The Grammar(s) of Code-Mixing:
A corpus-based approach to multilingual first language acquisition

Antje Endesfelder Quick
University of Leipzig

Stefan Hartmann
University of Düsseldorf

Nikolas Koch
LMU Munich

Code-Mixing

und dann magic air
'And then magic air'

Look at the Ampel, it's kaput
'Look at the traffic light, it's broken'

We have to beeil
'We have to hurry up'

Code-Mixing = one of the more salient phenomena found in bilingualism
→ use of two languages in one utterance

Recent years have seen increased interest in code-mixing from a usage-based perspective

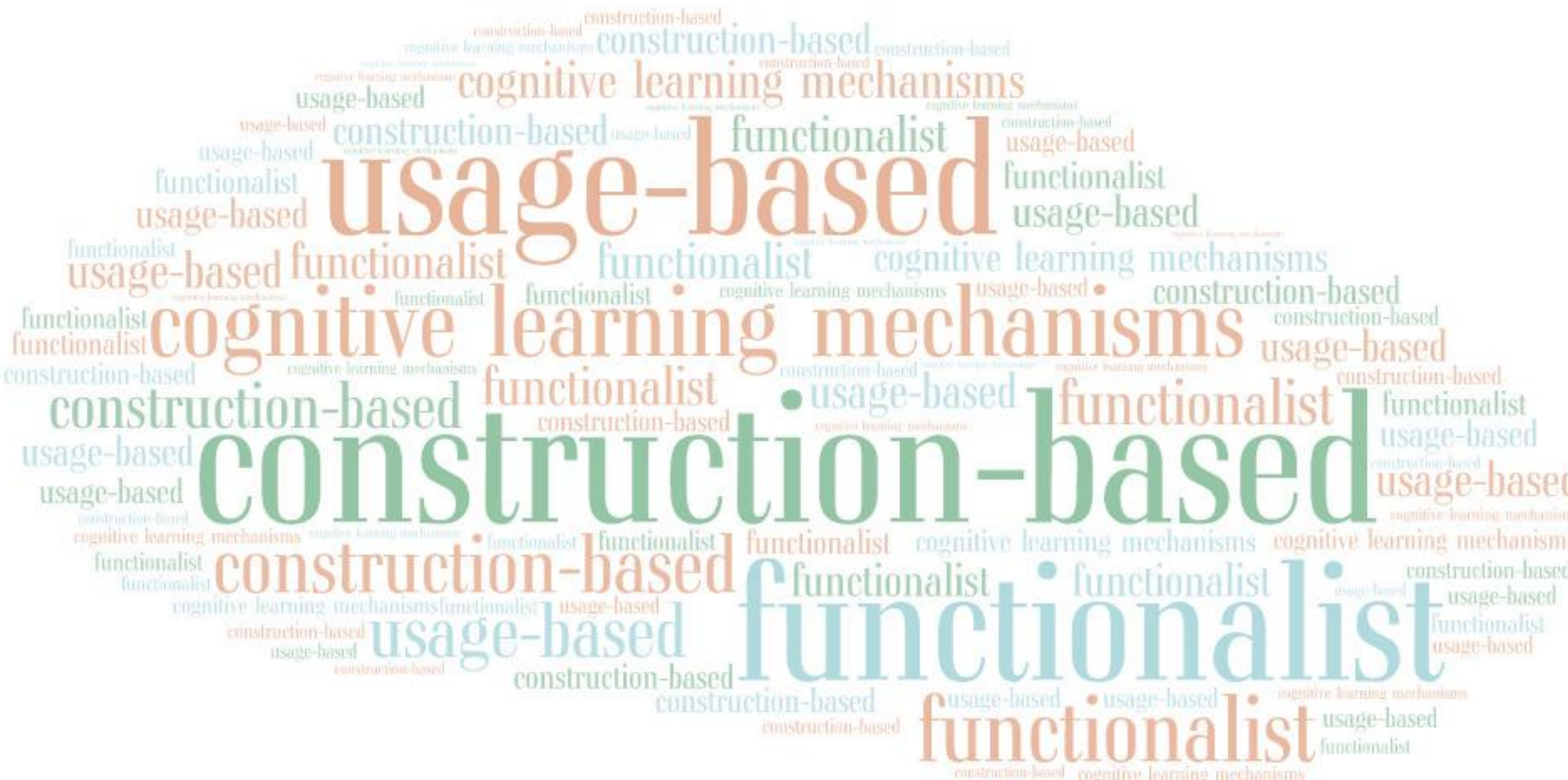
How do children acquire language(s)

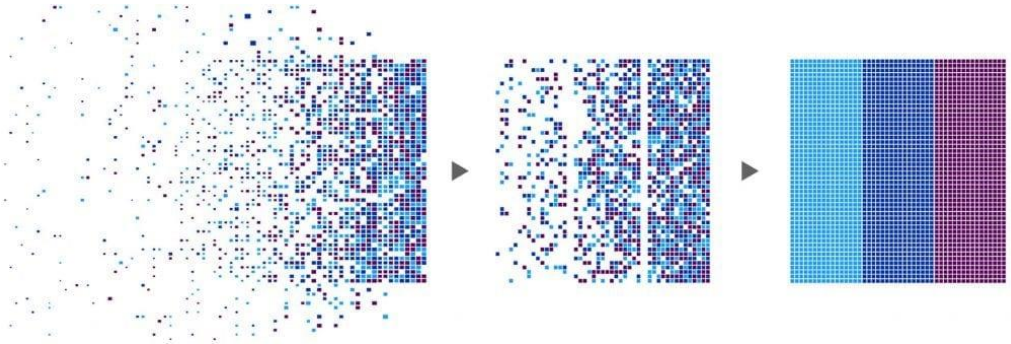
Complexity of languages is remarkable - Different research traditions

- ✓ children are equipped with pre-existing experience (e.g. Chomsky 1965)

