BUILDING A MODEL FOR A PHYLOLINGUISTIC INFERENCE OF GRAMMATICALIZATION

GERD CARLING, NOUR EFRAT-KOWALSKY, GOETHE UNIVERSITY, FRANKFURT AM MAIN

NEW ADVANCES IN PHYLOLINGUISTICS, MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY, LEIPZIG

11-12 JUNE, 2025

THE PROBLEM:
WHAT IS GRAMMATICALIZATION AND WHY IS IT IMPORTANT?

- ORIGINAL   I am going to the church
- SYNTACTIC REANALYSIS   I am going to tell you the truth   = **semantic bleaching**
- GRAMMATICALIZATION   I'm gonna tell you the truth   = **phonological erosion**

The grammaticality cline:

Content ———————————————————————————————————→ Function

Meillet (1912), Hopper & Traugott (2003), Haspelmath (2004)

THE EVOLUTIONARY ASPECT OF GRAMMATICALIZATION

- Grammaticalization is considered to be one of the most important processes of language evolution (Heine & Kuteva 2011)

- Grammaticalization is marked by a predictable directionality

- As such, it is used to reversely reconstruct the trajectory of language evolution (Heine & Kuteva 2007)
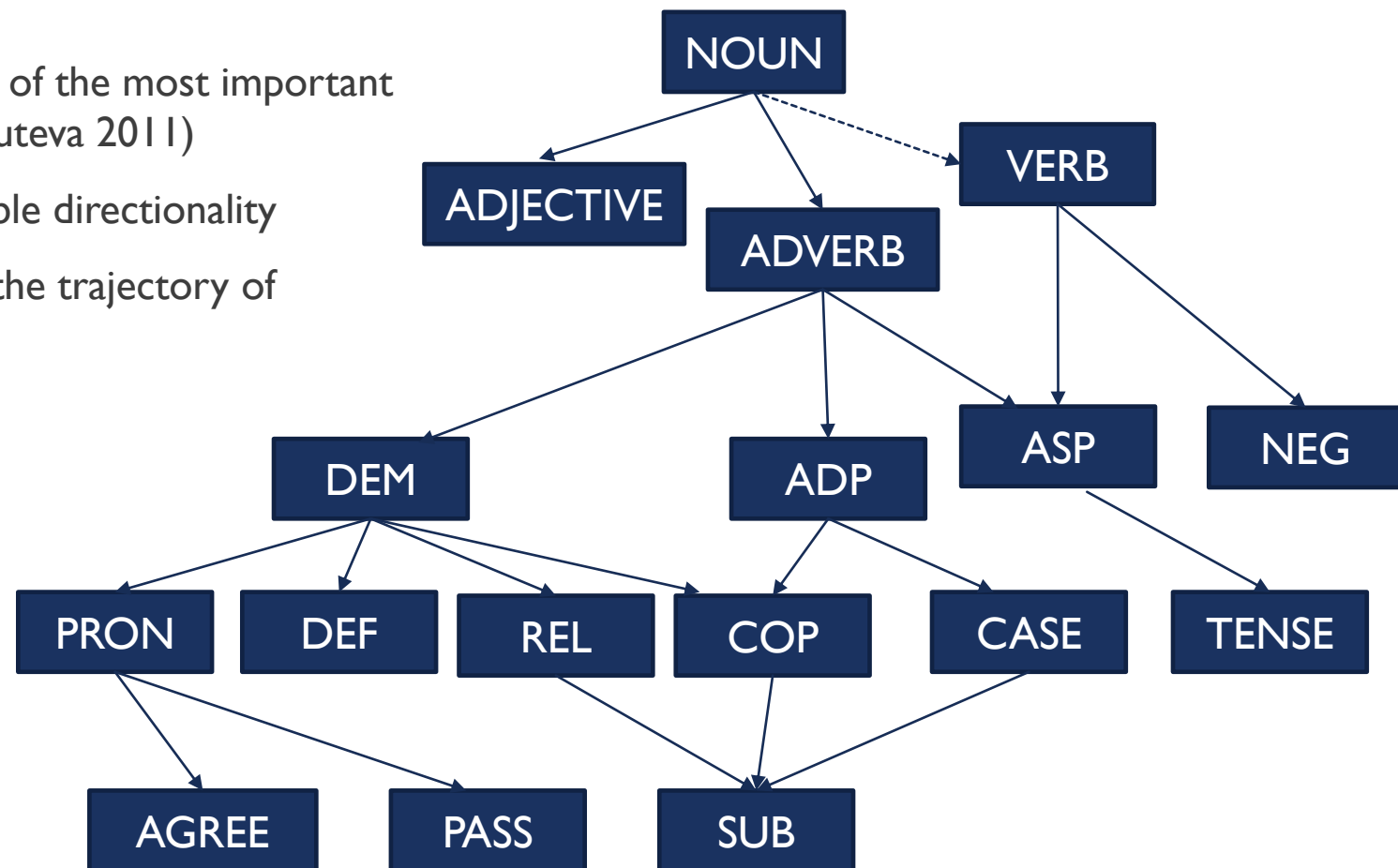
Language evolution by reverse reconstruction of grammaticalization (Heine & Kuteva 2007:111)

THEORIES ABOUT THE EVOLUTIONARY ASPECTS OF GRAMMATICALIZATION

- We know that grammaticalization increases the frequency of use of a lexeme (Bybee 2006, 2007, Rosemeyer 2016).

- Frequency has a conserving effect, as it stalls changes in the distribution of a construction (Bybee 2007)

FUNCTION WORD

CONTENT WORD   CONTENT WORD
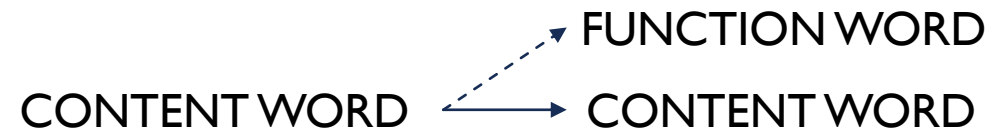
RESEARCH QUESTIONS

- **Conditions for grammaticalization:**
- Which words grammaticalize?
- Can any word grammaticalize?

- **Diversity of grammaticalization:**
- Do some languages grammaticalize more than others?

