)*

§2: Defining Yes/No Intelligibility: Intuition

$$L_A \lesssim L_B\text{:}$$

Language A is intelligible to B iff

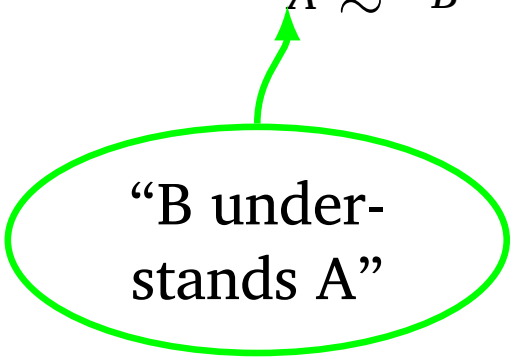
For all sentences [of A], for every morpheme therein that B does not understand, there is a path of explanations that ultimately allows B to grasp the original sentence

§2: Defining Yes/No Intelligibility: Formally #1

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

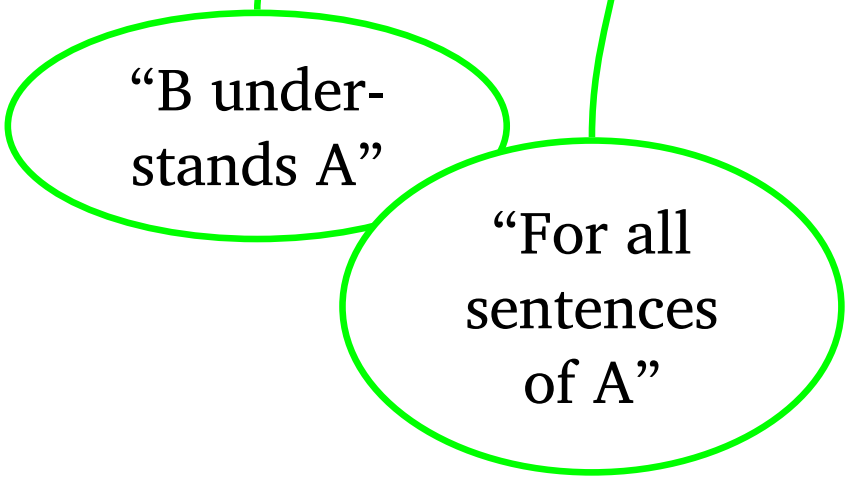
§2: Defining Yes/No Intelligibility: Formally #1

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$



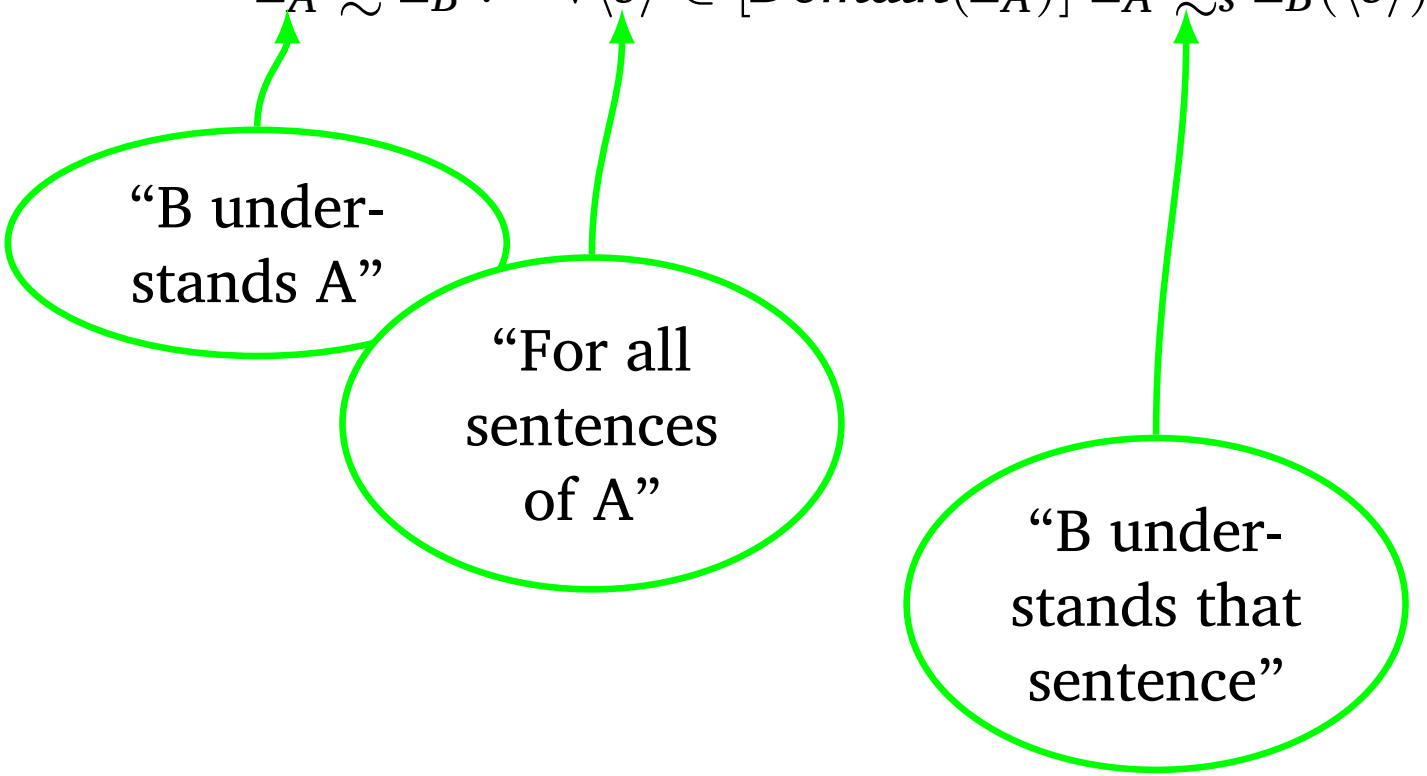
§2: Defining Yes/No Intelligibility: Formally #1

