The Menzerath-Altmann Law: Challenges We Are Facing

Radek Čech

MUNI ARTS

Masaryk University in Brno

University of Ostrava

UNIVERSITAS OSTRAVIENSIS

Acknowledgment




Ján Mačutek Tereza Motalová Kateřina Pelegrinová Michaela Nogolová

The Menzerath-Altmann Law (MAL)

- it expresses relationships between lengths of linguistic units

The Length of Linguistic Units

- length as a (non)arbitrary property

The Length of Linguistic Units

- length as a (non)arbitrary property

- arbitrariness
  - bed – Bett – letto – postel – postelja

The Length of Linguistic Units

- length as a (non)arbitrary property
- arbitrariness
  - bed – Bett – letto – postel – postelja
- non-arbitrarness

The Length of Linguistic Units

- length as a (non)arbitrary property

- arbitrariness
  - bed – Bett – letto – postel – postelja

- non-arbitrarness
  - frequency

The Length of Linguistic Units

- length as a (non)arbitrary property

- arbitrariness
  - bed – Bett – letto – postel – postelja

- non-arbitrarness
  - frequency
  - inventory of phonemes in a language

The Length of Linguistic Units

- length as a (non)arbitrary property

- arbitrariness
  - bed – Bett – letto – postel – postelja

- non-arbitrarness
  - frequency
  - inventory of phonemes in a language
  - grammatical constraints

The Length of Linguistic Units

- length as a (non)arbitrary property

- arbitrariness
  - bed – Bett – letto – postel – postelja

- non-arbitrarness
  - frequency
  - inventory of phonemes in a language
  - grammatical constraints
  - units of different levels → **The Menzerath-Altmann Law**

The Menzerath-Altmann Law

- relation between **length** of language units positioned in a **vertical hierarchy** according to their size

Hierarchy of Linguistic Units

grammatical/structural units

SENTENCE

CLAUSE

WORD

MORPHEME

PHONEME

Hierarchy of Linguistic Units

grammatical/structural units

SENTENCE

CLAUSE

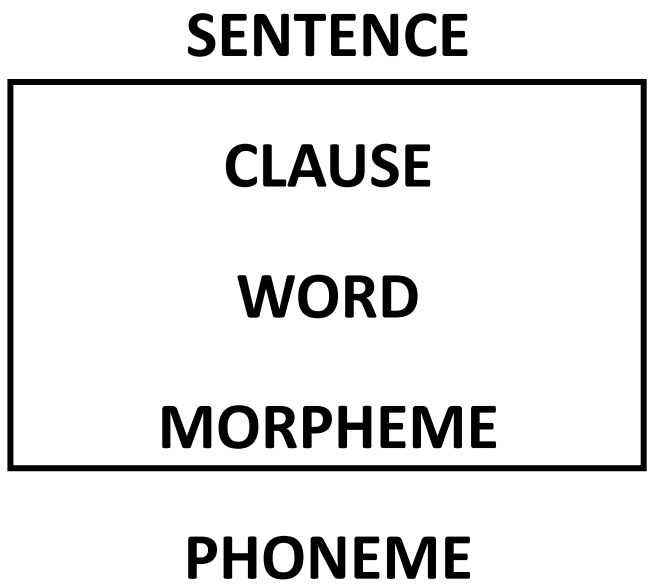
WORD

MORPHEME

PHONEME

Hierarchy of Linguistic Units

grammatical/structural units



Hierarchy of Linguistic Units

grammatical/structural units

SENTENCE

CLAUSE

WORD

MORPHEME

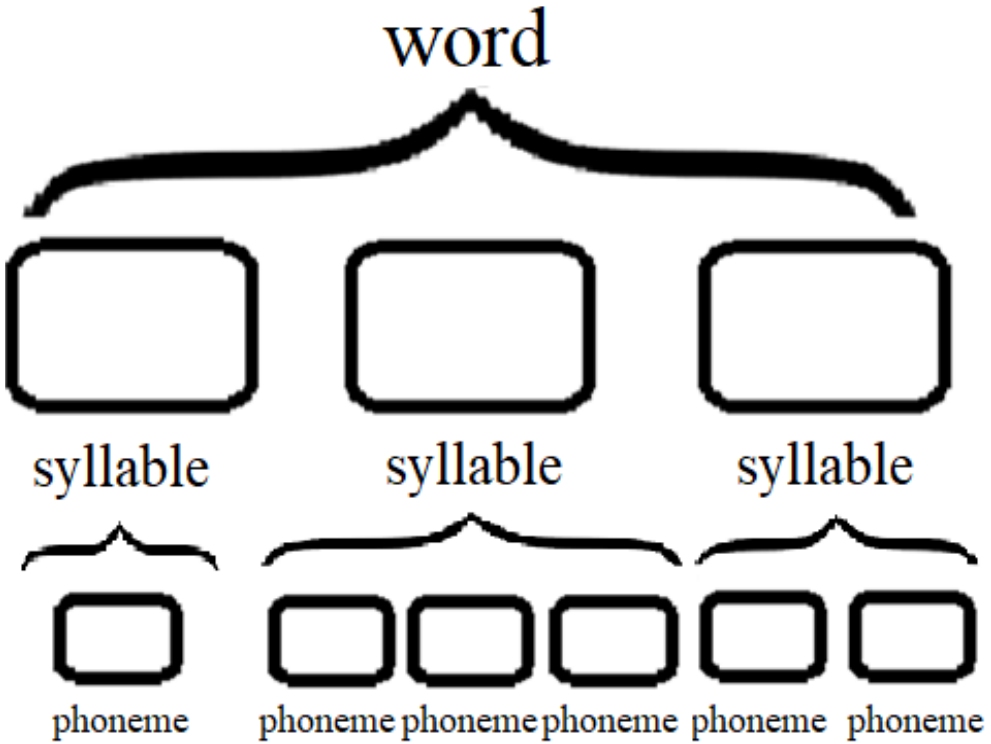
PHONEME

Grammatical/structural vs. sound units

grammatical/structural units	sound units
sentence	**utterence**
clause	**clause-unit**
word	**phonological word**
morpheme	**syllable**
phoneme	**sound**

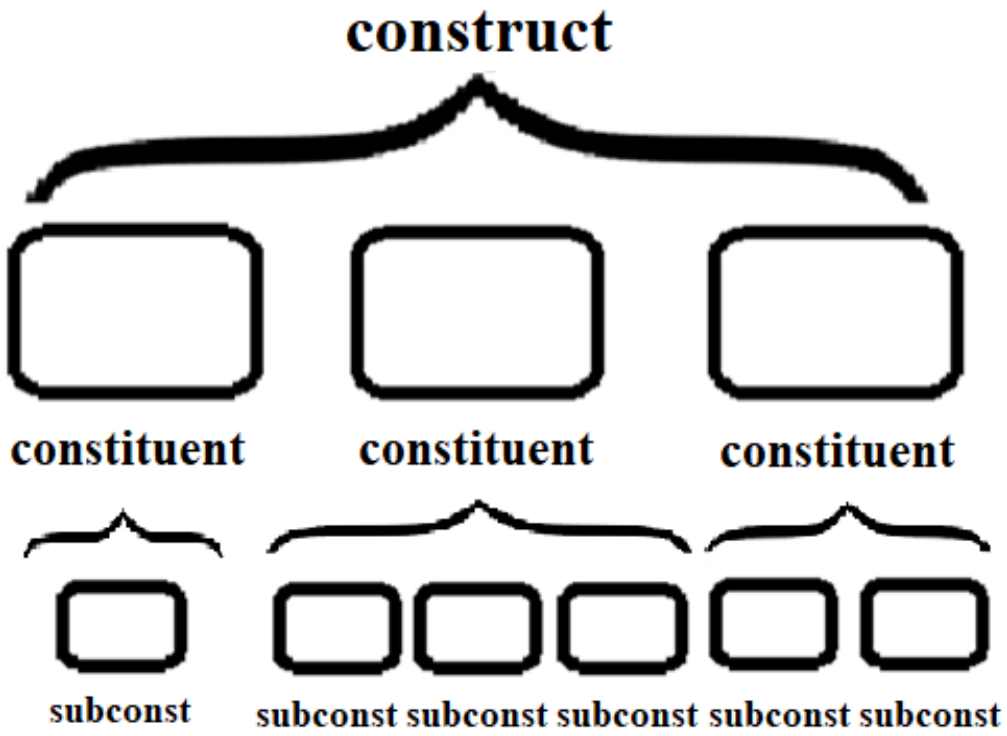
Menzerath-Altmann Law

- relation between **length** of language units positioned in a **vertical hierarchy** according to their size



Menzerath-Altmann Law

- relation between **length** of language units positioned in a **vertical hierarchy** according to their size



The Menzerath-Altmann Law (MAL)

- "[t]he longer a language construct the shorter its components (constituents)"

$$y = ax^b$$

where x is a construct length, y is a constituent length, and a, b, c are parameters.