- **Diachrony of grammaticalization:**
- Can languages be in different stages of grammaticalization?
- What are the evolutionary dynamics of grammaticalization?

HYPOTHESES CONCERNING EVOLUTION:

1. Increased productivity of grammaticalized items → **higher change rate**

2. Frequency has a conserving effect → **lower change rate**

HOW CAN THE TWO HYPOTHESES BE TESTED BY A PHYLOLINGUISTIC MODEL?
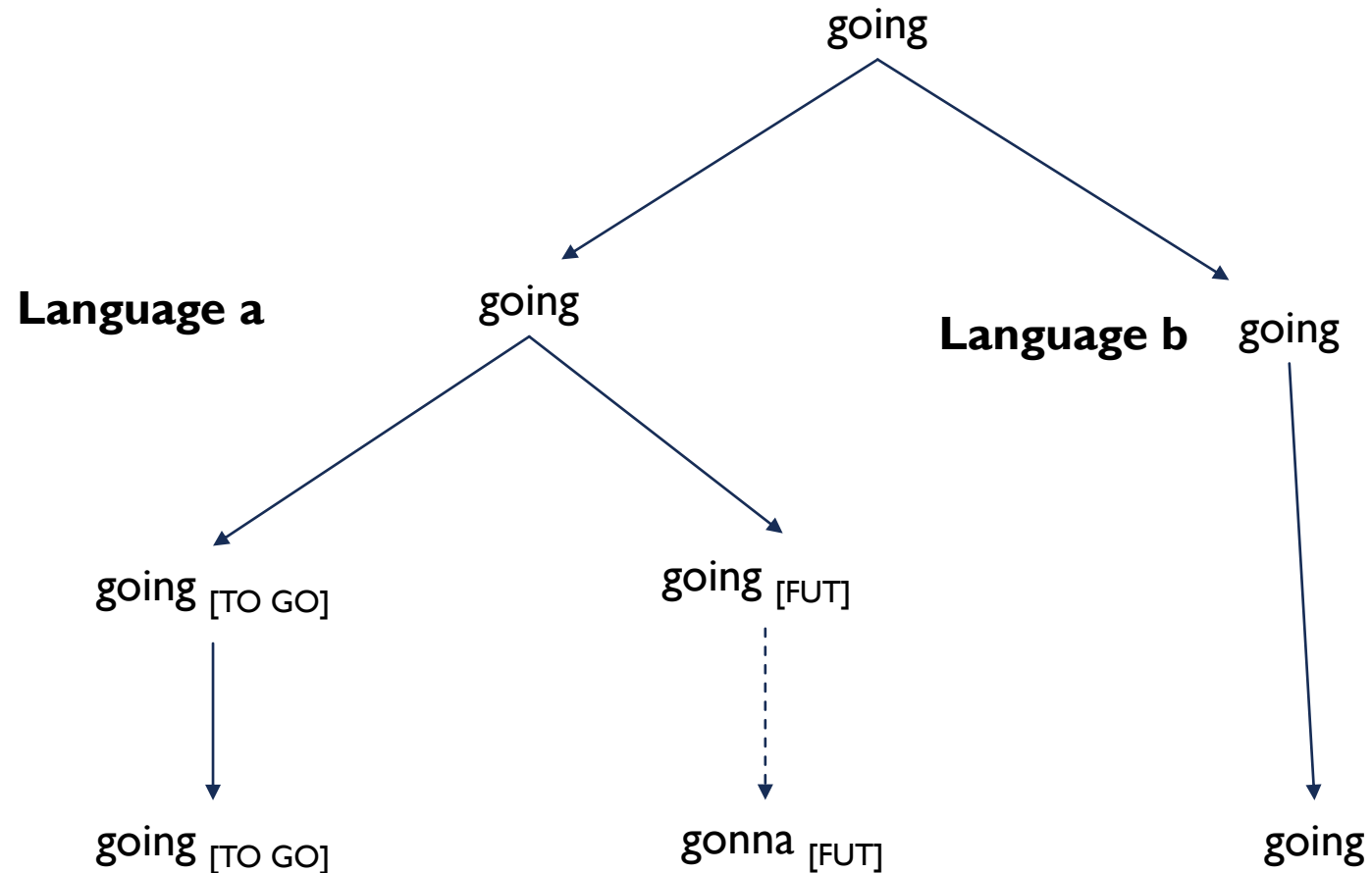
STEP 1: LEXICAL SEMANTICS OF GRAMMATICALIZATION

- Build a grammaticalization database, based on

- Lexical and functional concepts of attested grammaticalization events
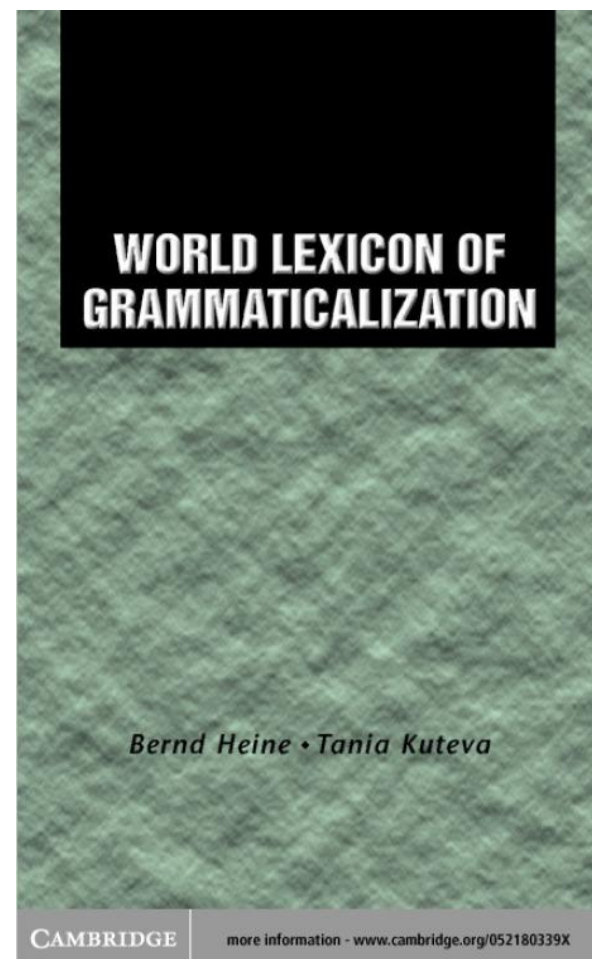
COLUMNS:
SOURCE (CONTENT)
SOURCE (FUNCTION)
SOURCE WORD CLASS
SOURCE NOUN
TARGET (CONTENT)
TARGET (FUNCTION)
TARGET WORD CLASS

WORLD LEXICON OF GRAMMATICALIZATION

Bernd Heine • Tania Kuteva

CAMBRIDGE   more information - www.cambridge.org/052180339X

THE ORGANIZATION OF THE LEXICON

- The lexicon is based on a data set of typologically defined grammaticalization events, each with examples from various languages.