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$



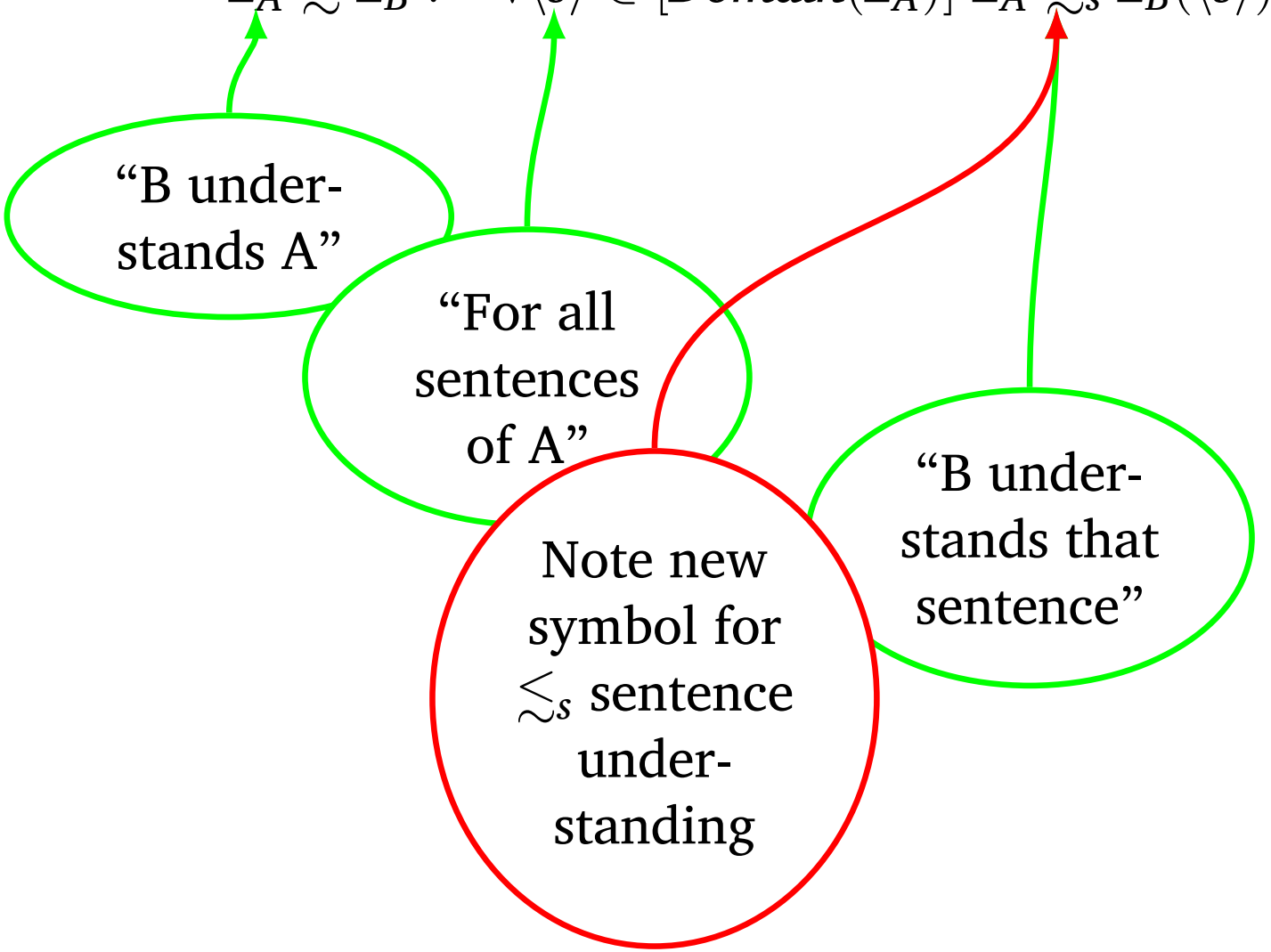
§2: Defining Yes/No Intelligibility: Formally #1

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$



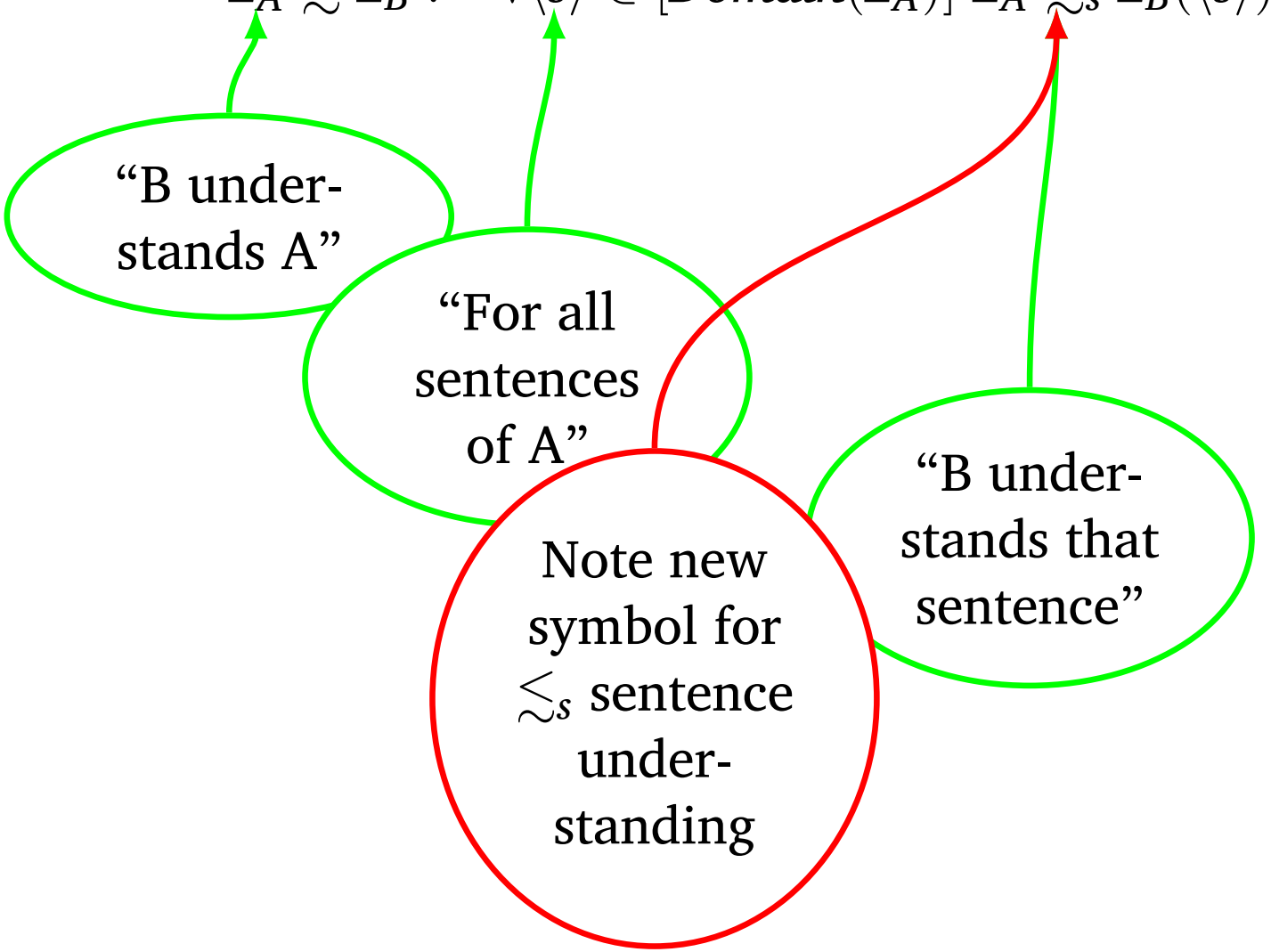
§2: Defining Yes/No Intelligibility: Formally #1

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$



§2: Defining Yes/No Intelligibility: Formally #1

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$



§2: Defining Yes/No Intelligibility: Formally #2

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneqq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

§2: Defining Yes/No Intelligibility: Formally #2

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneqq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