The Menzerath-Altmann Law (MAL)

- "[t]he longer a language construct the shorter its components (constituents)"

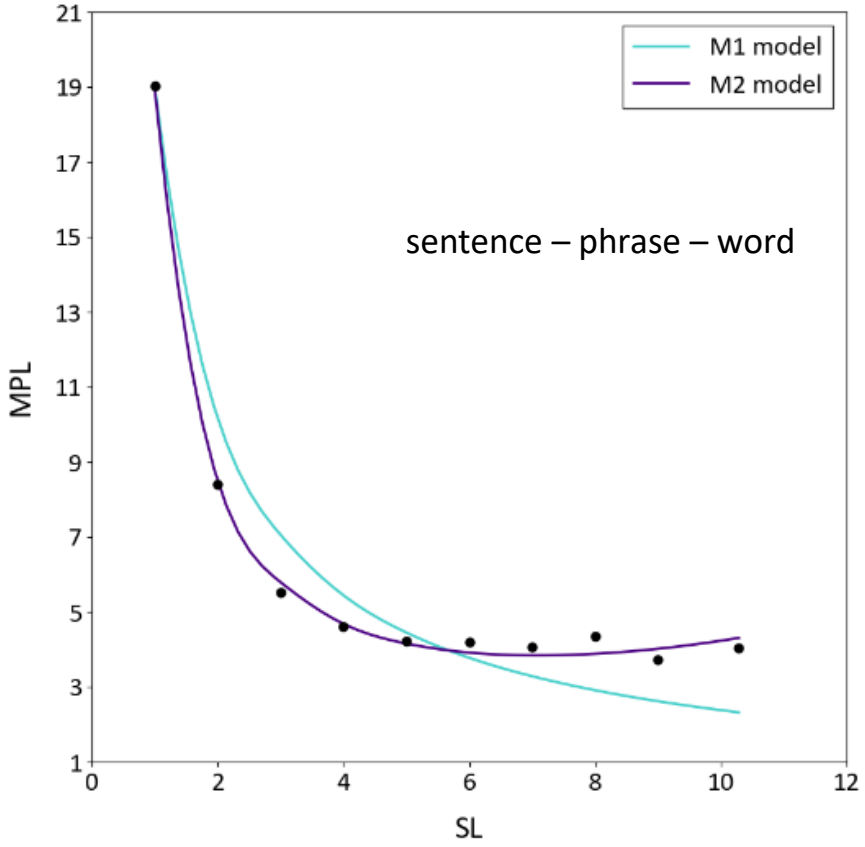
$$y = ax^b$$

- "[t]he length of the components is a function of the length of language constructs"

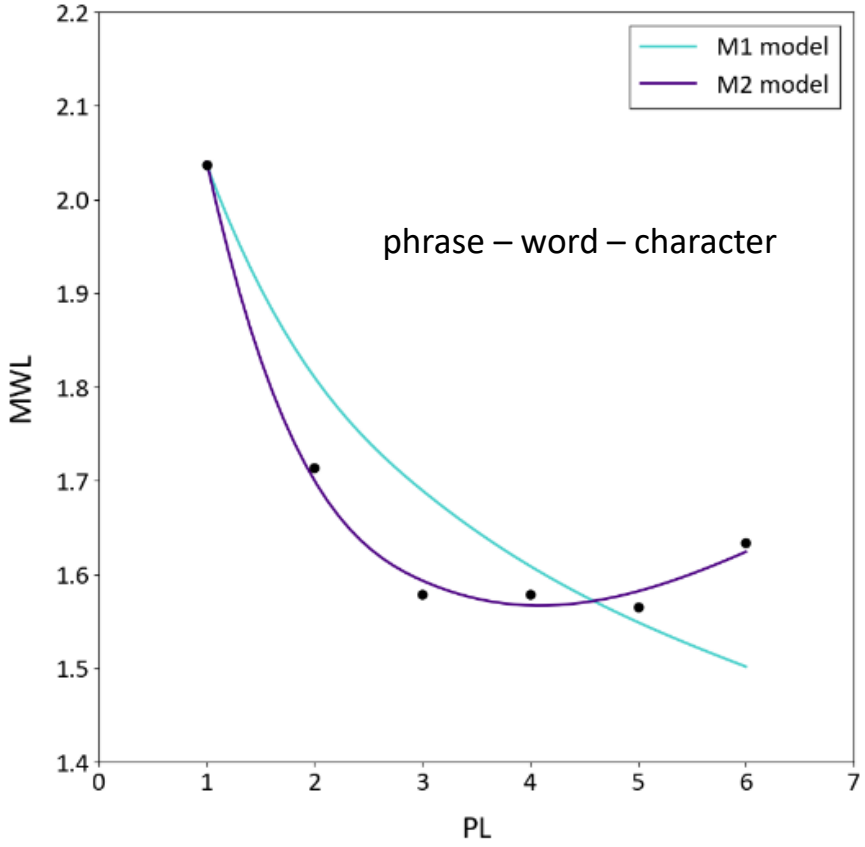
$$y = ax^b e^{-cx}$$

where x is a construct length, y is a constituent length, and a, b, c are parameters.