- The grammaticalization events are supported by data from more, unrelated languages

ALLATIVE > (2) DATIVE
Tamil *-iṭam* 'to' (directional bound postposition) > bound postposition marking the indirect object. Ex.

Tamil (T. Lehmann 1989: 41)

kumaar	raajaa·v-iṭam		oru	pustakam
Kumar	Raja-	LOC	a	book

koṭu-tt-	aa<u>n</u>.
give- PAST-3:SG:M	

'Kumar gave Raja a book.'

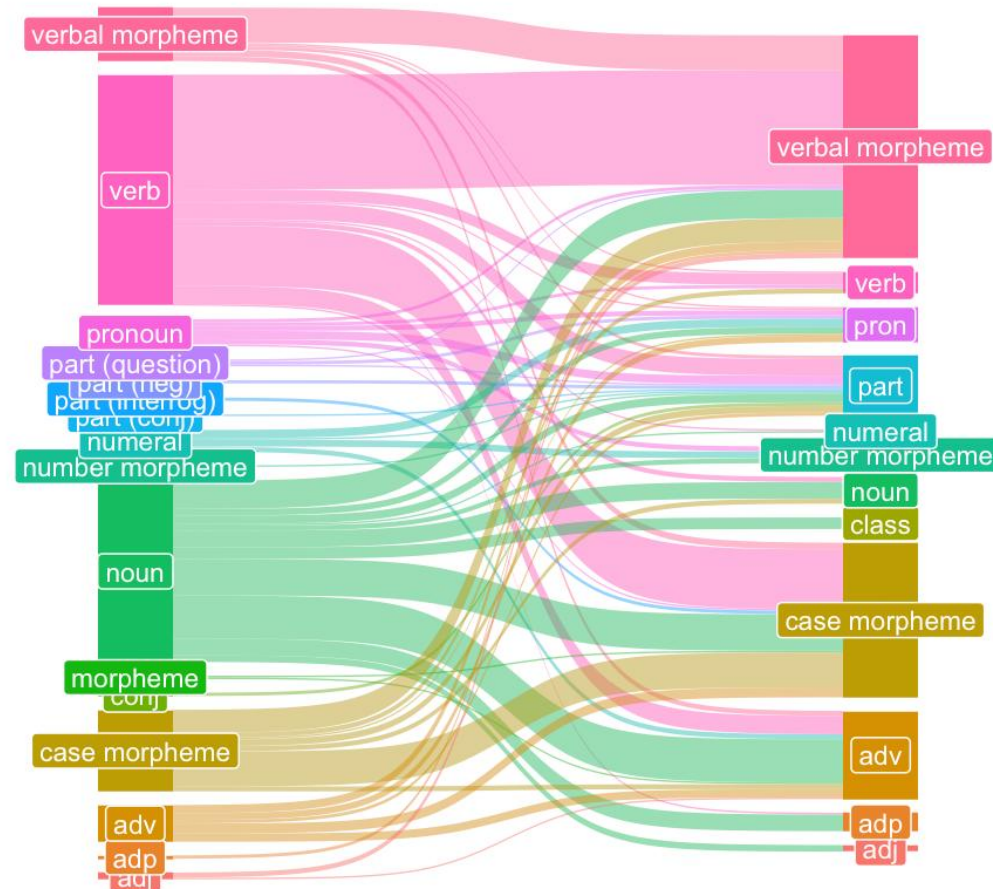# ORGANIZATION OF THE GRAMMATICLIZATION DATABASE

- **COLUMNS:**
- SOURCE (CONTENT)
- SOURCE (FUNCTION)
- SOURCE WORD CLASS
- SOURCE NOUN: SEMANTIC CLASS (taken from semantic class noun class/classifier/gender data)
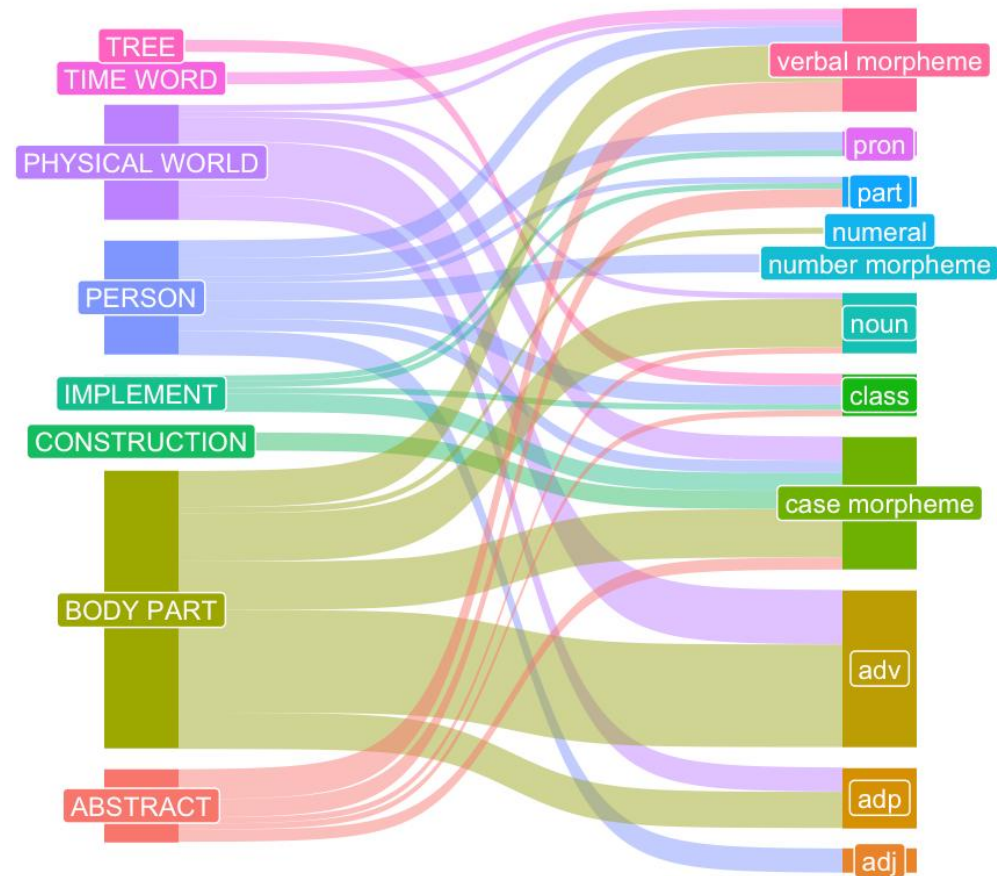
- TARGET (CONTENT)
- TARGET (FUNCTION)
- TARGET WORD CLASS