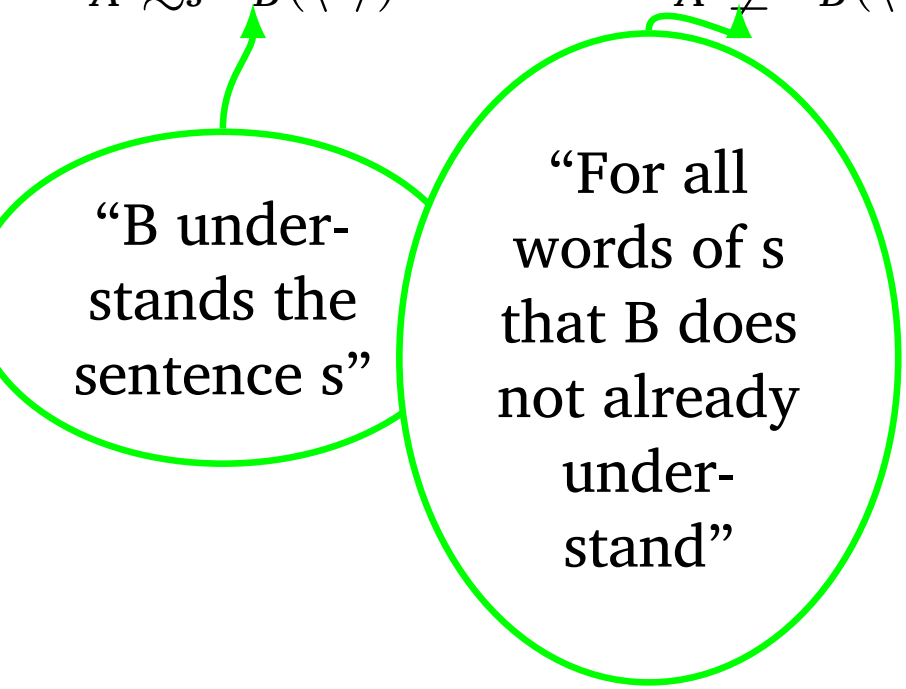
"B understands the sentence s"

§2: Defining Yes/No Intelligibility: Formally #2

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

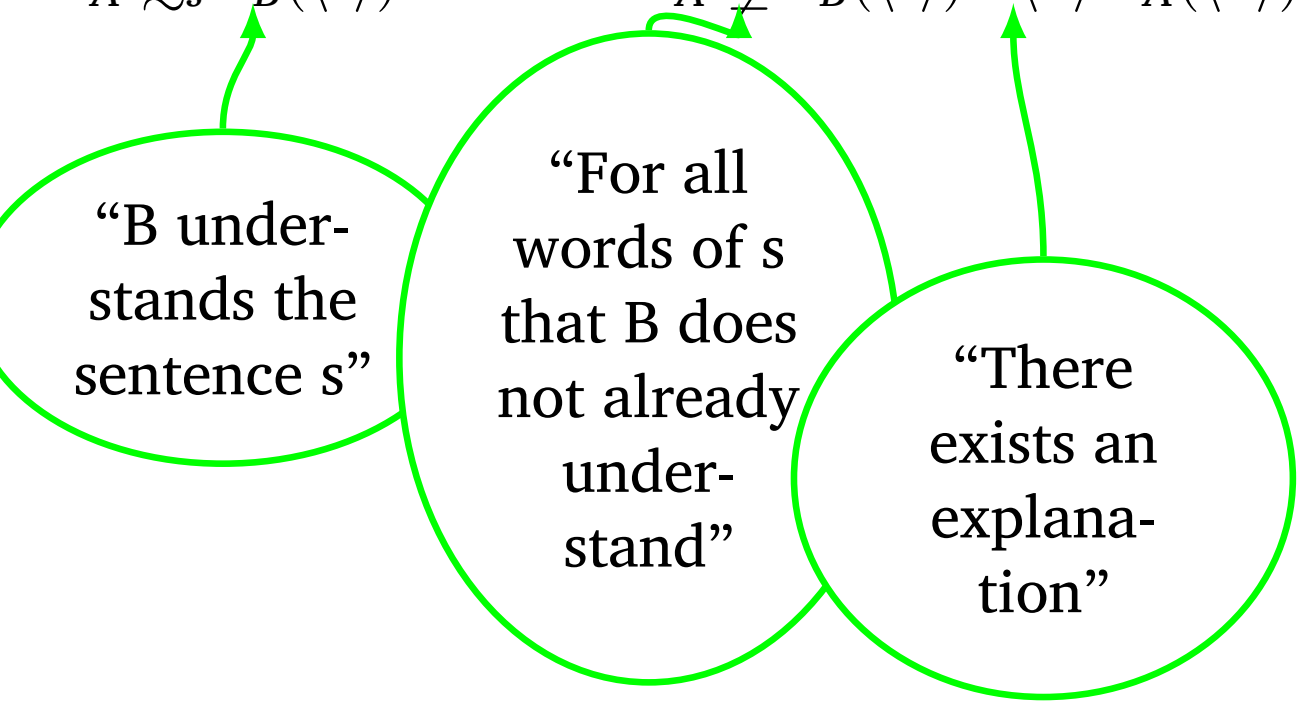


§2: Defining Yes/No Intelligibility: Formally #2

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

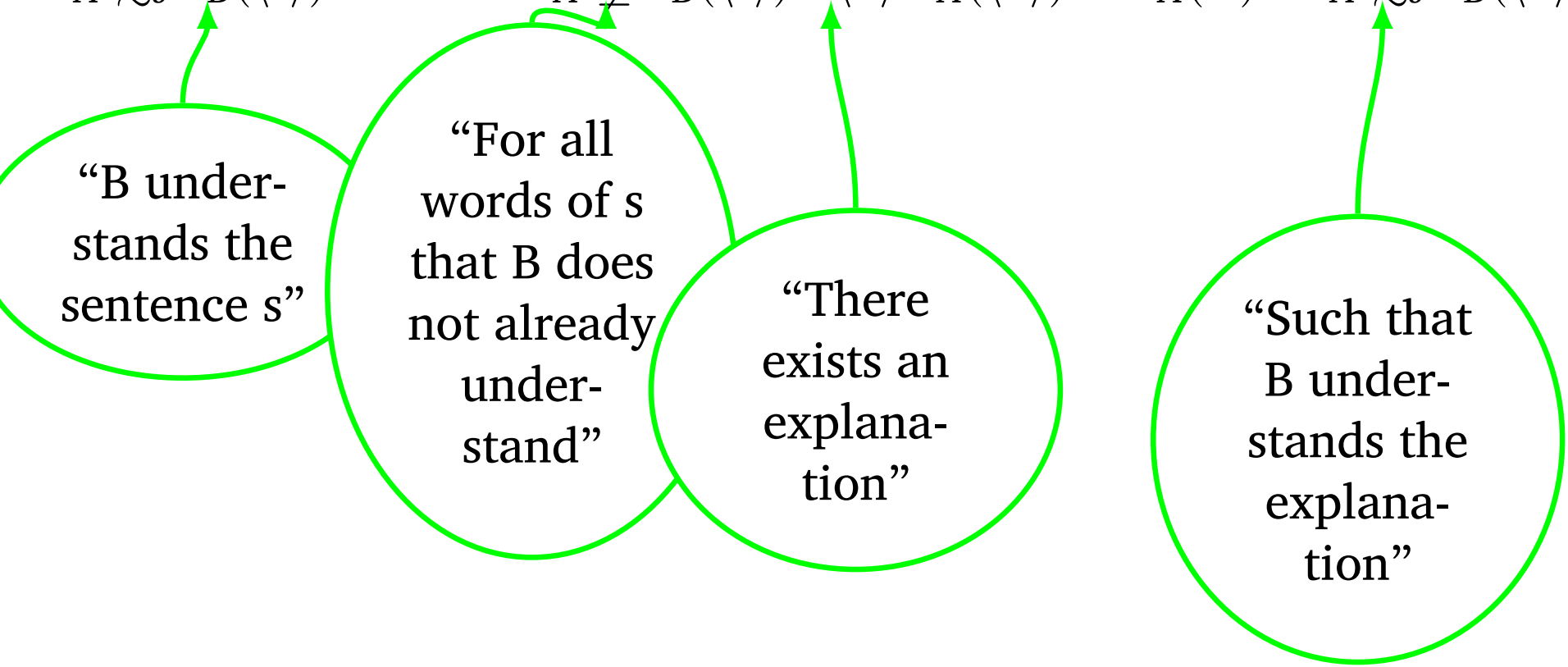


§2: Defining Yes/No Intelligibility: Formally #2

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$



§2: Defining Yes/No Intelligibility: Formally

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneqq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