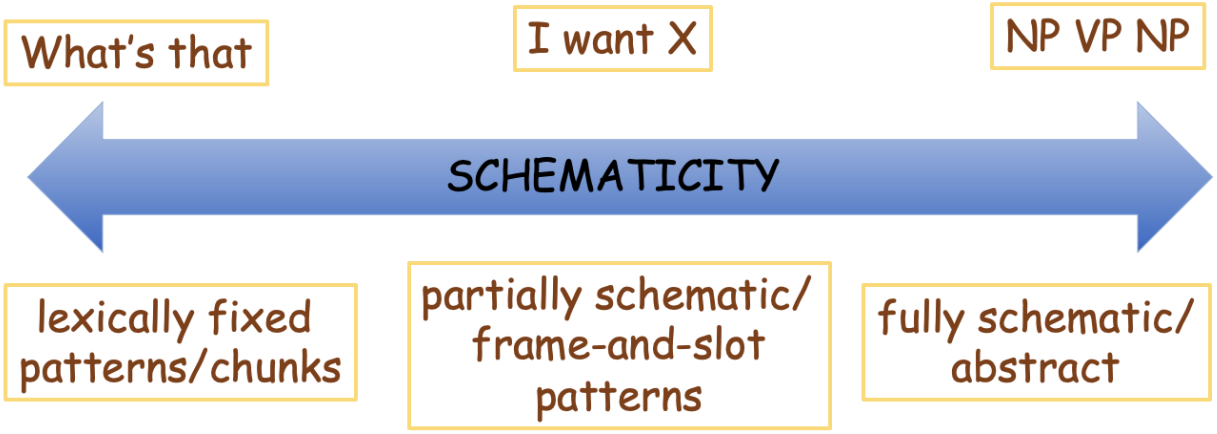
- ✓ children acquire languages by actively constructing complexity (e.g. Tomasello 2003) → piecemeal acquisition

Snapshot of the Usage-Based approach





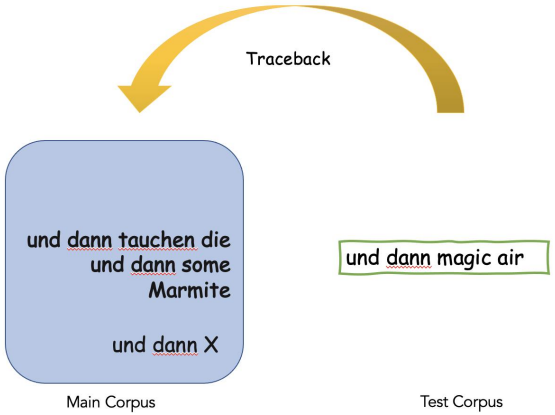
Blurring the line between lexicon and grammar



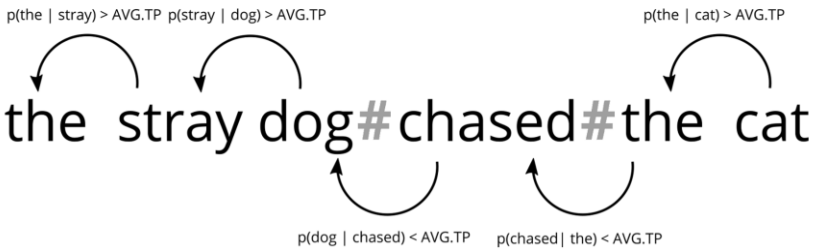
How to account for patterns ?

- Hypothesis: language acquisition is strongly item-based – early child language is highly formulaic

Traceback



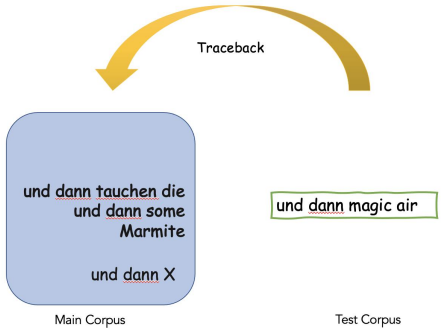
Chunk-based learner



How to account for patterns ?