The MAL: Formulas



short: $R^2 = 0.91$
long: $R^2 = 0.99$



short: $R^2 = 0.76$
long: $R^2 = 0.99$

Interpretation of MAL - parameters

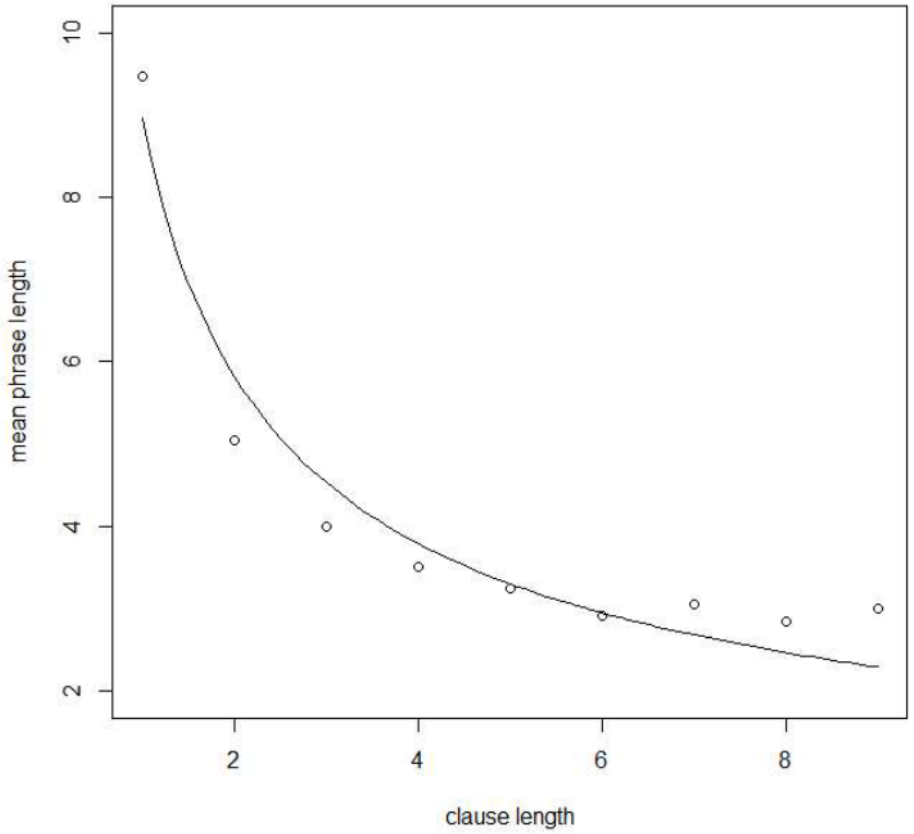
$$y = ax^b$$

$$y = ax^b e^{-cx}$$

Interpretation of MAL – parameter a

$$y = ax^b$$

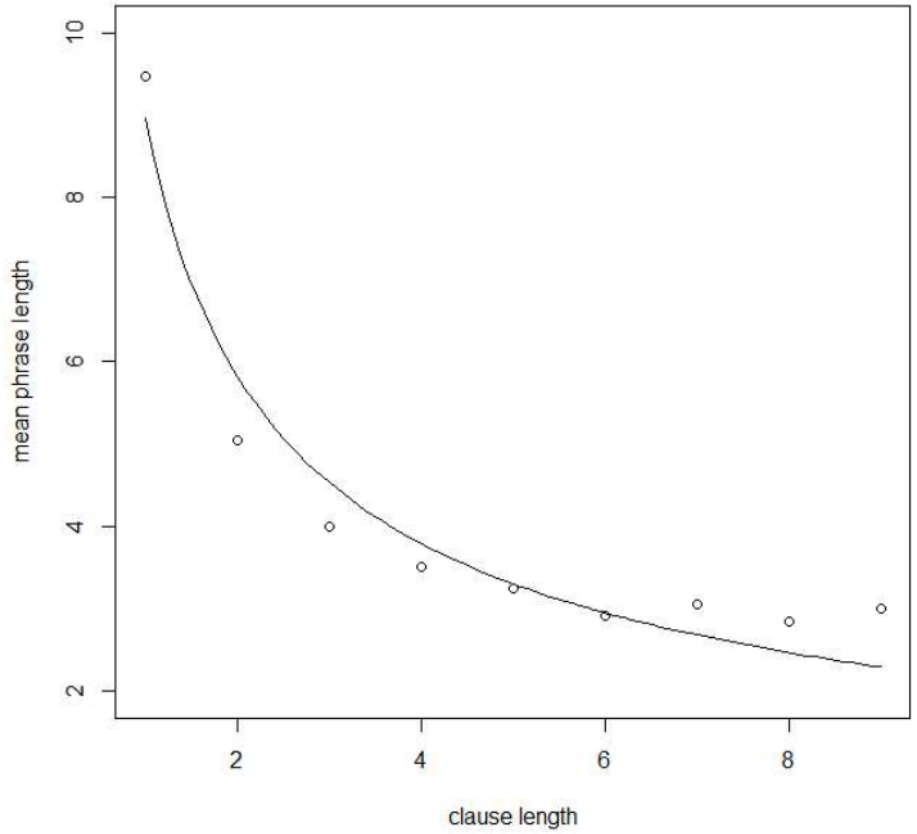
- approximately equivalent to the mean size of constituents belonging to a one-constituent construct

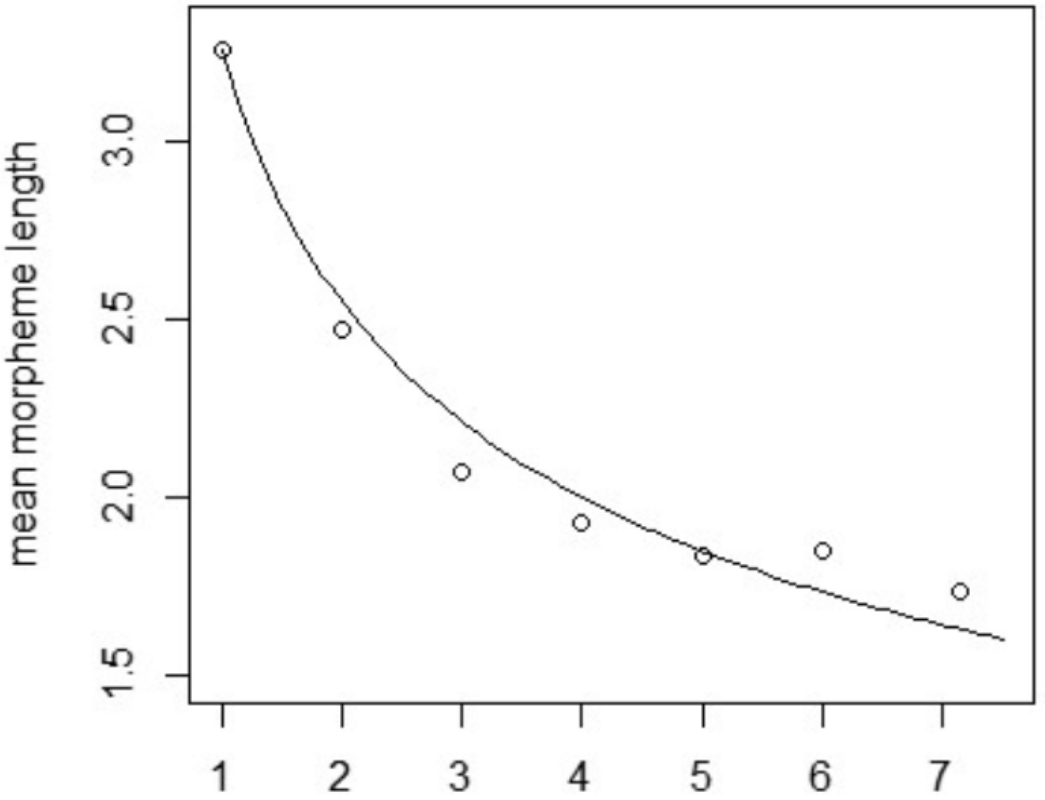


Interpretation of MAL – parameter a

$$y = ax^b$$

- approximately equivalent to the mean size of constituents belonging to a one-constituent construct





Interpretation of MAL – parameter *a*

- $x = 1$

$$y = ax^b = a1^b = a$$

(Köhler 1982)

Interpretation of MAL – parameter *a*

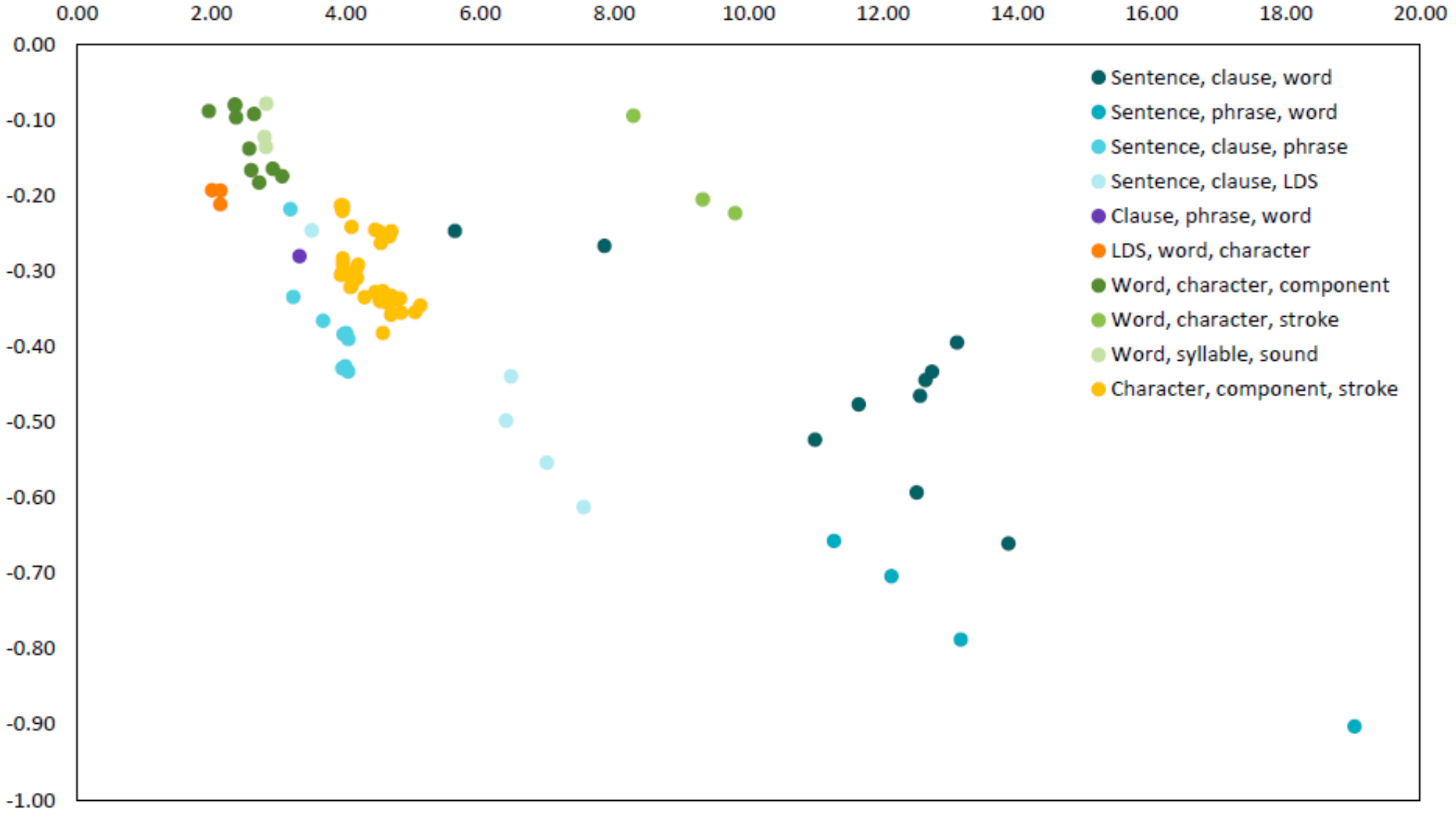
- $x = 1$

$$y = ax^b = a1^b = a$$

(Köhler 1982)