and

$$L_A \subsetneqq L_B(\langle w_1, \ldots, w_n \rangle) := \{w_i | L_A(w_i) \neq L_B(w_i)\}$$

§2: Defining Yes/No Intelligibility: Formally

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneqq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

and

$$L_A \subsetneqq L_B(\langle w_1, \ldots, w_n \rangle) := \{w_i | L_A(w_i) \neq L_B(w_i)\}$$

"B fails to understand the words of A that are not the same"

§2: Defining Yes/No Intelligibility: Formally

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneqq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

and

$$L_A \subsetneqq L_B(\langle w_1, \ldots, w_n \rangle) := \{w_i | L_A(w_i) \neq L_B(w_i)\}$$

Words of that are the same serve as base cases for the recursively defined sentence intelligibility relation

"B fails to understand the words of A that are not the same"

§2: Defining Yes/No Intelligibility: Formally #3

$$L_A \lesssim L_B := \forall \langle s \rangle \in [Domain(L_A)]\ L_A \lesssim_s L_B(\langle s \rangle)$$

where

$$L_A \lesssim_s L_B(\langle s \rangle) := \forall w \in L_A \subsetneqq L_B(\langle s \rangle)\ \exists \langle x \rangle\ L_A(\langle x \rangle) = L_A(w) \wedge L_A \lesssim_s L_B(\langle x \rangle)$$

and

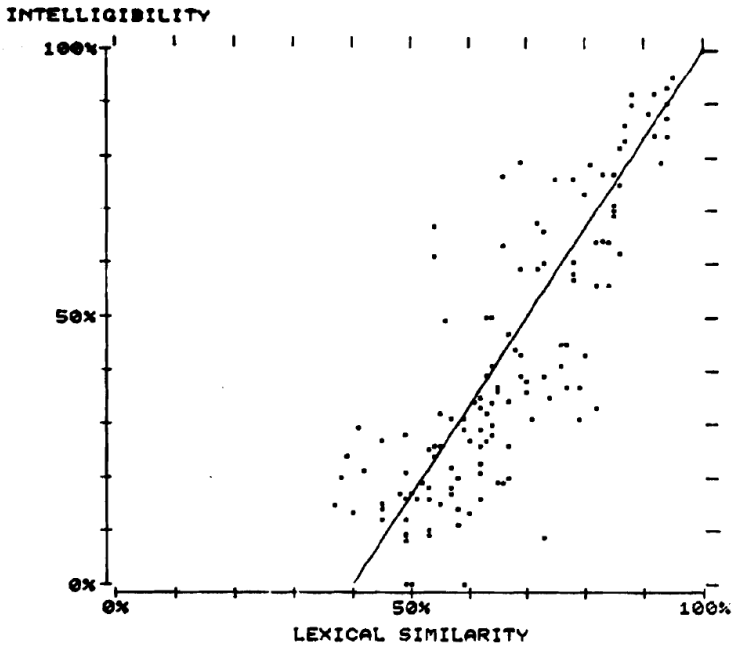
$$L_A \subsetneqq L_B(\langle w_1, \ldots, w_n \rangle) := \{w_i | L_A(w_i) \neq L_B(w_i)\}$$

§2: Notes on Yes/No Intelligibility

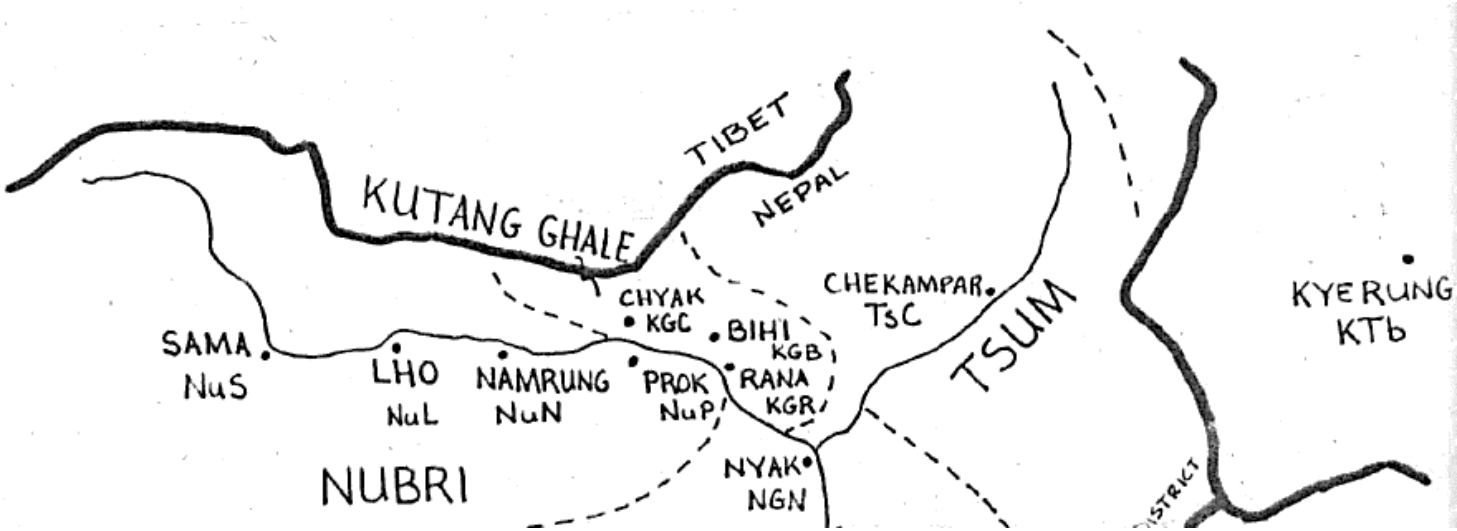
- The definition has no threshold value in it!
- You cannot do this trick for translating a gradient similarity to a yes/no relation on just any kind of object, cf. the situation in biology
- It only works if the objects are languages, i.e., objects that are built up from atomic parts that can be combined to describe the same atomic parts
- The MDP, crucially, makes sure that the first part of the intelligibility criterion namely $\exists \langle x \rangle L_A(\langle x \rangle) = L_A(w)$ is always inhabited
- Substitutability makes sure that understanding the parts, i.e., words/morphemes of the sentence is sufficient to understand the sentence itself
- One can not hope to ever be able to apply this defition in practice since it would require the complete description of a language