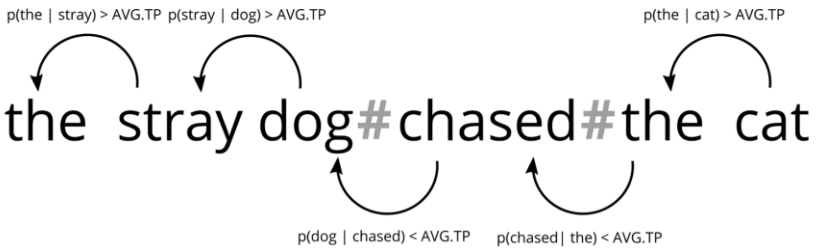
Traceback

focuses on frame-and-slot patterns



Chunk-based learner

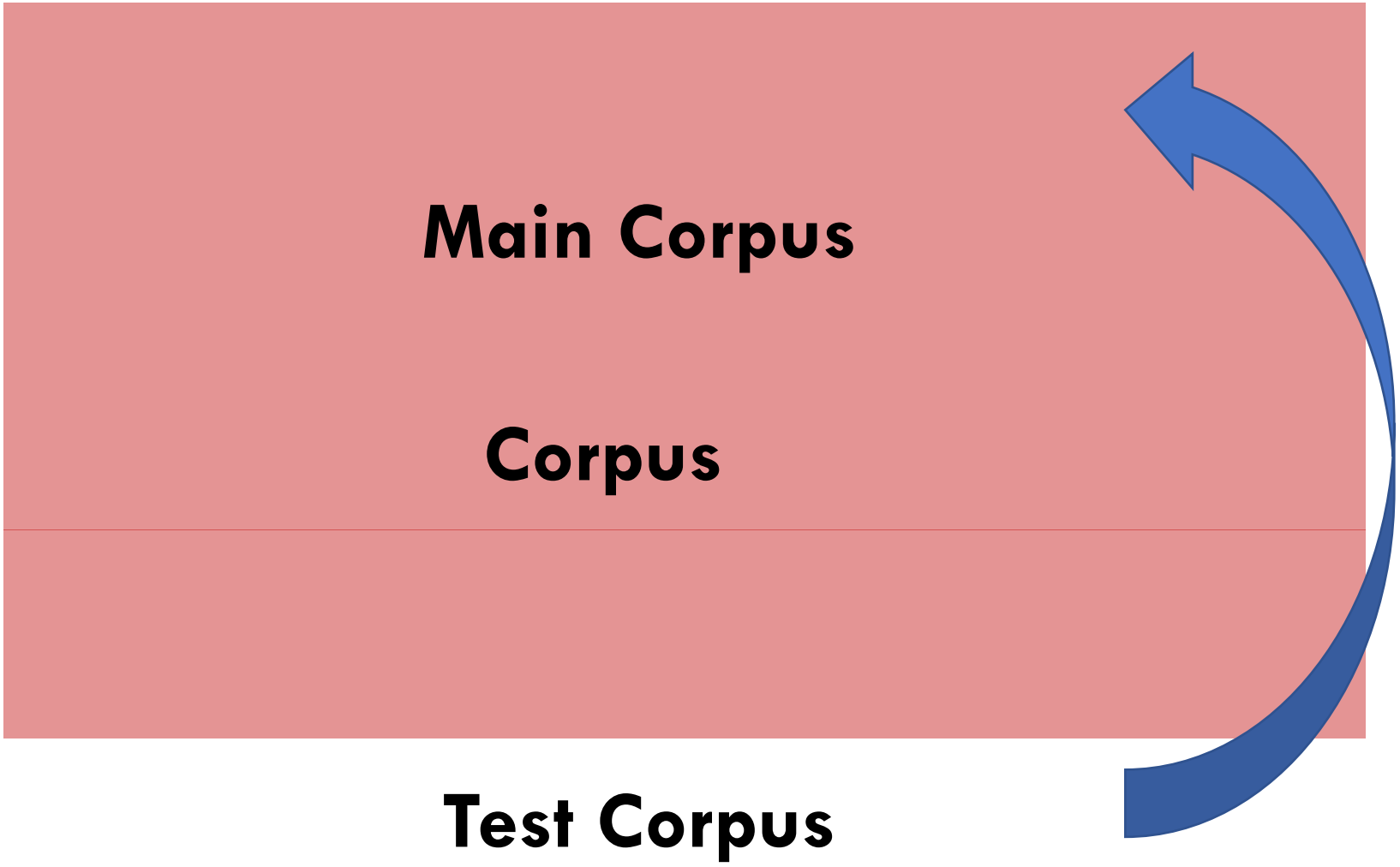
focuses on chunking processes



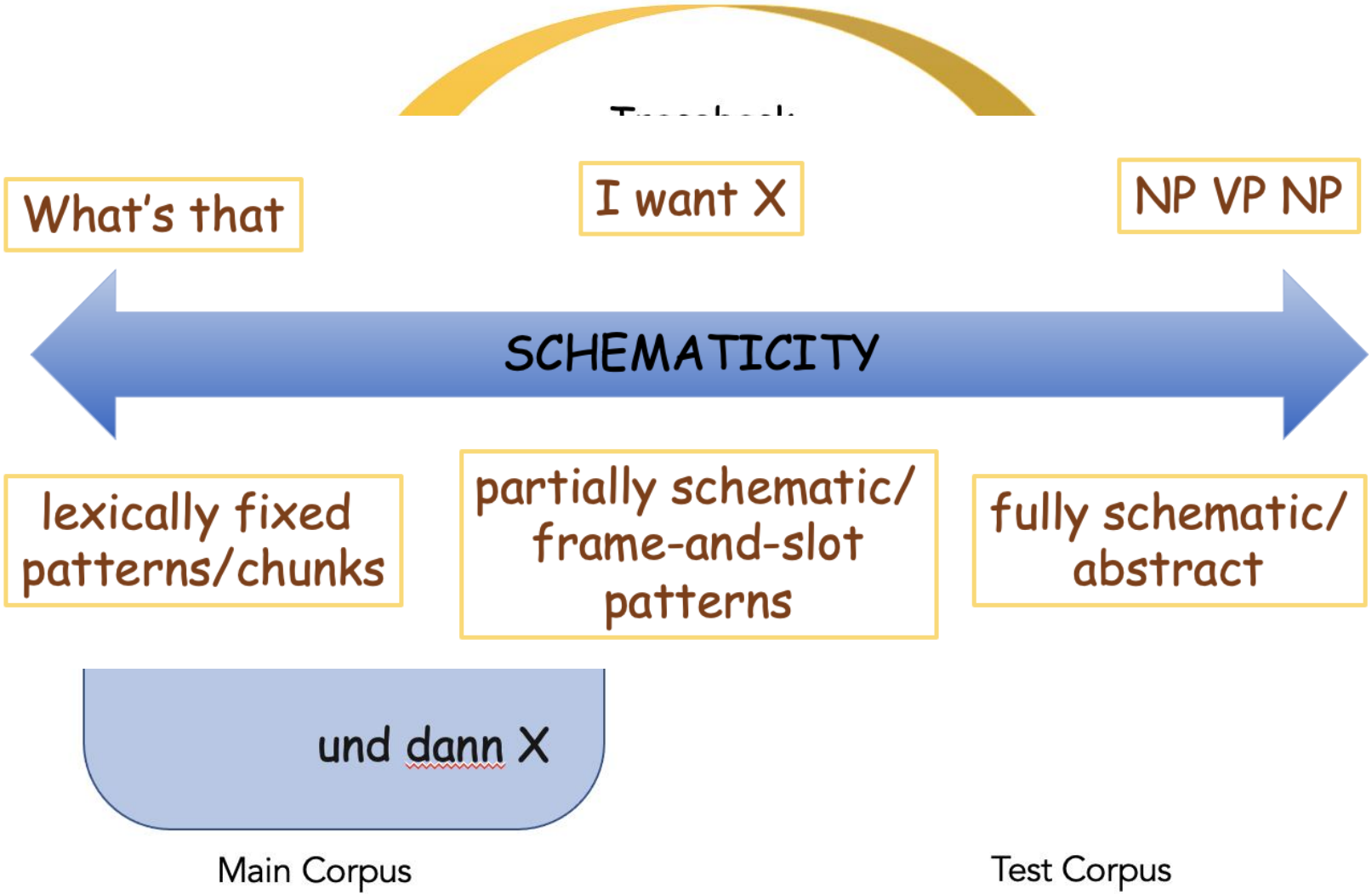
The Traceback method

Corpus

The Traceback method



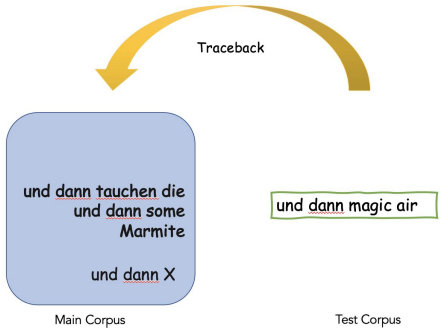
The Traceback method



How to account for patterns ?