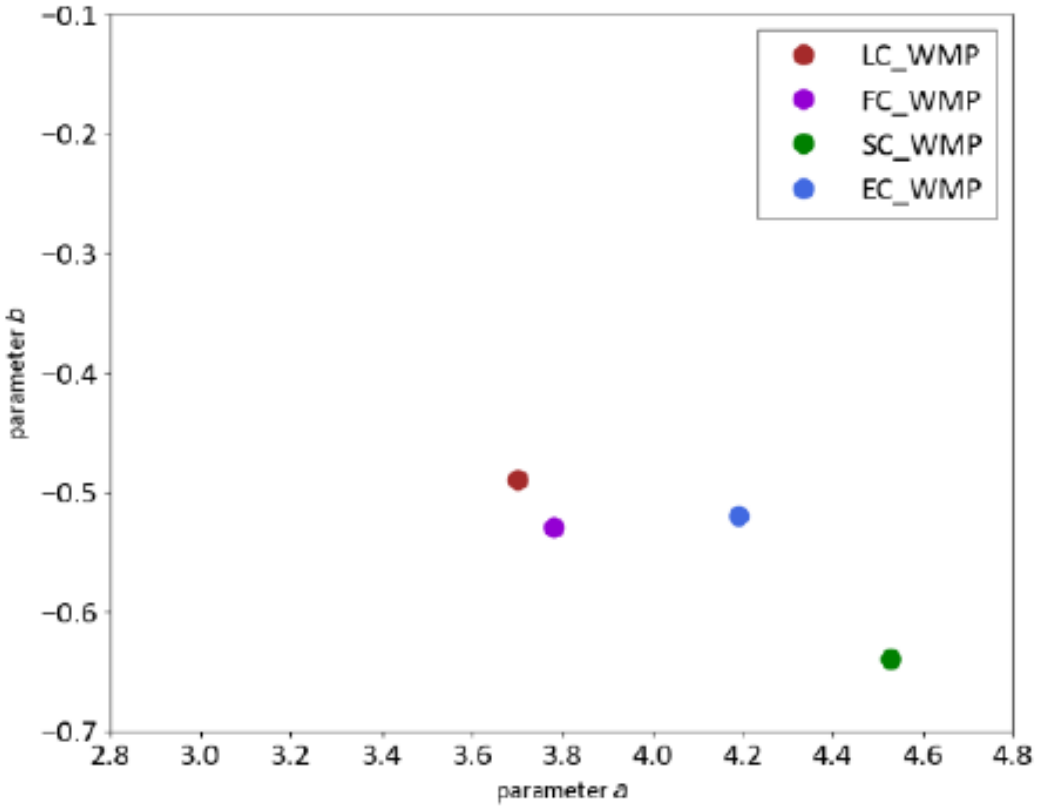
- *a* is (probably) influenced by the **language, level, genre, author**…

Interpretation of MAL – parameter *a*

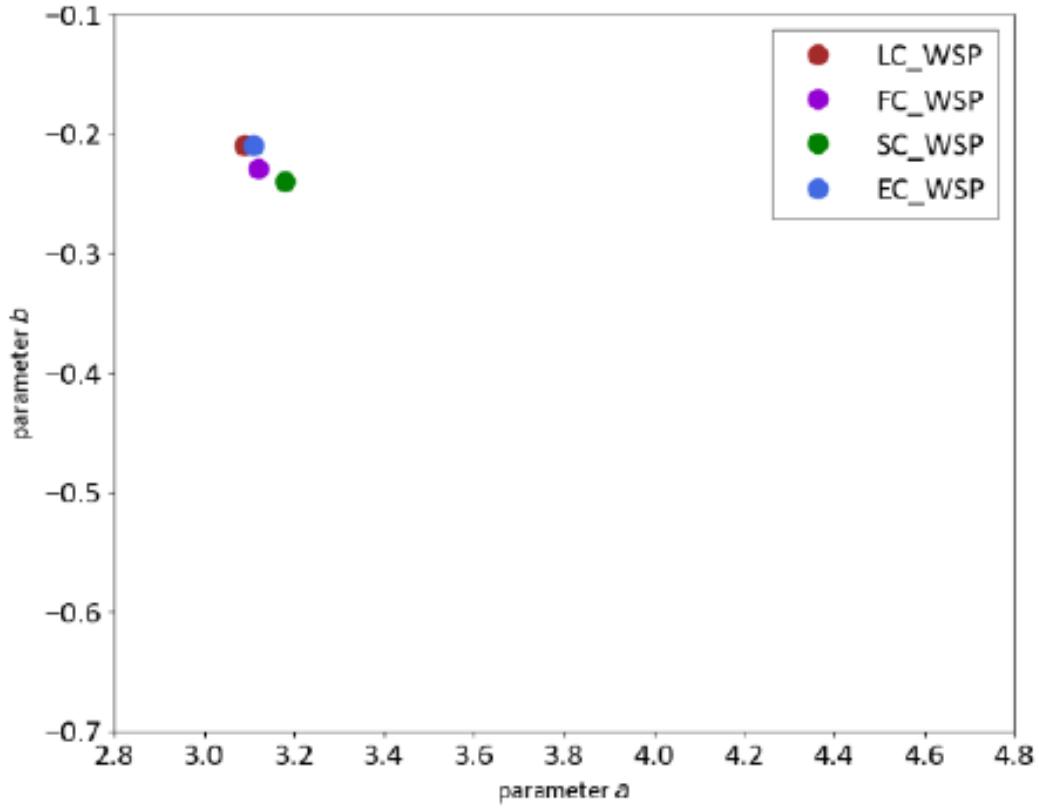


Interpretation of MAL – parameter *a*

word-morpheme-phoneme



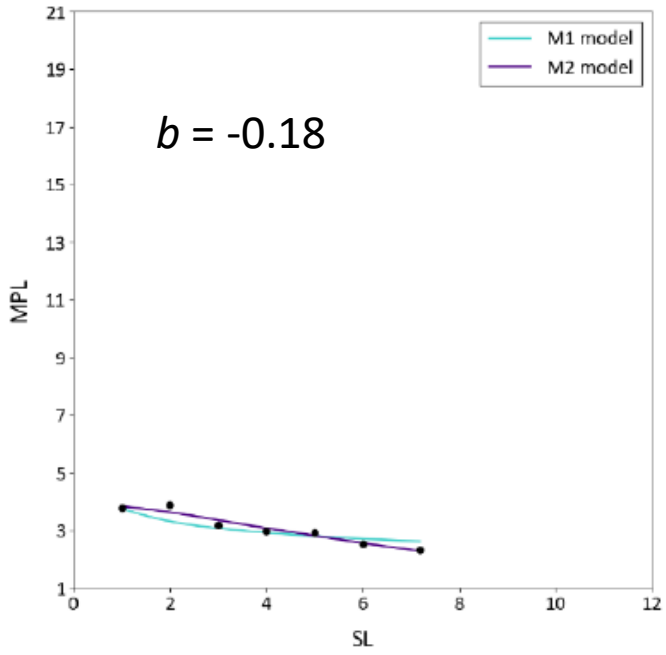
word-syllable-phoneme



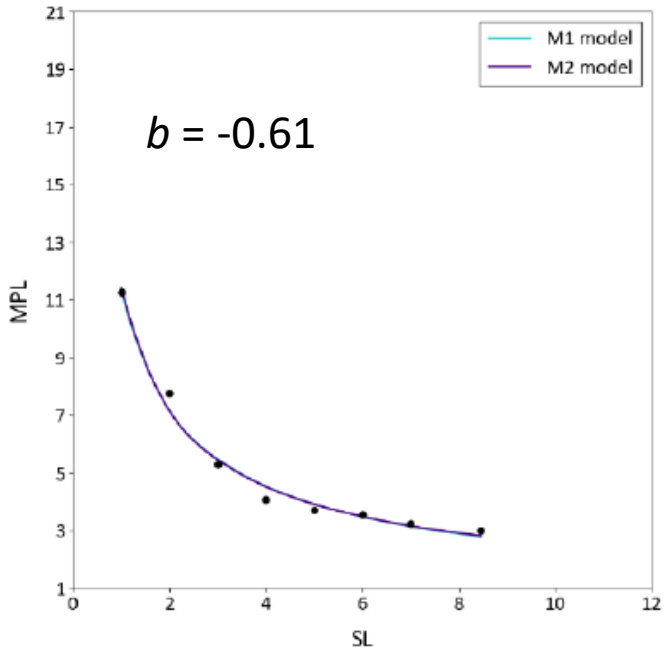
Interpretation of MAL – parameter *b*

$$y = ax^b$$

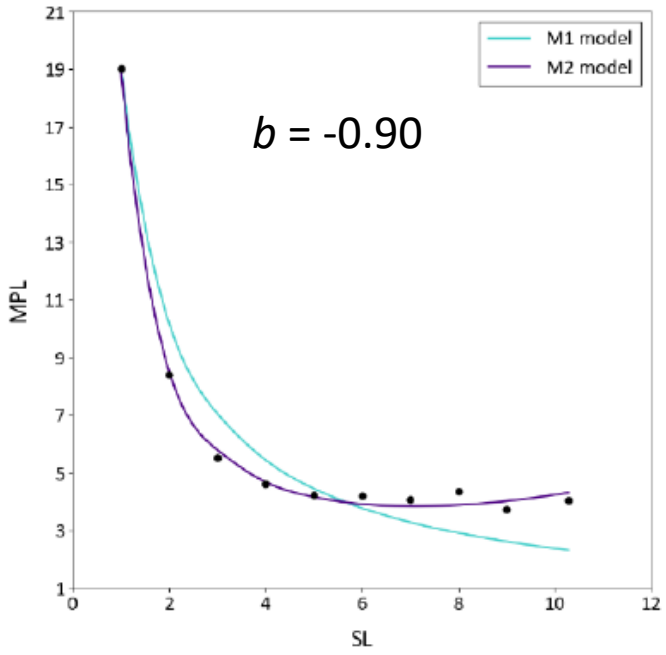
- shows a shortening tendency
- sentence, phrase and word