RESULTS: SANKEY DIAGRAM OF SOURCE AND TARGET WORD CLASS IN GRAMMATICALIZATION EVENTS
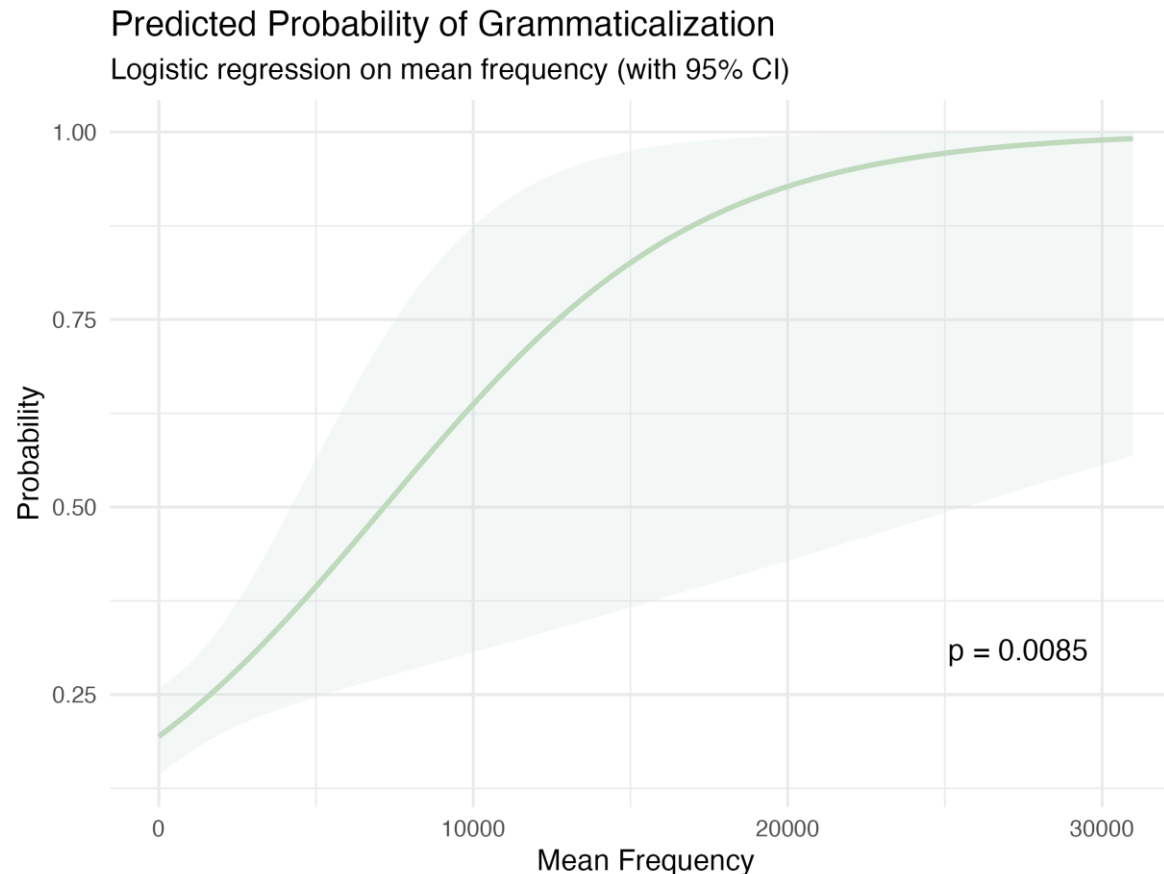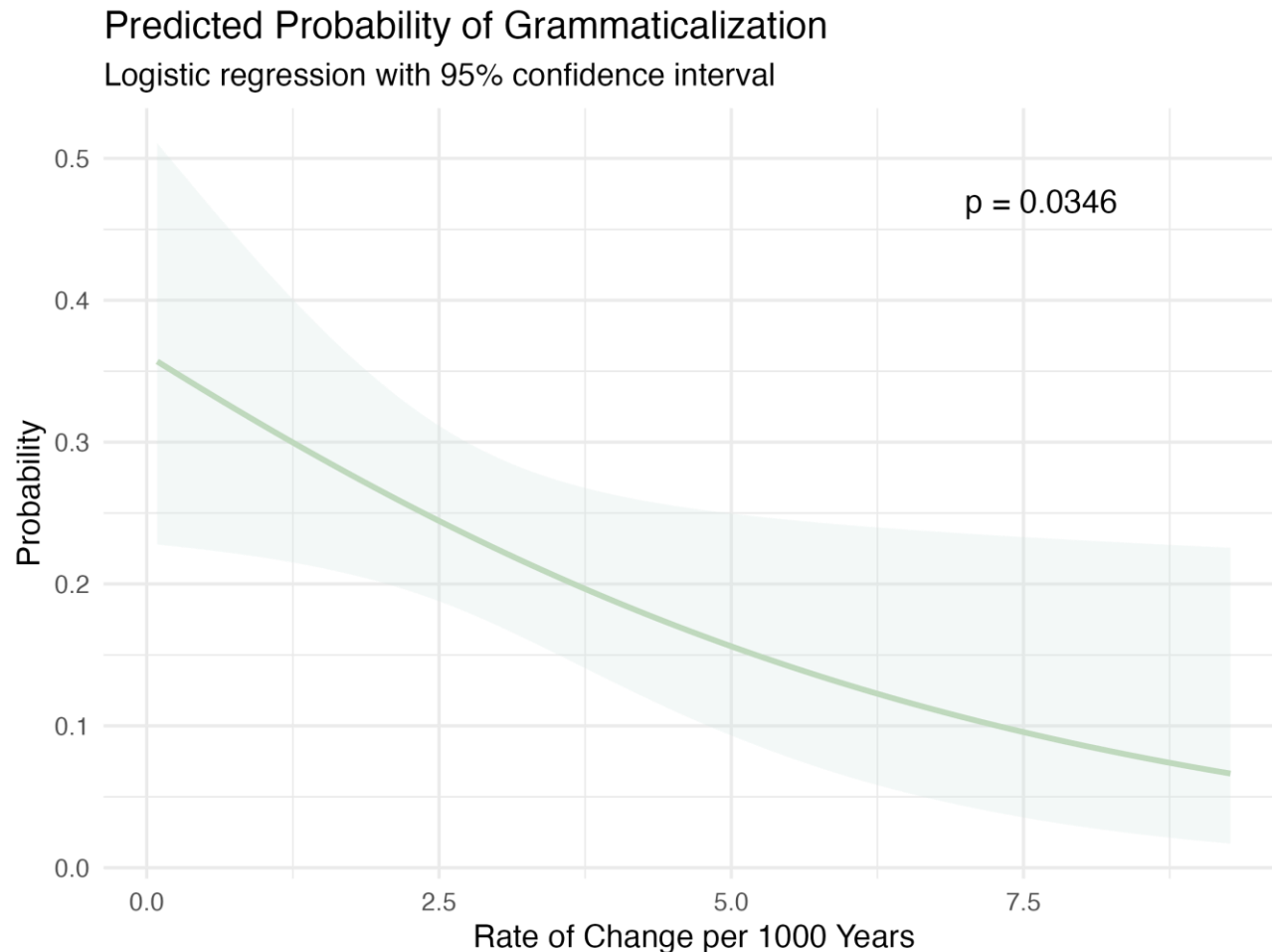
Word Class Grammaticalization