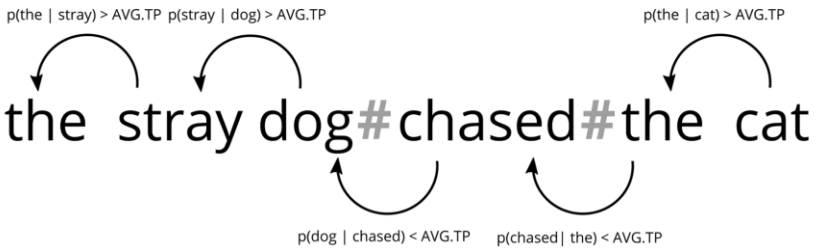
Traceback

focuses on frame-and-slot patterns



Chunk-based learner

focuses on chunking processes



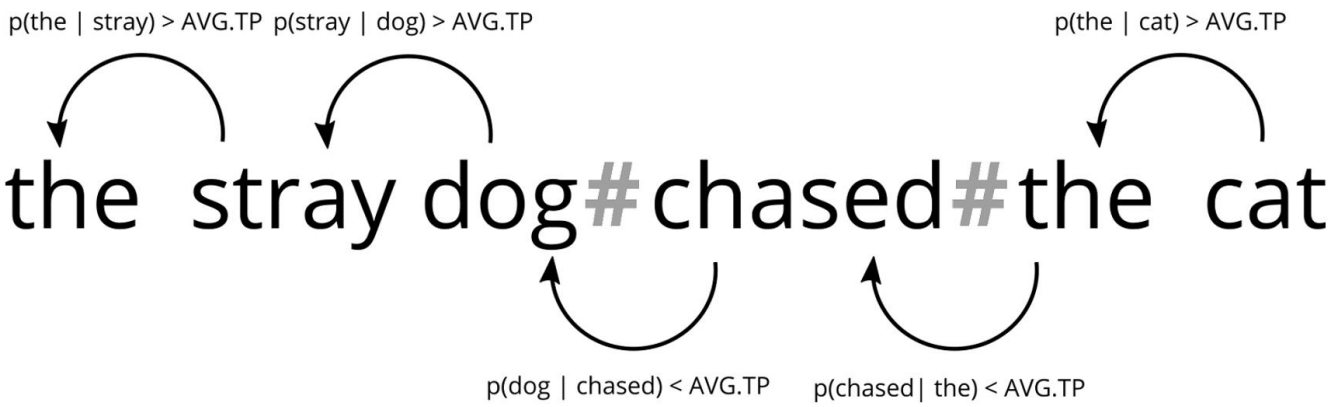
Chunk-Based Learner (CBL)

Corpus

Chunk-Based Learner (CBL)

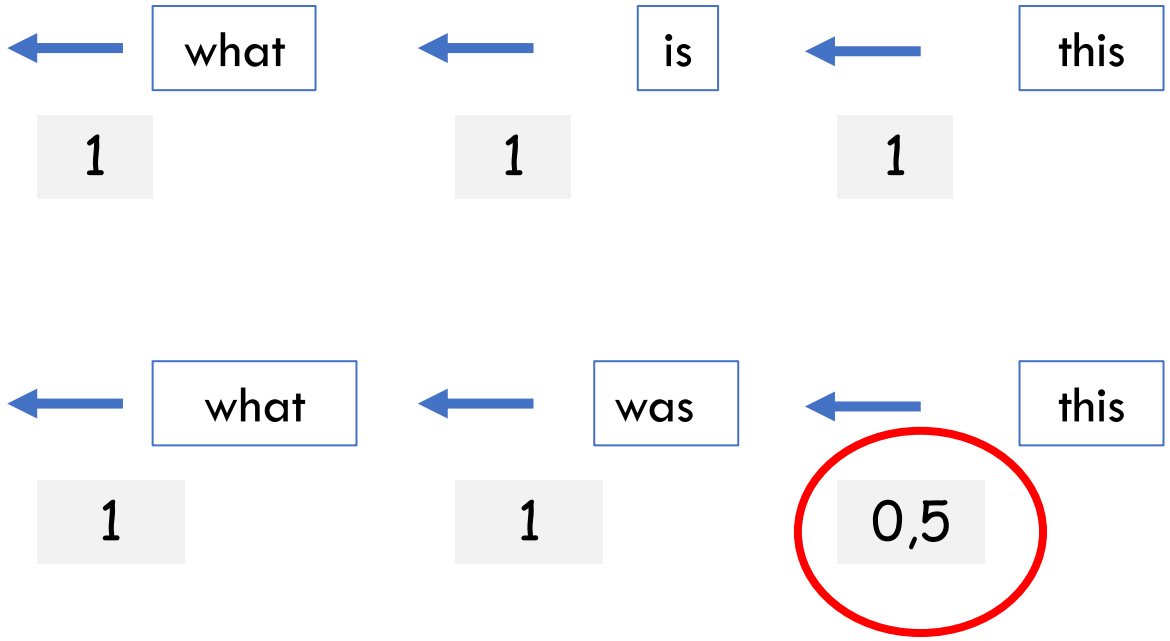
Corpus

detecting chunks incrementally → recognizing multiword chunks by using backward transitional probabilities (BTP)