HK-P



PUD



GSD

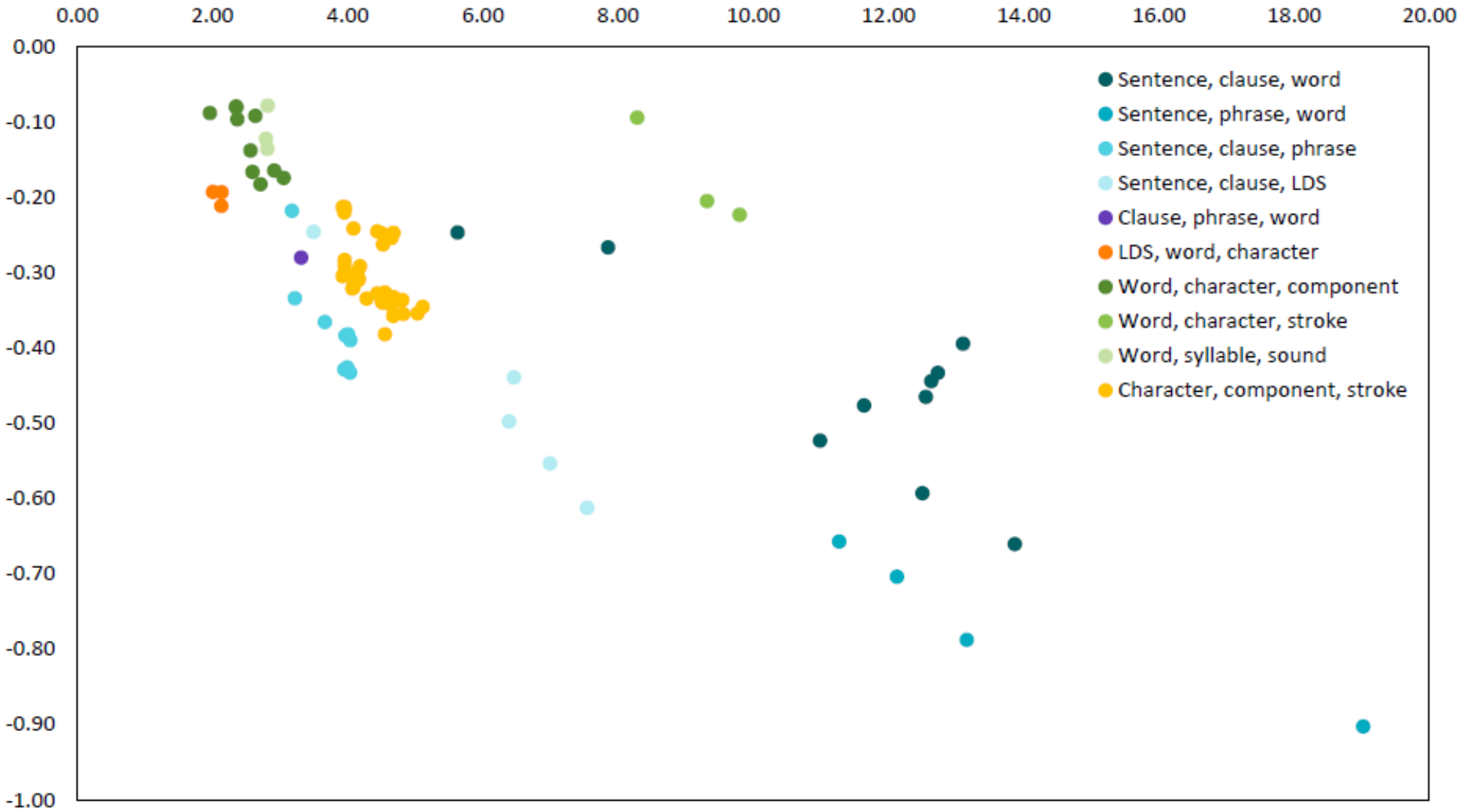
Interpretation of MAL – parameter *b*

$$y = ax^b$$

- shows a shortening tendency

- depends on
  - a linguistic level (Köhler 2012, Cramer 205)

Interpretation of MAL – parameter *b*



Interpretation of MAL – parameter *c*

$$y = ax^b e^{-cx}$$

- the least analysed parameter

- more relevant to lower linguistic levels (e.g. phonetic or word level) while being irrelevant to higher ones (e.g. syntactic level)
  - (Vulanović and Köhler, 2005; Andres et al. 2012; Köhler, 2012)

- however: lack of linguistic interpretation

Interpretation of MAL – general

- result of economy rules
  - ensure manageability of the whole
  - (Menzerath. 1954, p. 101)

- the principle of least effort
  - it balances lengthening and shortening tendencies
  - (Altmann, 1980, p. 5)

Interpretation of MAL – general

- information theory
  - the longer the information, the higher the amount of noise in the channel
  - the higher the degree of activation of the central nervous system (CNS)
  - In order to compensate for this burden and ensure the reliability of the transmitted information, the processing system shortens the information by splitting it into smaller segments
  - (Schwibbe, 1984)

Interpretation of MAL – general

- short-term memory and amount of structural information

(Köhler, 1984)

- cognitive capacity

(Jin and Liu, 2017; Jiang and Ma, 2020; Jiang and Jiang, 2022)

- breathing and lung capacity

(Torre, Dębowski and Hernández-Fernández, 2021, p. 2)

Interpretation of MAL – general

- cognitive capacity
  - achieving the goal – limited sources
  - working memory

The Menzetarth-Altmann law

- one of the most analyzed law in quantitative linguistics
- corroborated for many languages, levels

The Menzetarth-Altmann law

- one of the most analyzed law in quantitative linguistics
- corroborated for many languages, levels

- however…

MAL – challenges/problems

- what should be analyzed?
  - tokens / types / lemmas

MAL – challenges/problems

- what should be analyzed?
  - tokens / types / lemmas
- how the law should be analyzed?
  - look to whole structure

MAL – challenges/problems

- what should be analyzed?
  - tokens / types / lemmas
- how the law should be analyzed?
  - look to whole structure
- do we know/analyze proper levels?
  - clause - ??? - word

MAL – challenges/problems

- what should be analyzed?
  - tokens / types / lemmas
- how the law should be analyzed?
  - look to whole structure
- do we know/analyze proper levels?
  - clause - ??? - word

MAL: tokens – types – lemmas

- **tokens** = analysis of **language use** (language behaviour)
  - impact of frequency

- **types** / **lemmas** = analysis of **lagnguage structural property**

MAL: tokens – types – lemmas

- **tokens** = analysis of **language use** (language behaviour)
  - impact of frequency

MAL: tokens – types – lemmas

- Menzerath (1954)
  - German dictionary
- Altmann and Schwibbe (1989)
  - argued in favour of counting a unit only once
- Stave (2020, p. 4)
  - “Menzerath’s Law is expected to be due to an intrinsic trade-off between the components and the carrier, and not to the frequency of usage of the specific carrier”

MAL: tokens – types – lemmas

- **word token** analysis failing the MAL
  - Alekseev, 1998; Motalová and Matoušková, 2014; Benešová, Faltýnek and Zámečník, 2015; Chen and Liu, 2016, 2019, 2022

MAL: tokens – types – lemmas

- **word token** analysis failing the MAL
  - Alekseev, 1998; Motalová and Matoušková, 2014; Benešová, Faltýnek and Zámečník, 2015; Chen and Liu, 2016, 2019, 2022

- **word types** abides by the MAL & **word tokens** do not abide by the MAL
  - Mikros and Milička (2014), Milička (2014), and Rovenchak (2015)
  - mean syllable length is counted

MAL: tokens – types – lemmas

- MAL & word length in Chinese
  - ambiguous results

- Motalová, T., Mačutek, J., Čech, R. (2024). **Word length in Chinese: The Menzerath-Altmann law is valid after all**. *Journal of Quantitative Linguistics*.
  - https://www.tandfonline.com/doi/full/10.1080/09296174.2023.2259937