Noun Semantic Class Grammaticalization

POSSIBLE PREDICTORS OF GRAMMATICALIZATION:
FREQUENCY

Predicted Probability of Grammaticalization
Logistic regression on mean frequency (with 95% CI)


p = 0.0085

- Predicted probabilities of grammaticalization against Frequency (Source words, mean of frequency of English, Spanish, Russian, and Greek)
- Based on a data set of 200 Swadesh words (46/200 Grammaticalization source words), from Indo-European
- Data from Pagel et al (2007)

POSSIBLE PREDICTORS OF GRAMMATICALIZATION:
LEXICAL SUBSTITUTION RATES

Predicted Probability of Grammaticalization
Logistic regression with 95% confidence interval

- Predicted probabilities of grammaticalization (Source words) against lexical substitution rates
- Based on a data set of 200 Swadesh words (46/200 Grammaticalization source words)
- Data from Pagel et al (2007)

PREDICTED PROBABILITIES, RATES FROM FREQUENCY

Logistic Regression Model Predictions for Grammaticalization
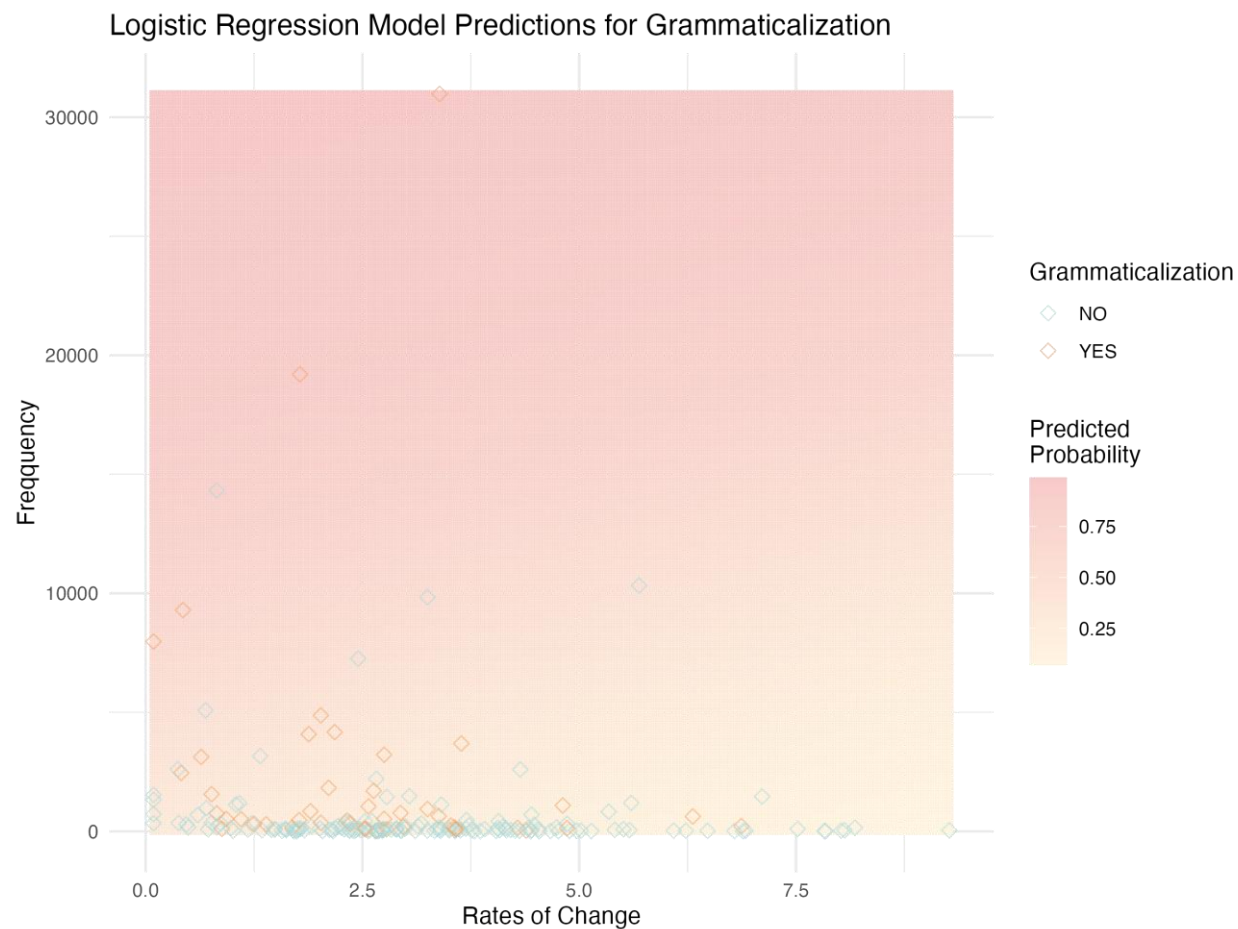
- Heatmap of predicted probabilities of grammaticalization (Source words) in relation to Rates of change and Frequency
- Data from Pagel et al (2007)