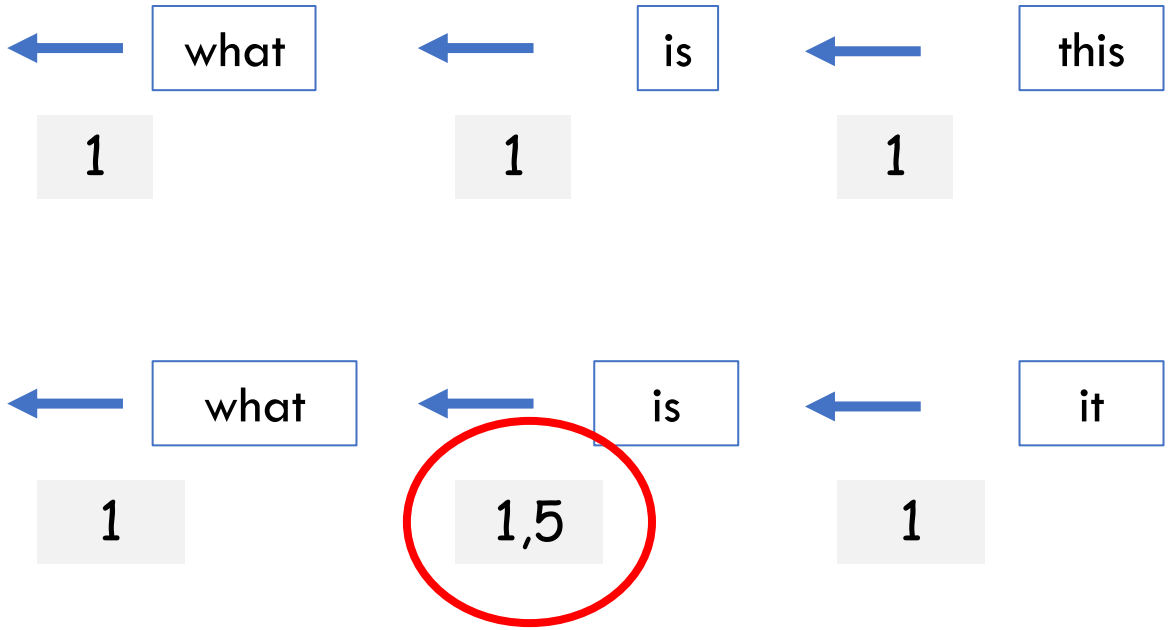
CBL – calculating BTP

What is this?
What was this?

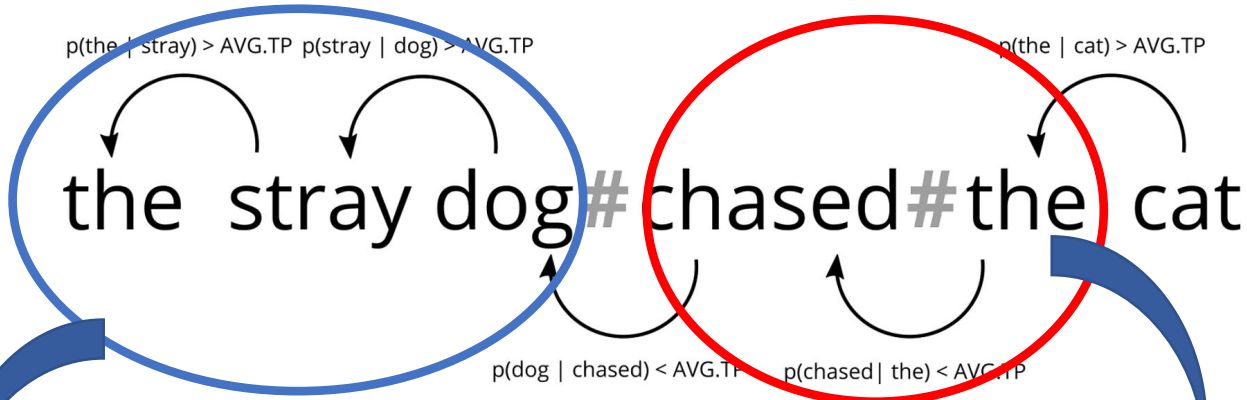


CBL – calculating BTP

What is this?
What is it?