Word length in Chinese & MAL

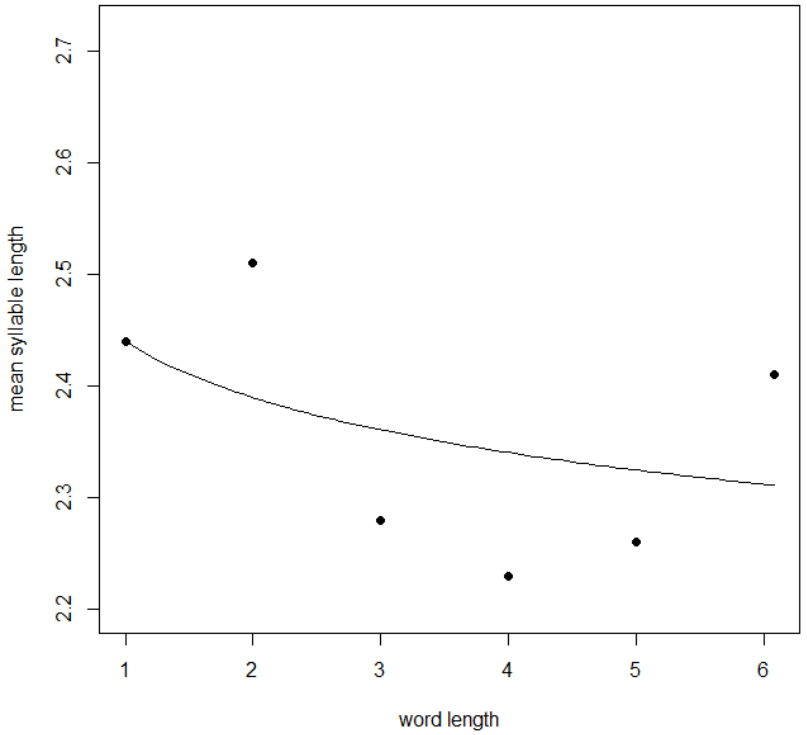
- two translations of the New Testament

1. in pinyin transcription
   - word – syllable – phoneme

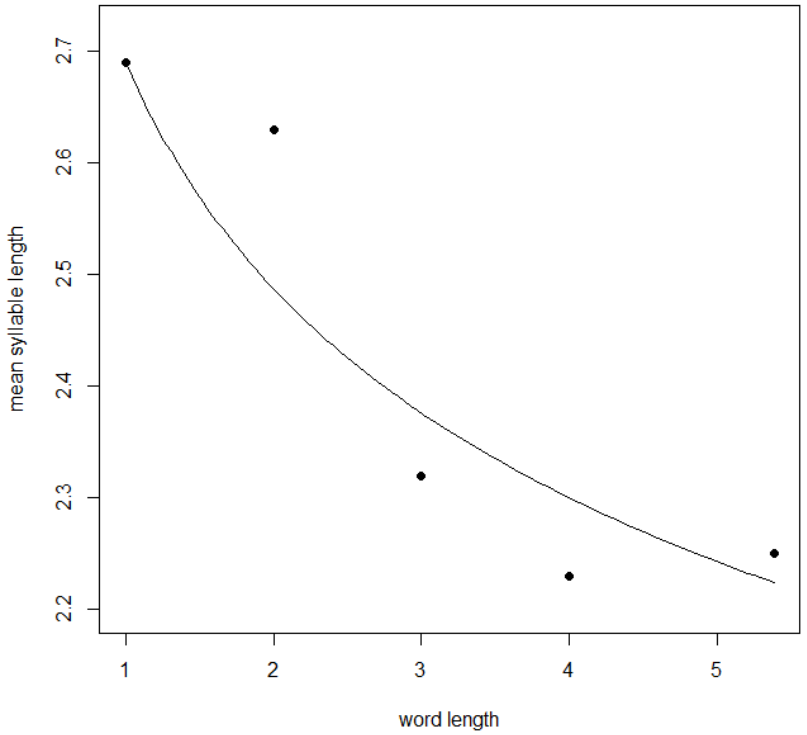
2. in Chinese simplified characters
   - word – character – components / strokes