☞ A frequent word is more likely to grammaticalize, regardless of its rate of change

CONCLUSION

- Rate of change explains the likelihood for a lexeme (content, function) to be source of grammaticalization

- Frequency explains even better the likelihood for a lexeme (content, function) to be source of grammaticalization

- Frequency also explains the propensity for colexification (Xu et al 2020)

- Which means that:

- Colexification and grammaticalization may be influenced by similar conditions (colexification lacks the process of reanalysis and semantic bleaching)

PRECONDITIONS FOR A MODEL

FREQUENCY

Slower Rates of Change [A → B]

Higher Grammaticalization [A → A, a]

Higher Colexificaiton [A → A, B]
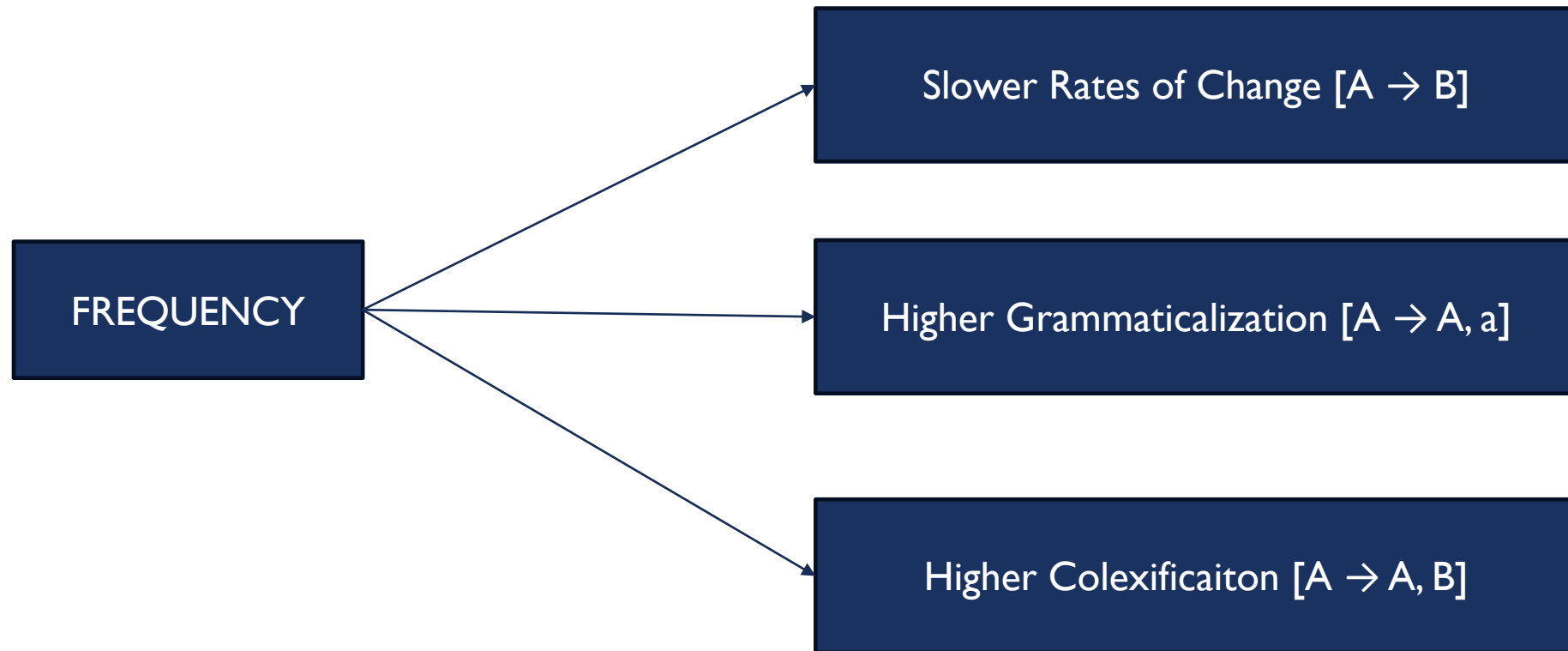
WHAT DO RATES OF CHANGE MEASURE?