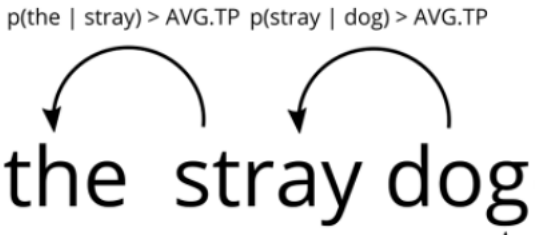
CBL – identifying chunks



BTP value is above the average BTP → chunk

BTP of the word pair falls below the current average BTP → word boundary

CBL – identifying chunks



CHUNKATORY

Data - Bilingual Corpus

DE GB

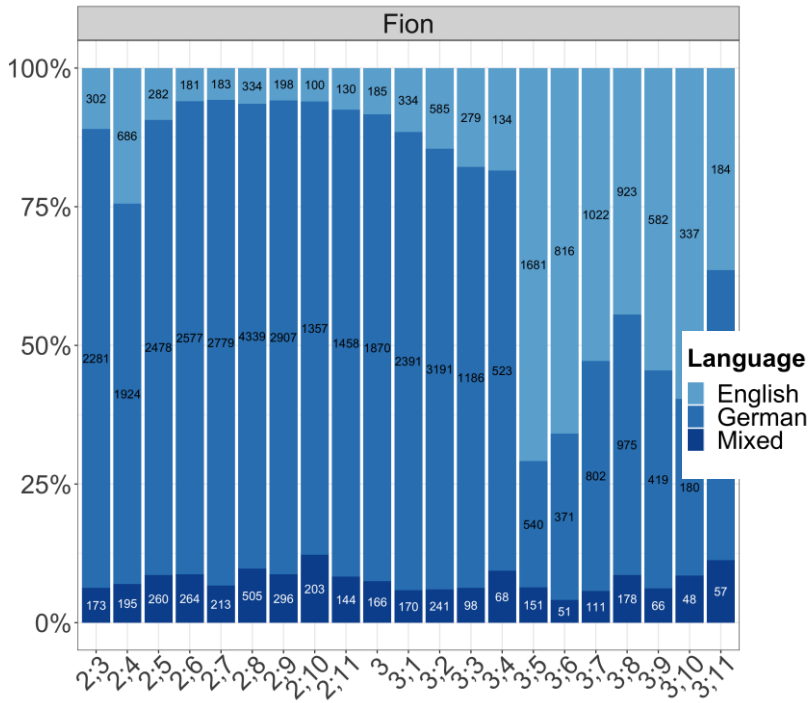
Fion

2;3 - 3;11

N_{CHI} = 47.928

N_{Input} = 180,293

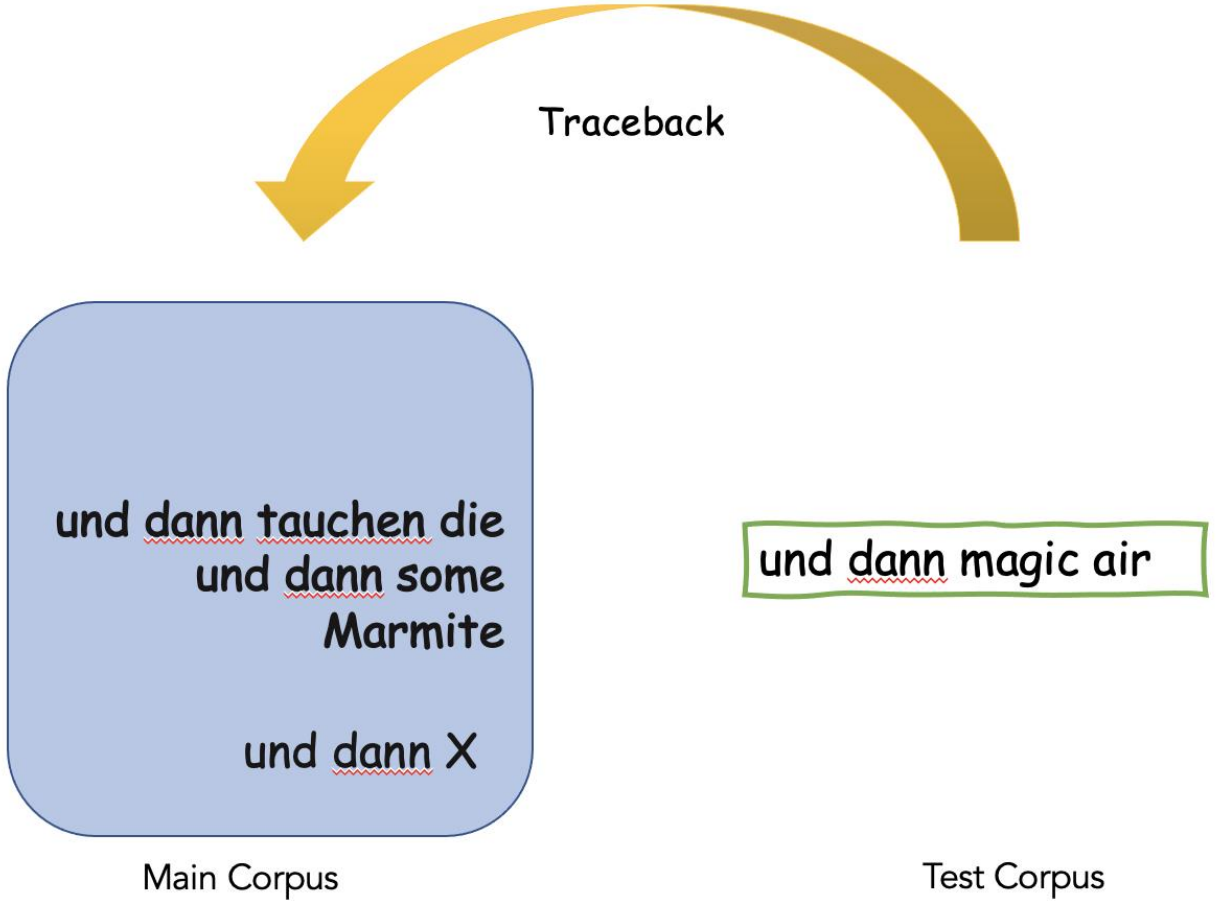
Language Proportions



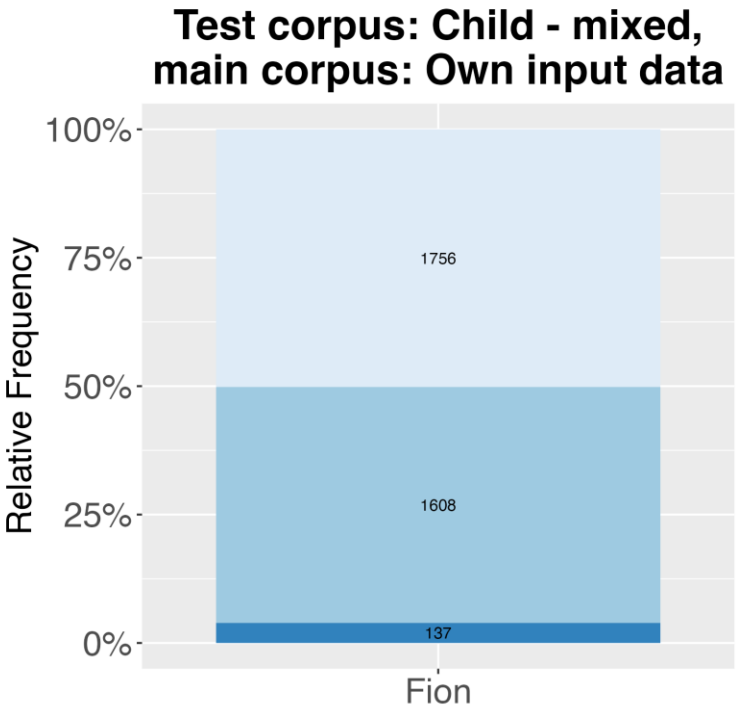
- ✓ entire dataset → monolingual utterances and input as input for the CBL algorithm
- ✓ focus on the "comprehension" side here, i.e., the chunks that the model identifies

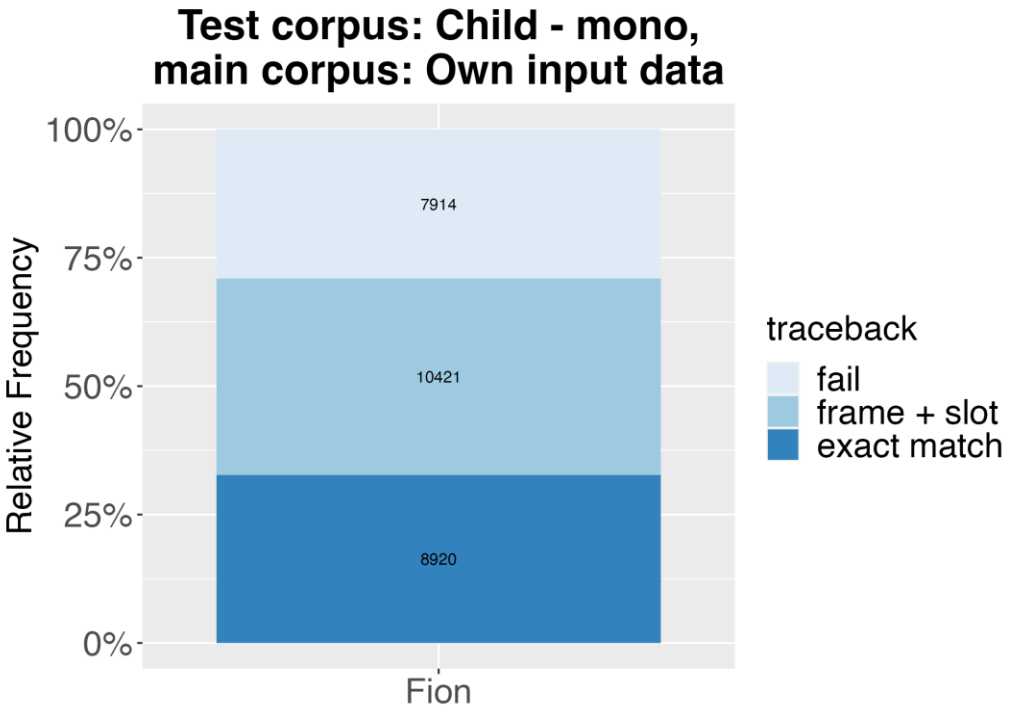
script for the CBL algorithms available @ https://github.com/StewartMcCauley/CBL/