§2: Mutual Intelligibility in Practice

- One can do intelligibility surveys on the ground that perhaps approximate the theoretically existing binary intelligibility relation
- If no intelligibility surveys are available, one can measure lexical similarity on a standardized word list ("Swadesh lists") which correlates with the results of intelligibility surveys (Simons 1979:87-99, Korjakov 2017, etc.)
- 75% cognates on a 200-word Swadesh list is a ML estimate of an intelligibility border



§2: Example Discretizing Languages in Practice



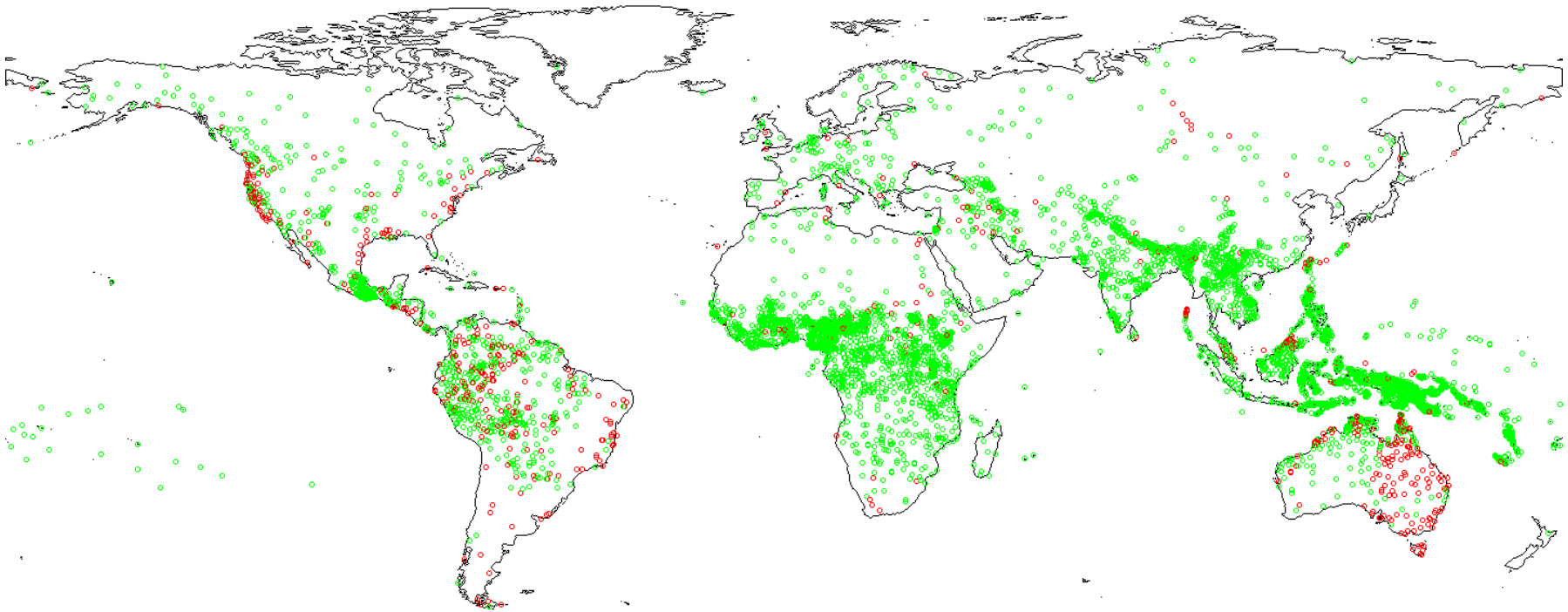
- Lexical similarity between villages (in %, 200-word Swadesh)

Lho						
93	Sama					
91	93	Namrung				
79	78	79	Prok			
77	76	78	71	Chekampar		
68	68	68	74	66	Kyerung	
63	63	60	64	60	59	Lhasa Tibetan

- Using 75% as the intelligibility border (Simons 1979:87-99)
- We get four languages: Lho-Sama-Namrung-Prok, Chekampar, Kyerung and Lhasa Tibetan

§3: MI-Languages: Lgs by Mutual (Un)intelligibility

There are approximately 6,500 attested spoken languages (plus 200 signed) that are (or were) mutually unintelligible



§3: Language Inventory Websites

Ethnologue: `http://www.ethnologue.com`

- Speaker numbers
- Detailed locations
- Further metadata
- Behind a paywall

Glottolog: `http://www.glottolog.org`

- Sources of data
- Principled classification
- Free and open

Both databases recognize something close to MI-languages but with deviations towards the political definition of language, and thus have inventories of around 7-8 000 languages.

§3: Spoken Language Inventory

- There are approximately 6 500 attested spoken languages that are (or were) *mutually unintelligible* with each other
- $\sim$ 1 000 of them are already extinct
- Notes on surveying:
  - As of 2026, the entire landmass has been surveyed for spoken languages at one time or another, with very few exceptions
  - The least well-surveyed areas include the northern and southern foothills of Indonesian Papua, the Nigeria-Cameroon borderland, the Javari river area (Brazil-Peru border area), pockets of the Democratic Republic of Congo and its border to Angola, the border area of Arunachal Pradesh (India) and China and the area around where Chad-Sudan-Central African Republic meet.
  - Many regions of the world are or were politically difficult to survey for western scholars and are thus known only mainly from older surveys, e.g., Myanmar and Libya.

§3: Coverage of the Language Inventory

- The language inventory as represented in the given numbers and databases is entirely dependent on there being a written record, if even by a traveller, for example
  - Many languages in the Amazon went extinct in the past few centuries for which we have scraps of data from travellers ascertaining their previous existence
  - For eastern Brazil, where the obliteration took place earlier, such information is much more scarce, leaving the list of assertable languages much shorter
  - But by analogy with the neighbouring regions, quite possibly in eastern Brazil there were many more that never made it into the written record
- Estimating languages in prehistory with a non-written record could be done
  - If a language is left alone it would take approximately 800 years for it to reach unintelligibility with its former self
  - 800 years is not a physical constant but some kind of average obtained from known cases (conditions under which it is faster or slower are known, cf. Bakker 2000).
- One could then take archaeological information and infer how many languages would have been spoken in any place in the past, and obtain a far higher number than 6 500 (Pagel 2000).

§3: Dynamics of the Language Inventory

- There are numbers, distribution and sometimes further information on peoples living without permanent contact with the outside world
  - one ethnic group in North Sentinel Island (Sarkar and Pandit 1994)
  - an unknown number in Indonesian Papua
  - some seventy ethnic groups in South America (Brackelaire and Azanha 2006)
- Languages completely new to the scientific community continue to be discovered every year, but these are typically languages spoken by a (usually aging) fraction of an ethnic group who otherwise speak a known language (that is how earlier surveys were never alerted to it)
- Apart from completely new languages, hundreds of revisions to the language inventory are made every year following newly collected information or more careful scrutiny of older data