word – syllable – phoneme

- tokens
  - $R^2 = 0.281$

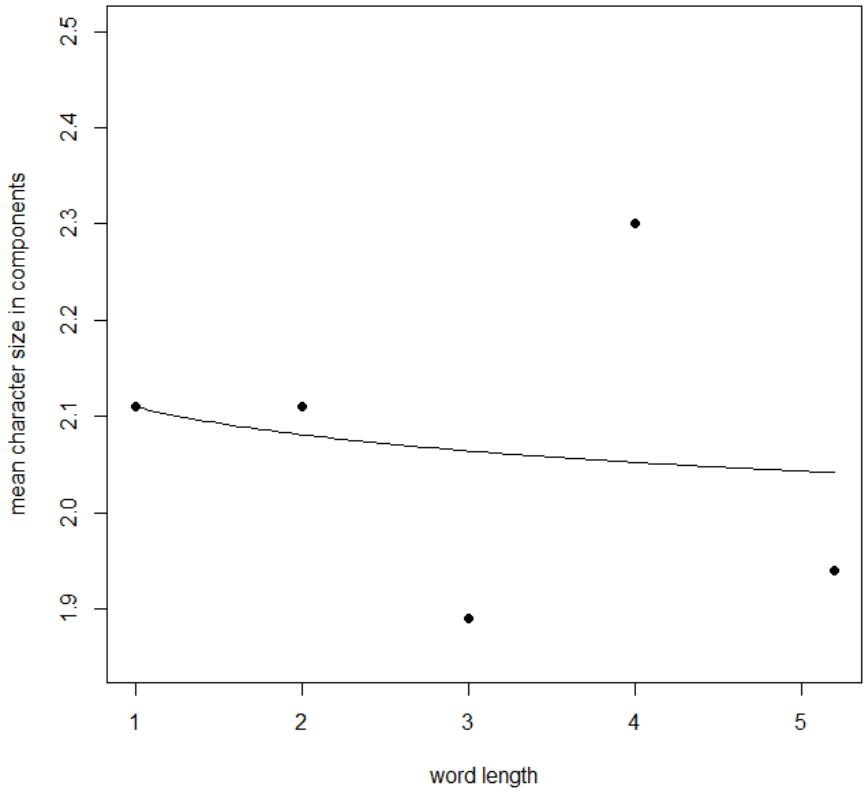


- types
  - $R^2 = 0.849$

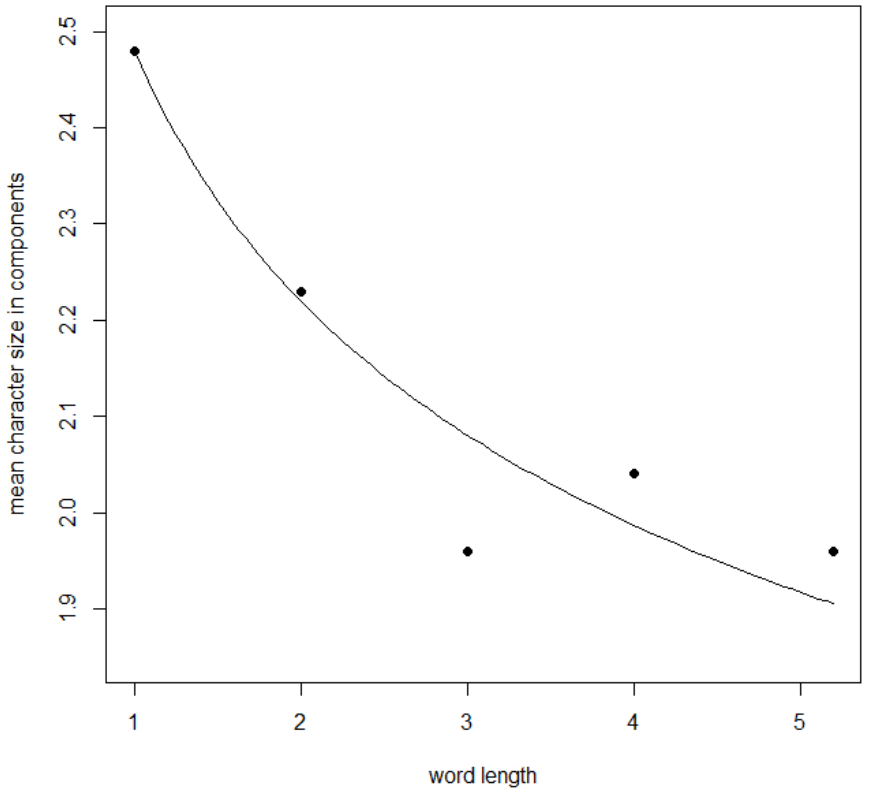


word – character – component

- tokens
  - $R^2 = 0.024$

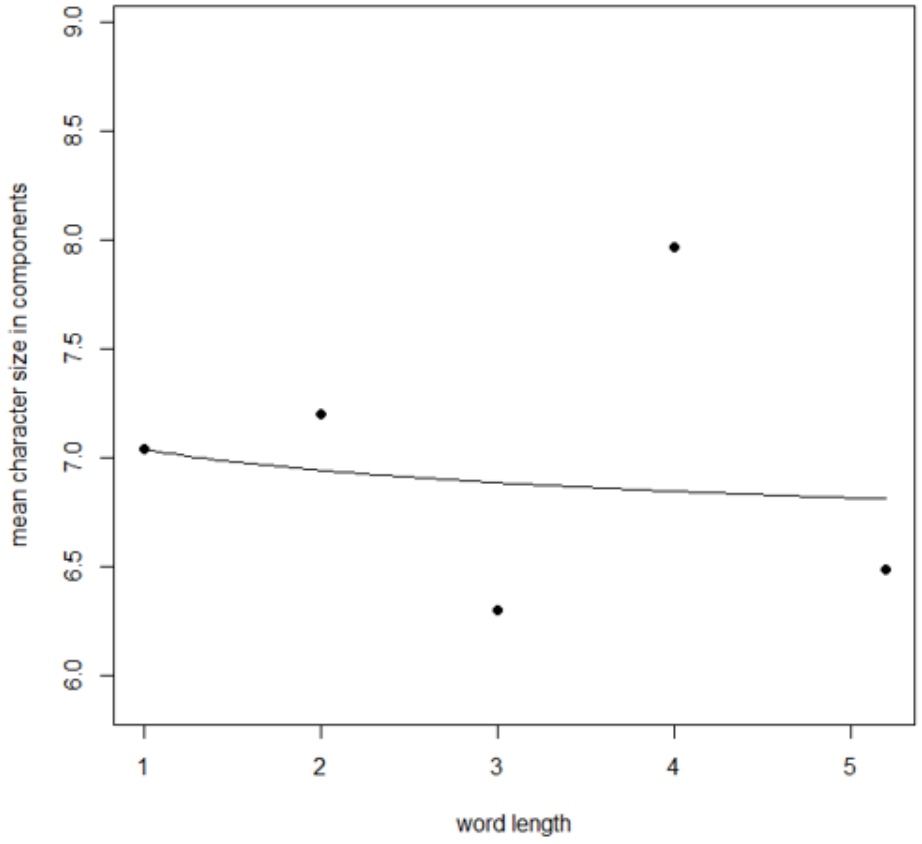


- types
  - $R^2 = 0.899$

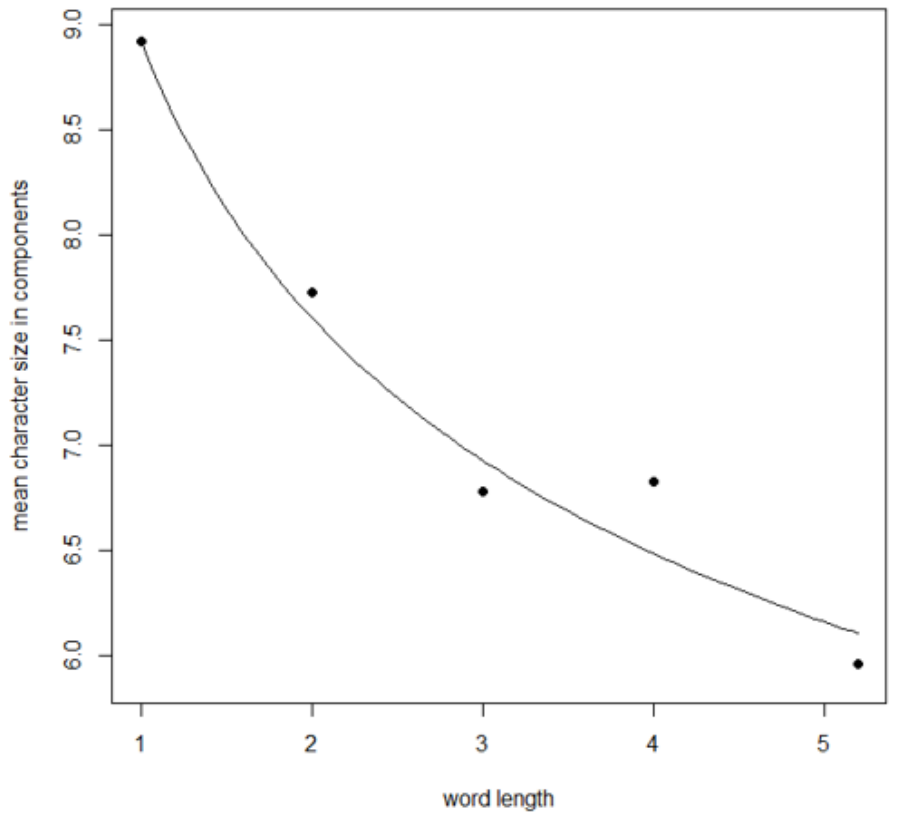


word – character – stroke

- tokens
  - $R^2 = 0.005$



- types
  - $R^2 = 0.966$



MAL: tokens – types

- the higher the level, the smaller the type-token ratio → the smaller impact of frequency → the lower deviation of the MAL

word

syntactic phrase

clause

sentence

MAL: tokens – types

- the higher the level, the smaller the type-token ratio → the smaller impact of frequency → the lower deviation of the MAL

word

syntactic phrase

clause

sentence

- type-token ratio could/should counted

MAL – challenges/problems

- what should be analyzed?
  - tokens / types / lemmas
- how the law should be analyzed?
  - look to whole structure
- do we know/analyze proper levels?
  - clause - ??? - word

MAL: how many levels to be analyzed?