- Rates of change inform us about the likelihood for the semantic prototype to be substituted, which is not what grammaticalization is

How frequently [1 → 0] and vice versa

GRAMMATICALIZATION IS MORE COMPLEX:
POSSIBLE MODEL FOR PRINCIPLES OF LEXICAL CHANGE

- Frquent words duplicate themselves by colexification and grammaticalization, but stay unchanged

- Infrequent words change
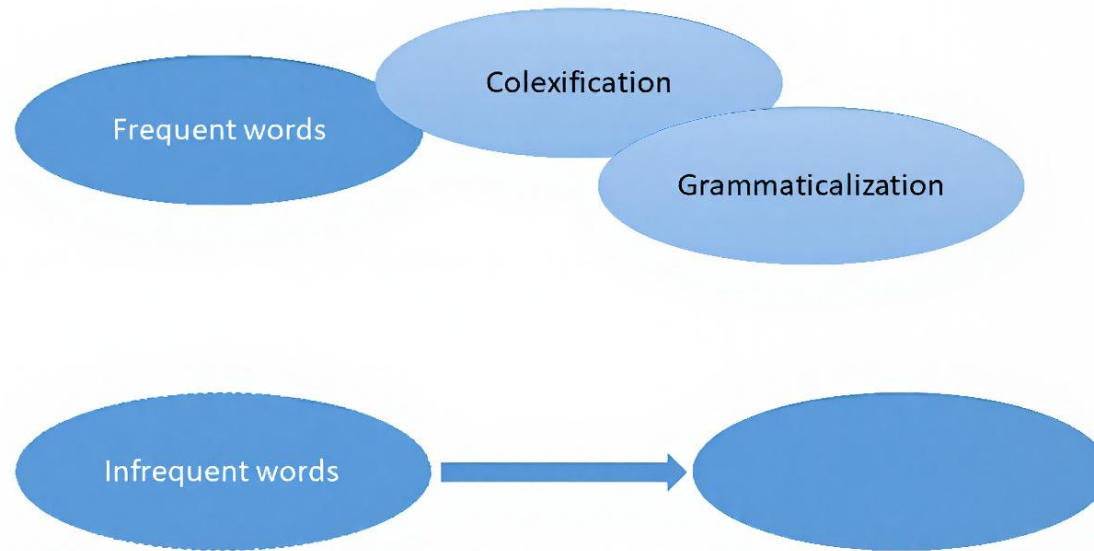
This article is part of the Research Topic

The Evolution of Meaning: Challenges in Quantitative Lexical Typology
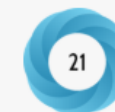
View all 9 articles >

The evolution of lexical semantics dynamics, directionality, and drift

Gerd Carling[1,2*]   Sandra Cronhamn[1]   Olof Lundgren[1]

Victor Bogren Svensson[1]   Johan Frid[3*]

[1] Centre for Languages and Literature, Lund University, Lund, Sweden
[2] Institute for Empirical Linguistics, Goethe University, Frankfurt, Germany
[3] Lund University Humanities Lab, Lund University, Lund, Sweden

BASIC IDEA

- Lexical dataset, coded for cognacy (7 Eurasian families)

- Inclusion of colexified and changed meaning of lexemes, i.e., "Etymologies"

- Phylogenetic comparative model, probabilistic reconstruction of each meaning separately at each internal node

- Markov process (Maslova 2000) + Glottolog trees

- Extracting the average gain and loss rates of meanings and concepts

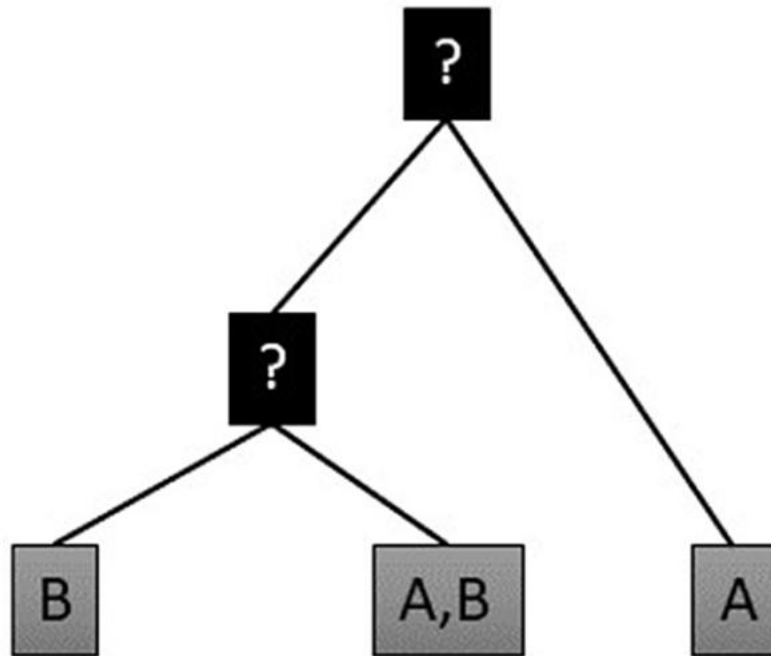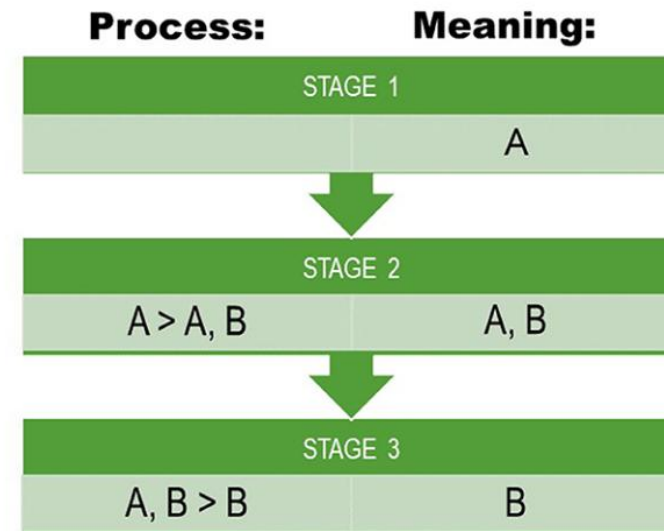# SCHEMATIC MODEL FOR THE PHYLOGENETIC RECONSTRUCTION

Figure 2

Figure 1

Figure 1. A schematic model evolution used in this paper. meaning(s) of lexemes in our hypothesized process at hidd attested meanings at attested 2013, p. 232–238).

POSSIBLE MODEL FOR A PHYLOLINGUISTIC MODEL OF GRAMMATICALIZATION 1: AUTOMATED MODEL (INDO-EUROPEAN)

SOURCE concepts:
1000 concept list
Lexemes analyzed by IPA
+ syllable structure

Typological data set of grammaticalization events

TARGET concepts:
Database of grammatical affixes
(Restriction?)