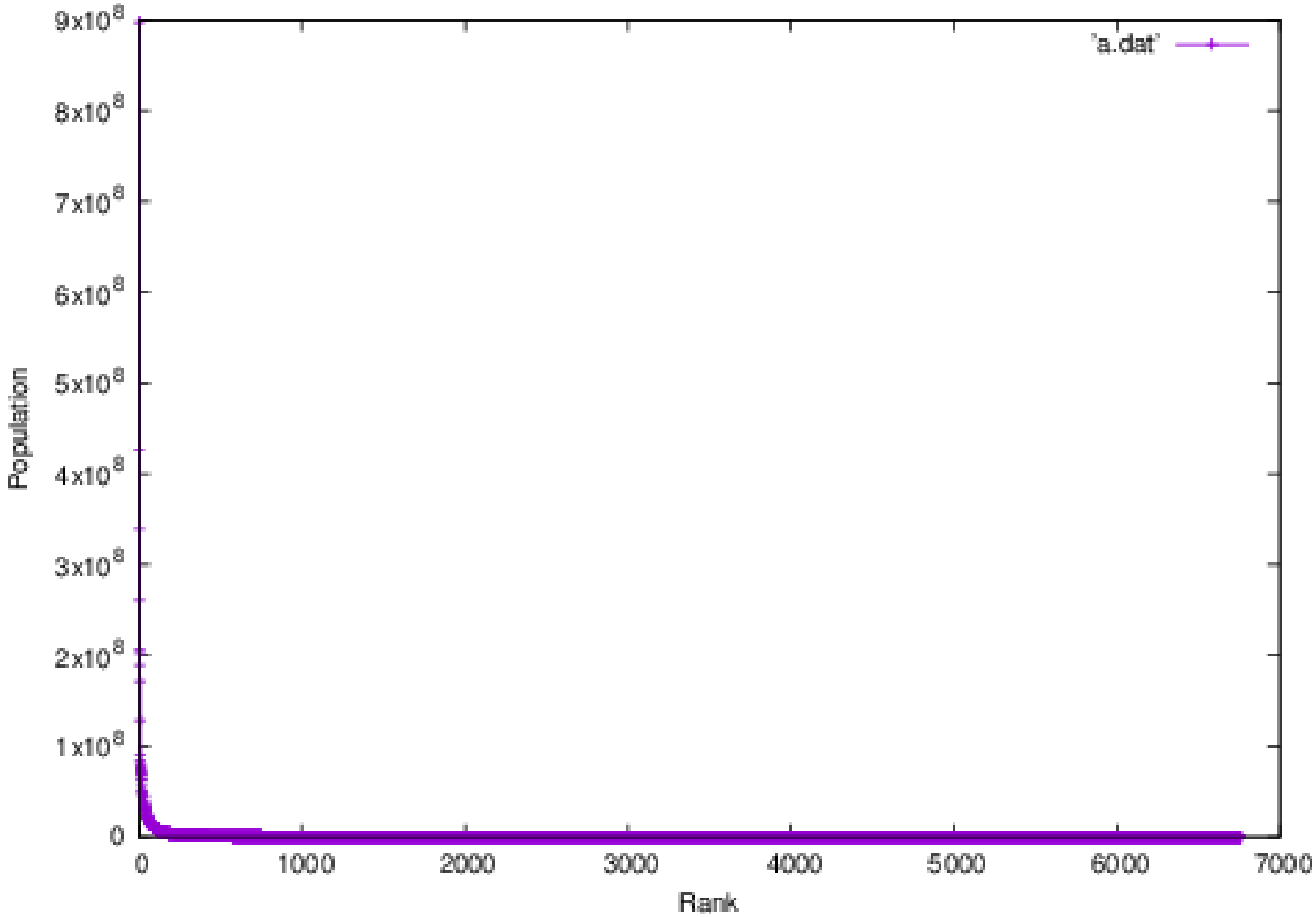
  More than 50,000 source documents underlie the language inventories of Ethnologue/Glottolog

§4: 15 Largest (L1 speakers, Ethnologue 27ed)

Rank	Language	ISO 639-3	Population
1	Mandarin Chinese	cmn	940 936 330
2	Spanish	spa	485 05 900
3	English	eng	380 196 920
4	Hindi	hin	345 088 150
5	Standard Arabic	arb	332 459 000
6	Bengali	ben	236 862 060
7	Portuguese	por	236 460 250
8	Russian	rus	147 566 020
9	Japanese	jpn	123 427 320
10	Yue Chinese	yue	86 133 890
11	Vietnamese	vie	85 429 000
12	Turkish	tur	84 077 680
13	Wu Chinese	wuu	83 397 340
14	Marathi	mar	83 247 270
15	Telugu	tel	82 795 890
…	…	…	…
105	Swedish	swe	10 048 870
…	…	…	…

§4: Most Languages are Small

- 94% of languages have a population of less than 1 000 000
- Median speaker number is 7 600 (50% have a higher population, 50% have a smaller population)



§4: 15 Smallest Languages

Language	ISO 639-3	Speakers
Amurdak	amg	1
Apiaká	api	1
Dhargari	dhr	1
Gagadu	gbu	1
Ganggalida	gcd	1
Gurdjar	gdj	1
Djeoromitxí	jbt	1
Kuyubi	-	1
Lake Miwok	lmw	1
Northeast Maidu	nmu	1
Northern Pomo	pej	1
Tagalaka	tgz	1
Tolowa-Chetco	tol	1
Taushiro	trr	1
Umotína	umo	1

These are all languages of bigger ethnic groups who have (almost completely) shifted language …

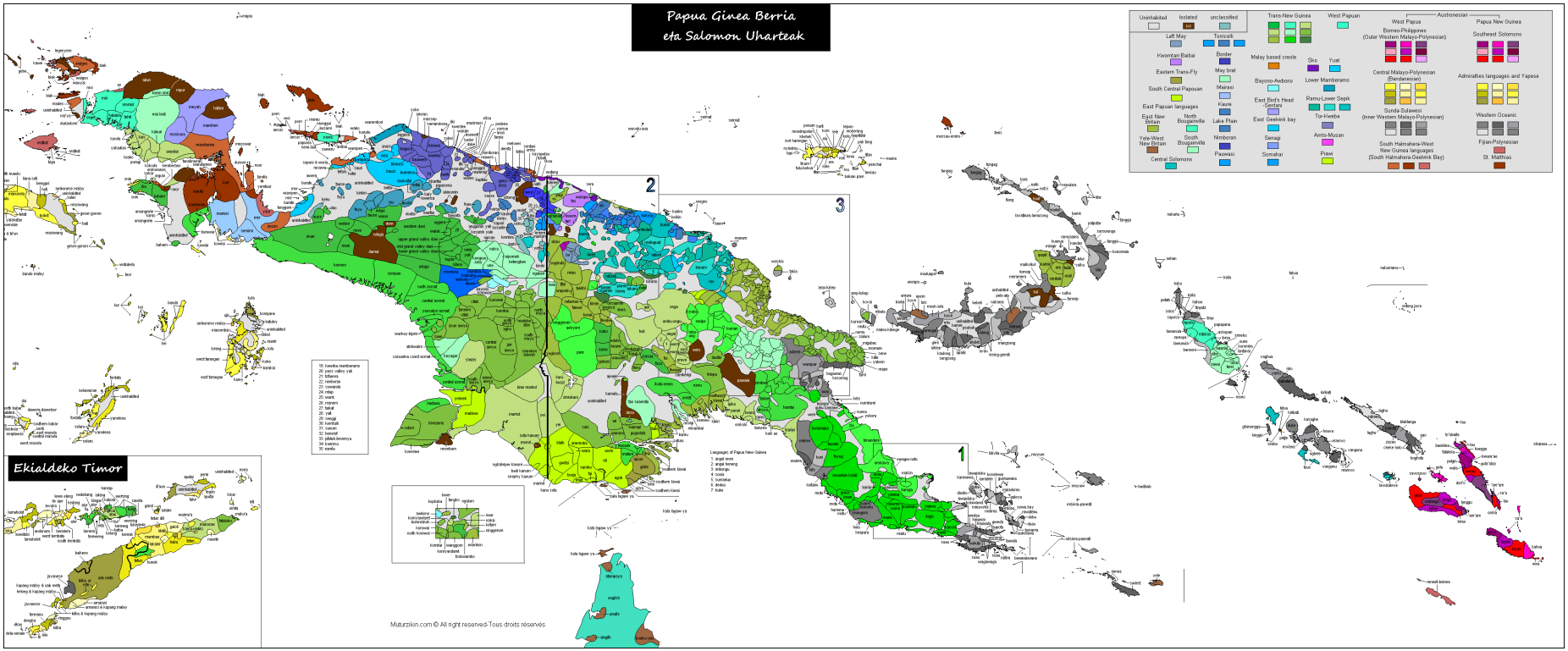
§4: How Small Can A Language Be?