Results - Traceback



Results - Traceback

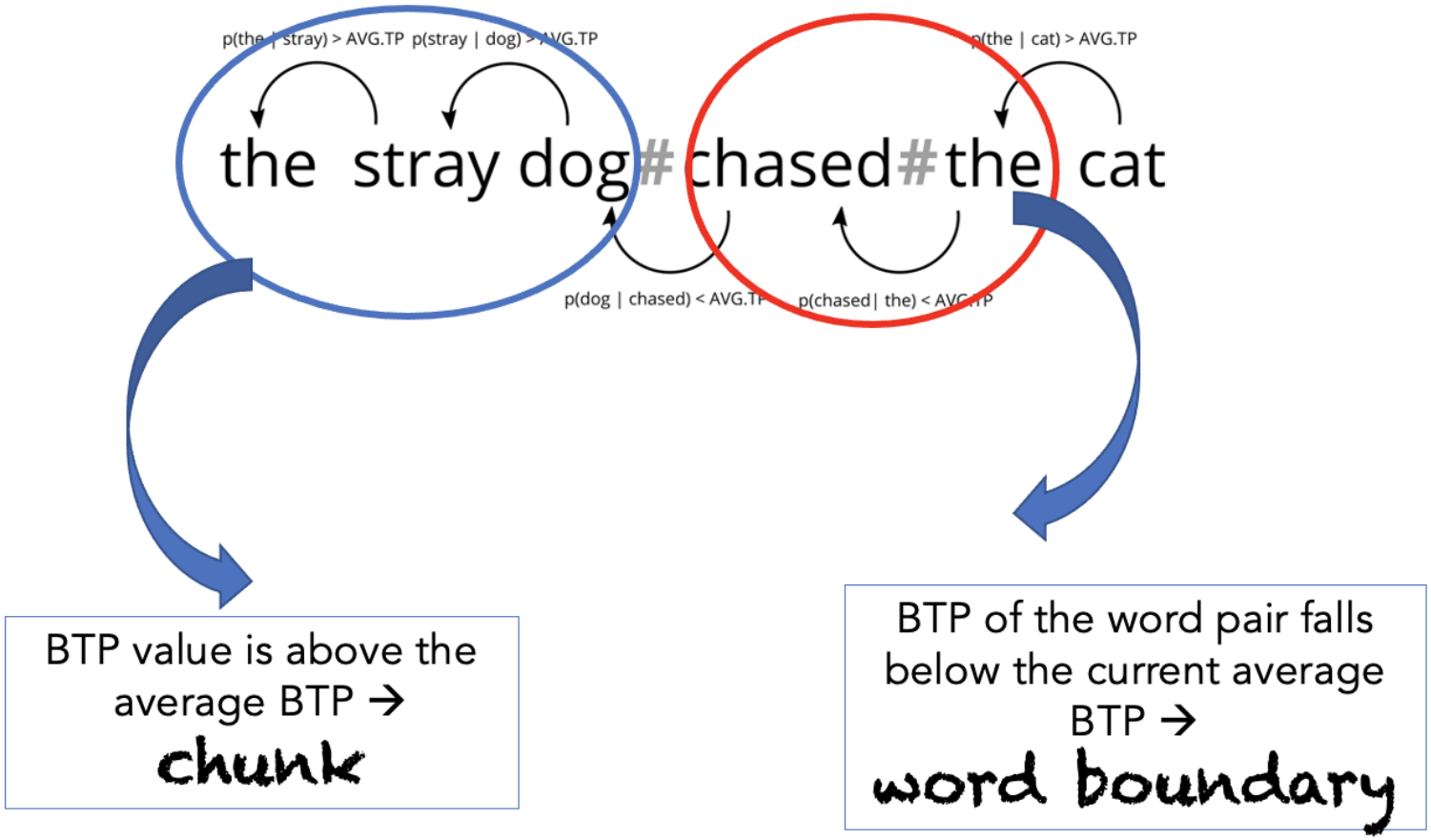




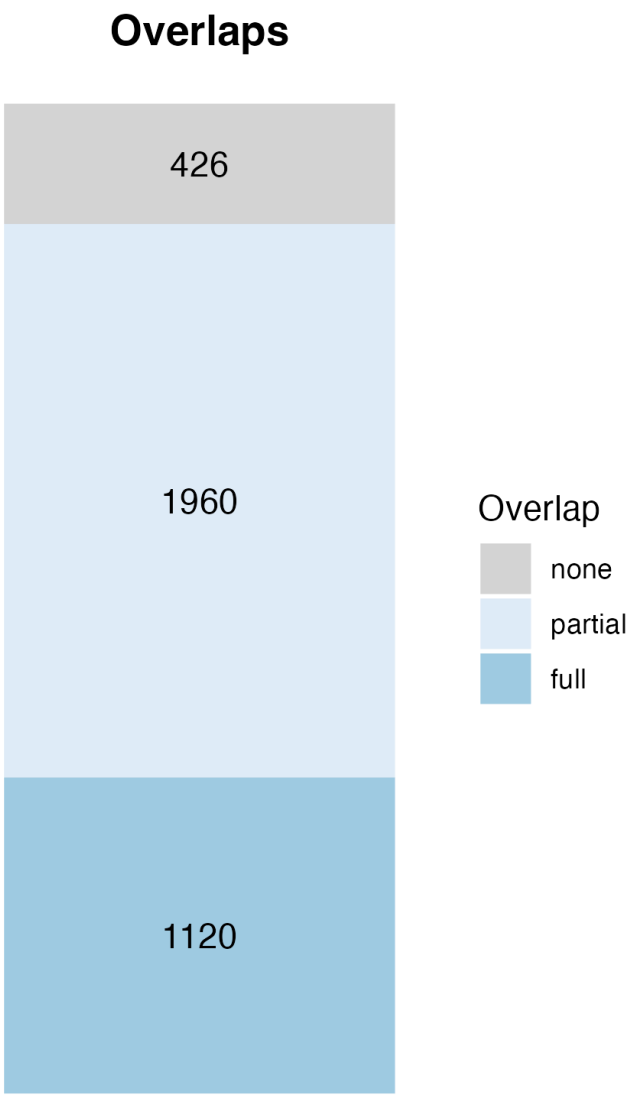
und dann magic air
We have to beeil
und this