Distance model

Distance model

Phonemic similarity between Source and Target =

Phylogenetic inference
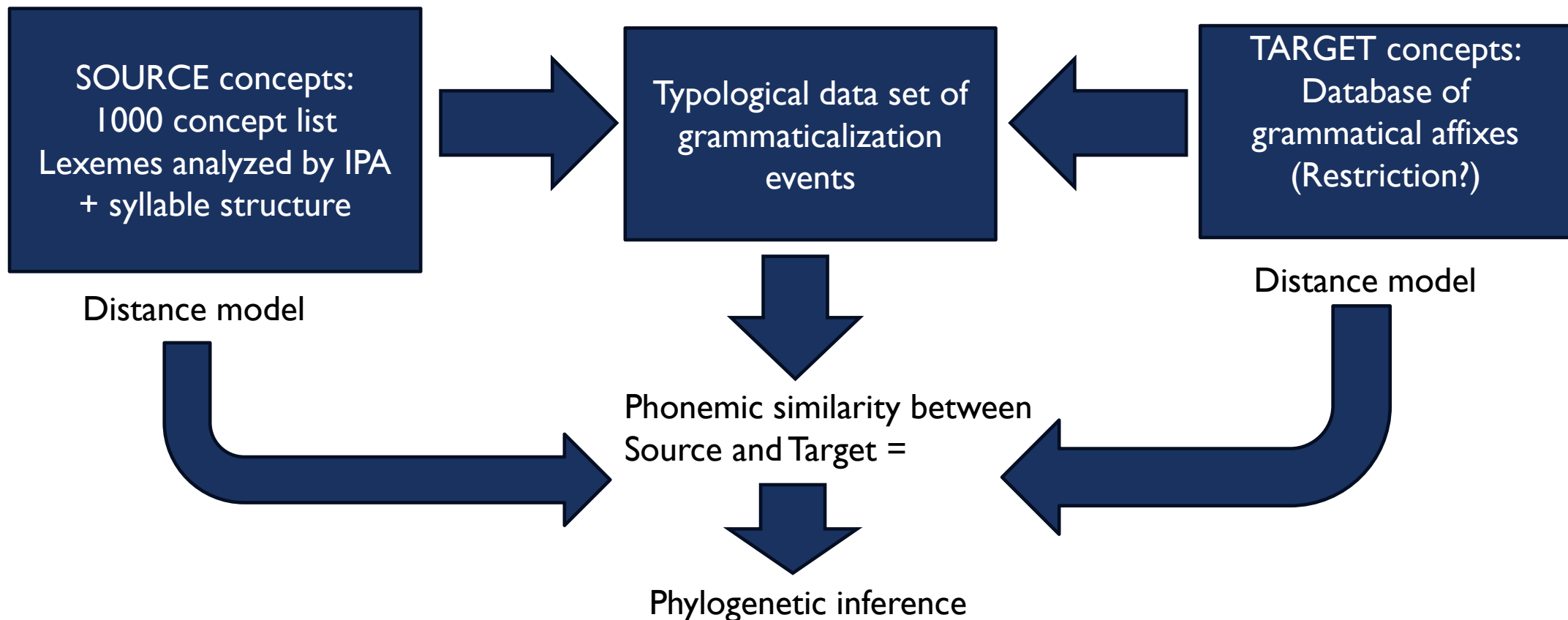
POSSIBLE MODEL 1: AUTOMATED

- **Disadvantages:**

- Innacurate and crude

- Many potential sources of error

- Much noise

- Only ending up with events leading from a lexeme to an affix (=much missing data)

- **Advantages:**

- We can test it without additional (complicated) coding

- Possibility to scale up data

POSSIBLE MODEL: 2. MANUAL CODING MODEL

200/400 Swadesh list (for Indo-European)

Code source words (grammaticalizing y/n) + target lexemes/content words

Binary file

Grammaticalization database with real data

Phylogenetic comparative model inferring events (based on a phylogenetic tree)

Advantage:

More accurate and reflecting grammaticalization events more completely

Disadvantage:

Very complex, manual coding procedure

PRELIMINARY CONCLUSION

- Grammaticalization is restricted to a specific number of source and target words (function and content morphemes)

- Among nouns, specific semantic classes dominate as source words

- Grammaticalization is restricted to a specific number of events

- Grammaticalization source words can be predicted by frequency (in Indo-European)

- Grammaticalization can also be predicted by lexical substitution rates (but more so by frequency)

- Therefore, it is likely that more frequent words remain stable but duplicate themselves by grammaticalization and colexification, whereas infrequent words change

- A phylogenetic inference of grammaticalization can be done either by an automated or a manual model

THANKS FOR THE ATTENTION!

All papers (including links to data and code) are available at:
https://gerdcarling.se/publications1/historical-linguistics-and-phylogenetics

Bybee, Joan. 2007. "Word Frequency in Lexical Diffusion and the Source of Morphophonological Change." In *Frequency of Use and the Organization of Language*, 1-18. Oxford University Press.

Haspelmath, Martin. 2004. "On directionality in language change with particular reference to grammaticalization." In *Up and down the Cline - The Nature of Grammaticalization*, edited by Olga Fischer, Muriel Norde and Harry Perridon. Amsterdam - Philadelphia: John Benjamins.

Heine, Bernd, and Tania Kuteva. 2007. *The genesis of grammar: a reconstruction*. Oxford - New York: Oxford University Press.

Heine, Bernd, and Tania Kuteva. 2011. "Grammaticalization theory as a tool for reconstructing language evolution." In *Oxford Handbook of Language Evolution*, edited by Maggie Tallerman Kathleen R. Gibson, 512-527. Oxford: Oxford University Press.

Hopper, Paul, and Elizabeth C. Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Meillet, Antoine. 1912. "L'évolution des formes grammaticales." *Scientio (Reivsta de scienza)* XII(XXVI, 6).

Pagel, Mark, Quentin D. Atkinson, and Andrew Meade. 2007. "Frequency of word-use predicts rates of lexical evolution throughout Indo-European history." *Nature* 449:717. doi: 10.1038/nature06176

https://www.nature.com/articles/nature06176#supplementary-information.

Rosemeyer, Malte. 2016. "Modeling frequency effects in language change." In *Experience Counts: Frequency Effects in Language*, edited by Behrens Heike and Pfänder Stefan, 175-208. Berlin, Boston: De Gruyter.