- Many languages today have a small (< 200) speaker numbers, but nearly all of these reflect language endangerment
  - The ethnic group is bigger
  - There is poor intergenerational transmission
- World record holders in *stable* small speaker communities are Masep (~ 40 , Indonesian Papua), Moriori (~ 50, Indonesian Papua), Mor (~ 60, Indonesian Papua) and Burarra (~ 60, Australia)
- But speakers of these are all (at least) bilingual
- In the Amazon one can find monolingual communities with small speaker numbers (Zuruwaha ~ 140 speakers, Yorá ~ 170)
  - The communities are probably runaways from political turmoil in the rubber boom era ($\sim$ 1900)
  - Their monolingualism probably reflects their recent post-rubber boom situation, not a longer tradition of monolingualism

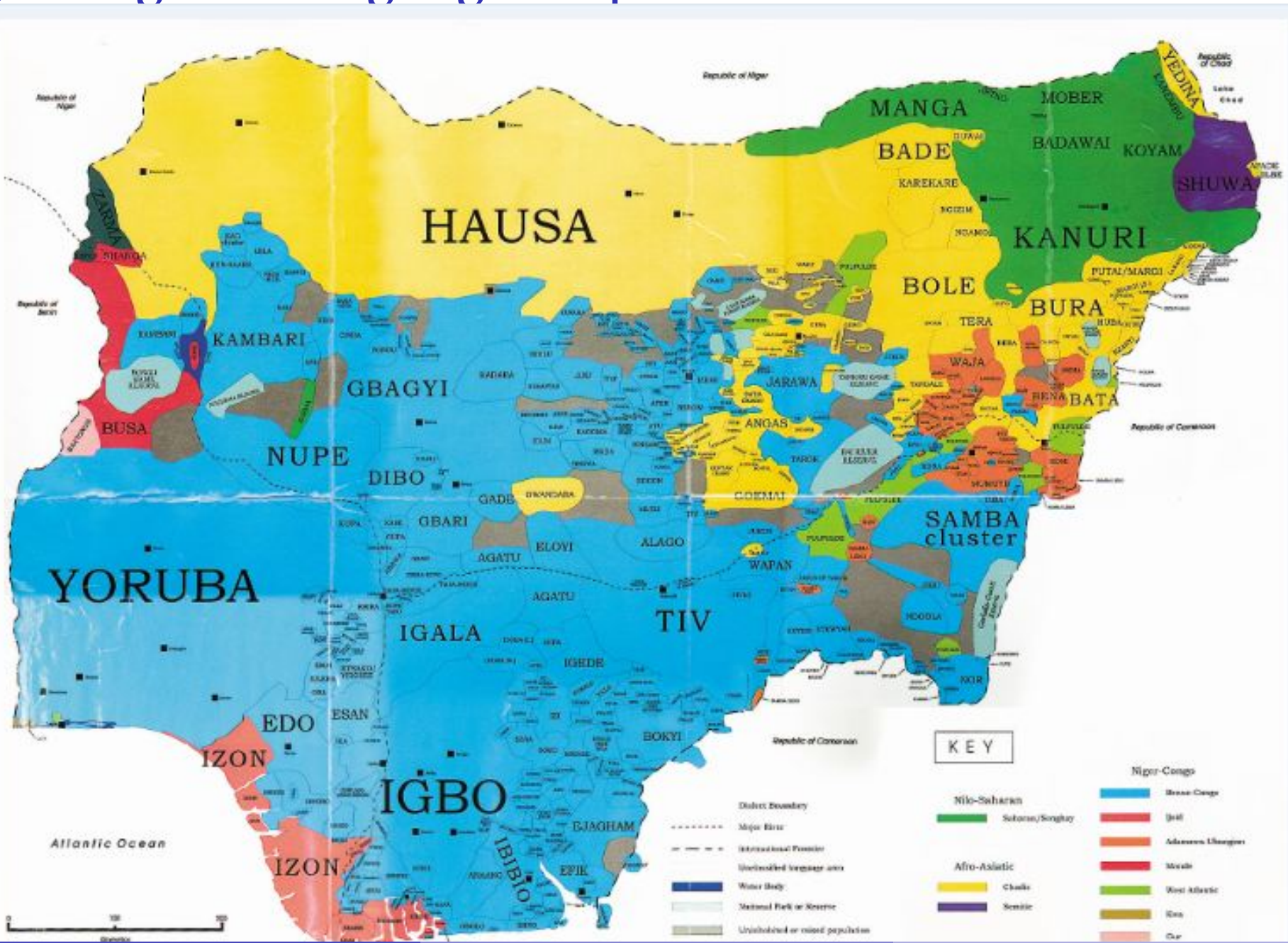
§4: Countries and Languages

Country	# lgs
Papua New Guinea	857
Indonesia (Papua)	714
Nigeria	532
India	423
Australia	312
China	339
Mexico	303
Cameroon	286
United States	268
Democratic Republic of the Congo	216
Philippines	185
Brazil	180
Russian Federation	152
Sudan	148
Malaysia	143
Chad	137
Vanuatu	126
Tanzania	122