- great reliance on lexically fixed patterns in CM
- temporary entrenched bilingual chunks
- frame-and-slot patterns also in the input

Results - CBL

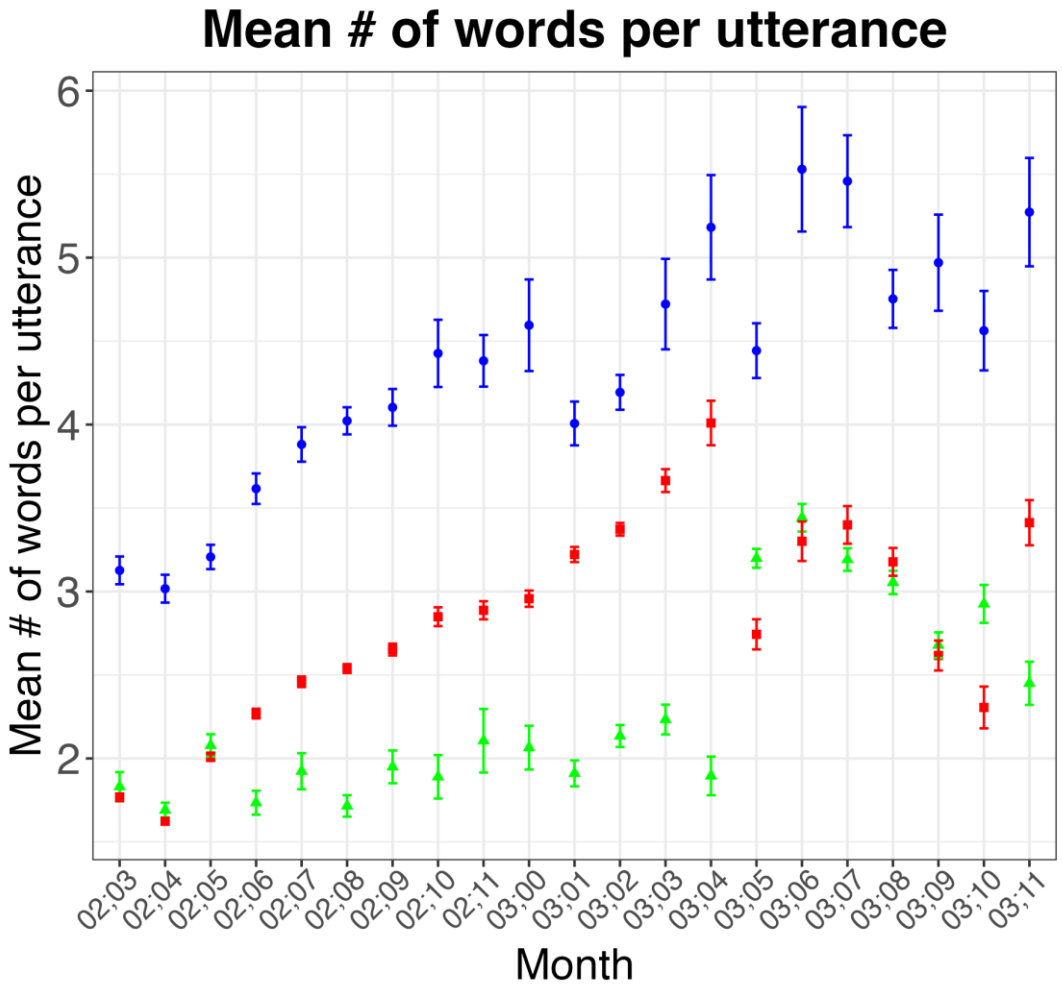

p(the | stray) > AVG.TP
p(stray | dog) > AVG.TP
p(the | cat) > AVG.TP
the stray dog # chased # the cat
p(dog | chased) < AVG.TP
p(chased | the) < AVG.TP
BTP value is above the average BTP →
chunk
BTP of the word pair falls below the current average BTP →
word boundary

Results CMed - CBL



- ✓ lots of CMed utterances contain a chunk or a partial chunk
- ✓ differences between the chunks identified in the child's code-mixed and his monolingual utterances → model does not make a difference between them (G-E-CM)

Results CMed - CBL



- ✓ number of chunks per utterance is much higher in the code-mixed than in the monolingual data
- ✓ mean number of words per chunk tends to be higher in monolingual utterances

Results - CBL

- ✓ position of chunk boundaries identified by the CBL algorithm often coincides with the position of code-mixes
  - ✓ because an English word is of course relatively unlikely to be preceded by a German one, and vice versa

i. *ein kleinen || shark* (a little shark)
ii. *nein || a || nein || a ice hockey player* (no a no a ice hockey player)
iii. *zeig || ice cream* (show ice cream)

Conclusion

- no method has been proven to be exhaustive and that depends on the ways how they implement the notion of patterns

- CBL results complement the TB results in a useful way: both models detect formulaic language use but make different predictions that we see in the data
  - TB → frame-and-slot patterns
  - CBL → chunks

- frequent code-mixed sequences like *nein this*, which were identified as fixed chunks by the TB method, are not identified as chunks by the CBL algorithm

Conclusion

- code-mixed utterances can be accounted for by lexically fixed patterns and emerging frame-and-slot patterns → generating/recycling (creative) utterances from 'bits and pieces' of already acquired constructions

- important to view code-mixing as a dynamic process instead of trying to find a 'one-fit-all' grammar to account for mixes across different populations of children and different languages

Dziękuję!

Thanks!

Danke!

oh das war too much

I am alle

Kiitos!

Спасибо!

das war x

I'm x

Merci!

nuqneH!