§4: New Guinea Language Map



§4: Nigeria Language Map



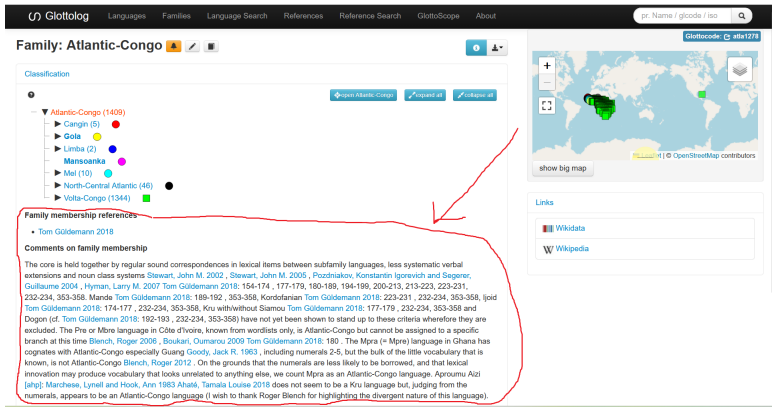
§5: Language Families

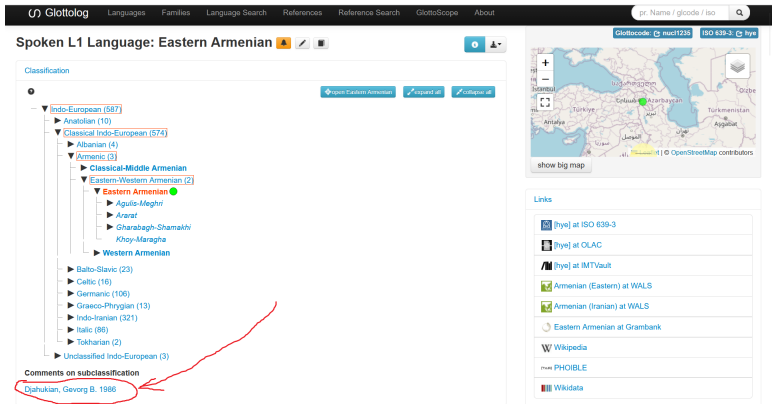
	Sanskrit	Greek	Latin
‘two’	*dvá*	*dýo*	*duo*
‘three’	*tráyas*	*treis*	*treːs*
…	…	…	…

The Sanscrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a **stronger affinity**, *both in the roots of verbs and the forms of grammar,* **than** *could possibly have been produced* **by accident**; *so strong indeed, that no philologer could examine them all three, without believing them to have* **sprung from some common source**, *which, perhaps,* **no longer exists**; *there is a similar reason, though not quite so forcible, for supposing that both the Gothic and the Celtic, though blended with a very different idiom, had the same origin with the Sanscrit; and the old Persian might be added to the same family (Jones 1786:422-423).*

§5: Language Family Trees

- The 7-8 000 languages fall into no less than 421 Language families
- Each family subdivides into sucessive subfamilies
- There are disagreements and unknowns
  - About top-level family membership
  - About subclassification
- Glottolog provides a scientific source for each judgment for all to scrutinize, see `https://glottolog.org/glottolog/glottologinformation`





Anderson, S. R. (2005). How many languages are there in the world? Answer to a FAQ by the Linguistic Society of America, Washington, D.C. Published on the web as `http://www.lsadc.org/pdf_files/howmany.pdf`. Accessed 1 September 2005.

Bakker, P. (2000). Rapid language change: Creolization, intertwining, convergence. In Renfrew, C., McMahon, A., and Trask, R. L., editors, Time depth in historical linguistics, volume 2 of Papers in the prehistory of languages, pages 585–620. Cambridge: McDonald Institute for Archaeological Research.

Beijering, K., Gooskens, C., and Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distance by means of the levenshtein algorithm. In van Koppen, M. and Botma, B., editors, Linguistics in the Netherlands, volume 25, pages 13–24. Amsterdam: John Benjamins.

Blust, R. A. (2001). Language, dialect, and riotous sound change: the case of sa'ban. In Thurgood, G. W., editor, Papers from the Ninth Annual Meeting of the Southeast Asian Linguistics Society, pages

249–360. Arizona State University, Program for Southeast Asian Studies.

Brackelaire, V. and Azanha, G. (2006). Últimos pueblos indígenas aislados en américa latina: Reto a la supervivencia. In Lenguas y tradiciones orales de la Amazonía. ¿diversidad en peligro?, pages 313–367. La Habana: Casa de las Américas, La Habana.

Casad, E. (1974). Dialect intelligibility testing. Norman, Oklahoma: SIL.

Comrie, B., editor (1987). The World's Major Languages. New York: London: Croom Helm, New York.

Egland, S., Bartholomew, D., and Ramos, S. C. (1983). La inteligibilidad interdialectal en México: resultados de algunos sondeos. México: ILV.

Garey, M. R. and Johnson, D. S. (1979). Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman, New York.

Gartner, B. and Streiter, O. (2006). Smart messages: The whistled languages of la gomera (spain), antia (greece) and kuşköy (turkey) –

state of research and open questions. In Abel, A., Stuflesser, M., and Putz, M., editors, Mehrsprachigkeit in Europa: Erfahrungen, Bedürfnisse, Gute Praxis. Tagungsband. Bozen: Eurac.

Gooskens, C. (2013). Methods for measuring intelligibility of closely related language varieties. In Bayley, R., Cameron, R., and Lucas, C., editors, Handbook of sociolinguistics, pages 195–213. Oxford University Press.

Hammarström, H. (2008). Counting languages in dialect continua using the criterion of mutual intelligibility. Journal of Quantitative Linguistics, 15(1):34–45.

Heine, B. and Nurse, D. (2000). Introduction. In Heine, B. and Nurse, D., editors, African Languages: An Introduction, pages 1–10. Cambridge: Cambridge University Press.

Hockett, C. F. (1960). The origin of speech. Scientific American, 203:88–111.

Jones, W. (1798 [1786]). Third anniversary discourse, delivered 2 february, 1786. Asiatic researches: Or, Transactions of the society instituted in Bengal, for inquiring into the history and antiquities, the arts, sciences, and literature, of Asia, 1:415–431.

Kendon, A. (1988). Sign languages of Aboriginal Australia: cultural, semiotic and communicative perspectives. Cambridge: Cambridge University Press.

Korjakov, Y. B. (2017). Problema ”jazyk ili dialekt” i popytka leksikostaticeskogo podxoda. Voprosy Jazykoznaniya, 2017(6):79–101.

Lynch, J. and Crowley, T. (2001). Languages of Vanuatu: A New Survey and Bibliography, volume 517 of Pacific Linguistics. Canberra: Research School of Pacific and Asian Studies, Australian National University, Canberra.

Meyer, J. (2015). Whistled Languages: A Worldwide Inquiry on Human Whistled Speech. Berlin: Springer.

Milliken, M. E. and Milliken, S. R. (1996). System relationships in assessing dialect intelligibility. Notes on Linguistics, 72:15–31.

Pagel, M. (2000). The history, rate, and pattern of world linguistic evolution. In Knight, C., Studdert-Kennedy, M., and Hurford, J., editors, The Evolutionary Emergence of Language, pages 391–416. Cambridge: Cambridge University Press.

Sarkar, J. K. and Pandit, T. N. (1994). Sentinelese. In Pandit, T. N. and Sarkar, B. N., editors, Andaman and Nicobar Islands, volume XII of People of India, pages 184–187. Madras: Anthropological Survey of India.

Simons, G. F. (1979). Language Variation and Limits to Communication, volume 3 of Working Papers for the Language Variation and Limits to Communication Project. [Ithaca]: Cornell University and Summer Institute of Linguistics.

Stern, T. (1957). Drum and whistle "languages": An analysis of speech surrogates. American Anthropologist, 59(3):487–506.

Thierry, . (2002). Les langages sifflés. Master's thesis, Paris: École Pratique des Hautes Études.

Voegelin, C. F. and Harris, Z. S. (1951). Methods for determining intelligibility among dialect of natural languages. Proceedings of the American Philosophical Society, 95(3):322–329.

Welmers, W. E. (1971). Checklist of african language and dialect names. In Sebeok, T. A., editor, Linguistics in Sub-Saharan Africa, volume 7 of Current Trends in Linguistics, pages 761–899. Berlin: Mouton de Gruyter.