SENTENCE

CLAUSE

SYNTACTIC/CLAUSE PHRASE

WORD

MORPHEME

PHONEME

MAL: number of levels to be analyzed

SENTENCE

CLAUSE

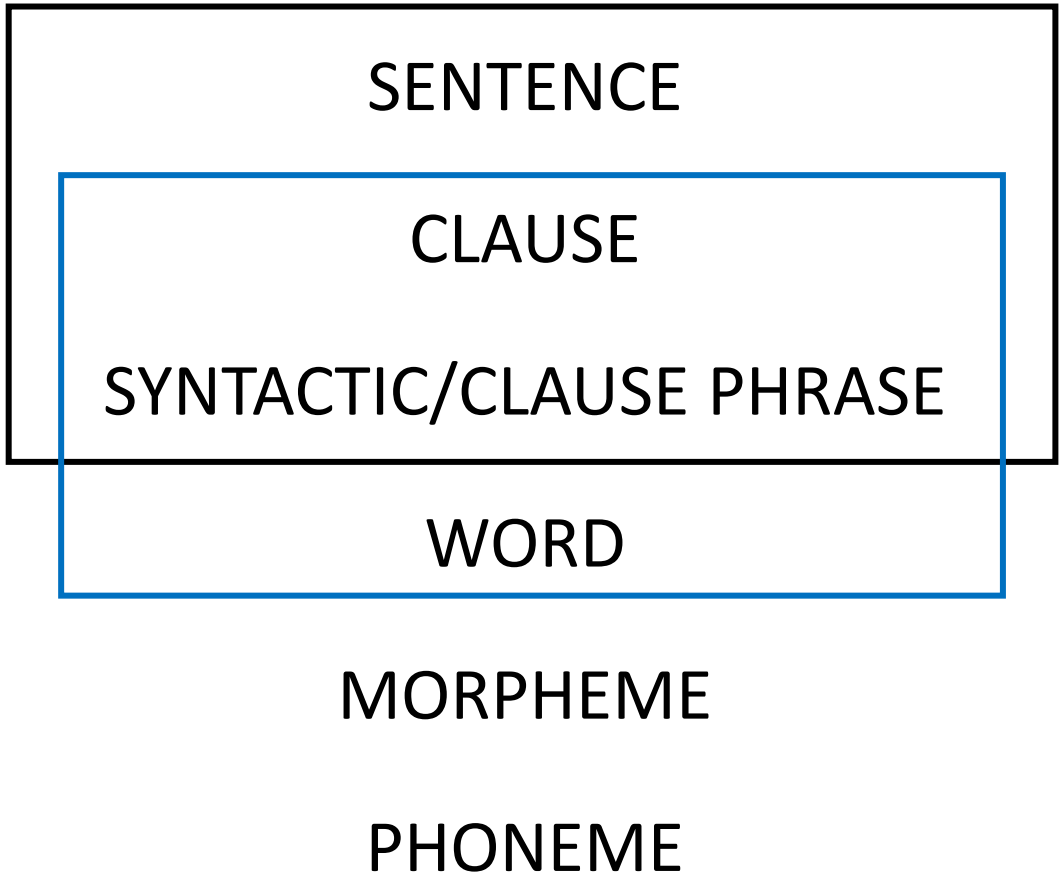
SYNTACTIC/CLAUSE PHRASE

WORD

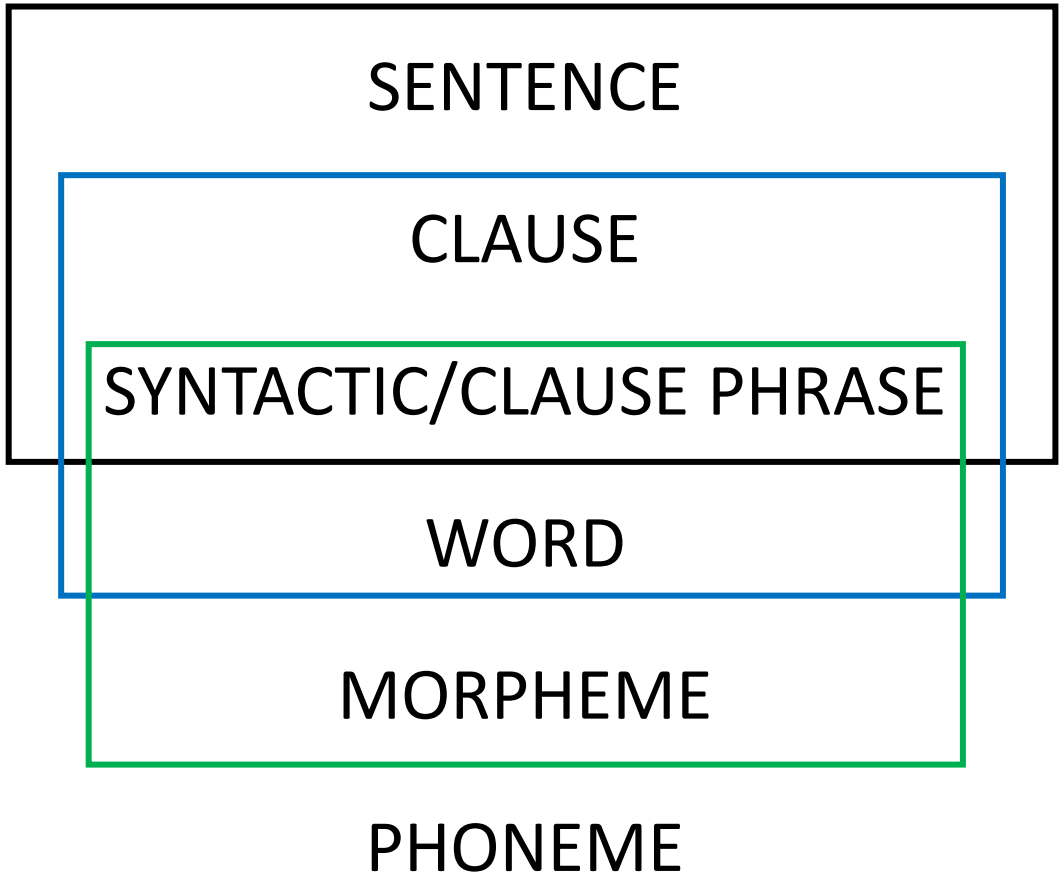
MORPHEME

PHONEME

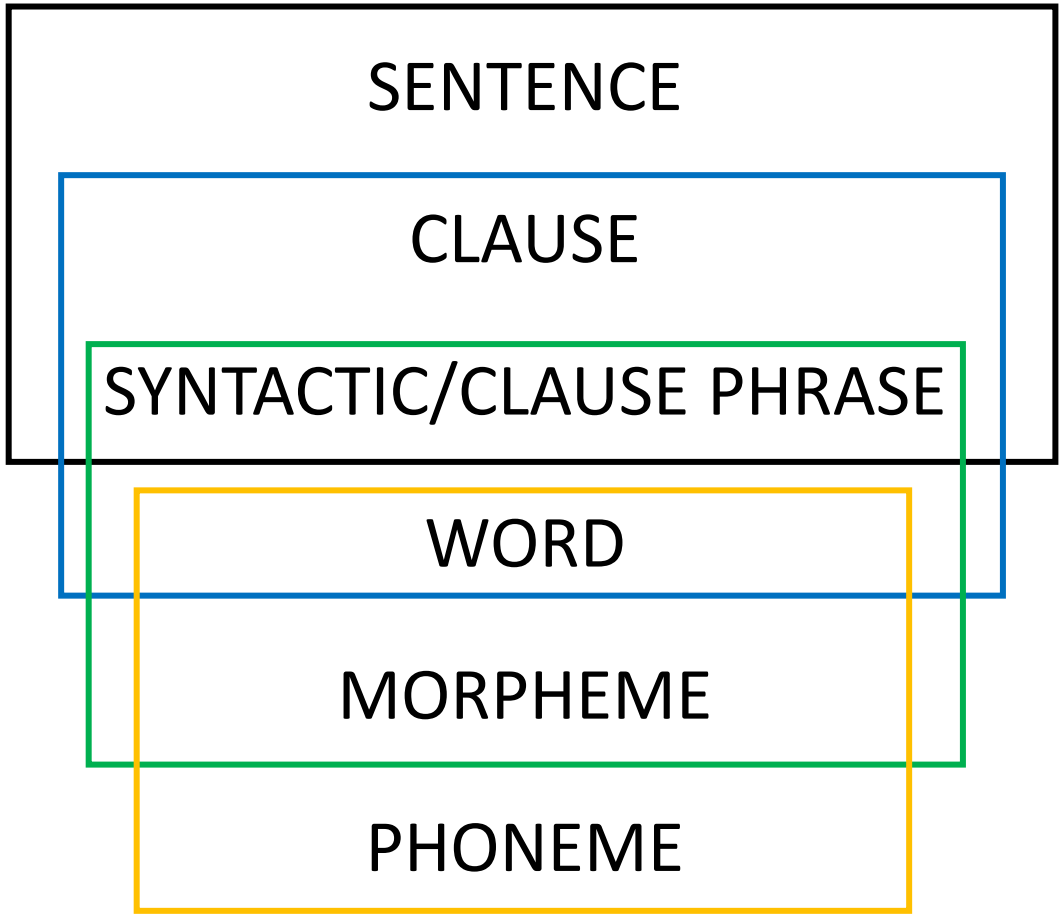
MAL: number of levels to be analyzed



MAL: number of levels to be analyzed



MAL: number of levels to be analyzed

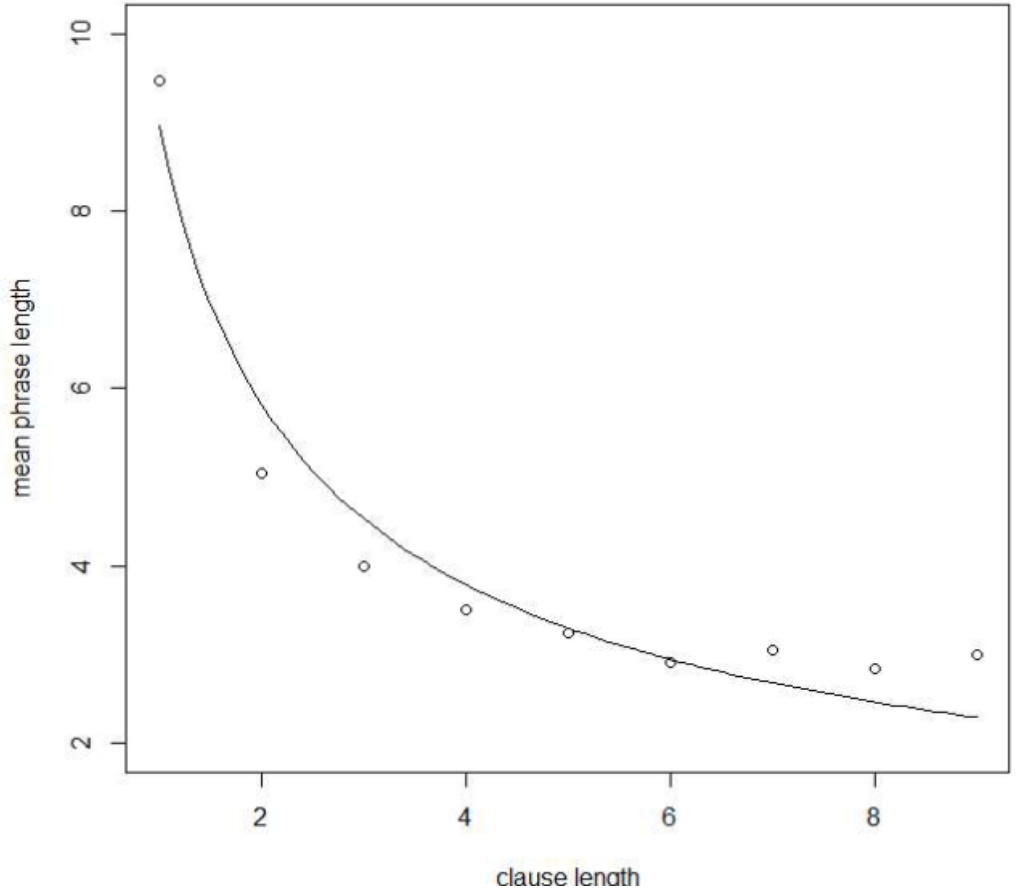


MAL: number of levels to be analyzed

- the MAL may be valid for one triplet, but not for the immediately following one

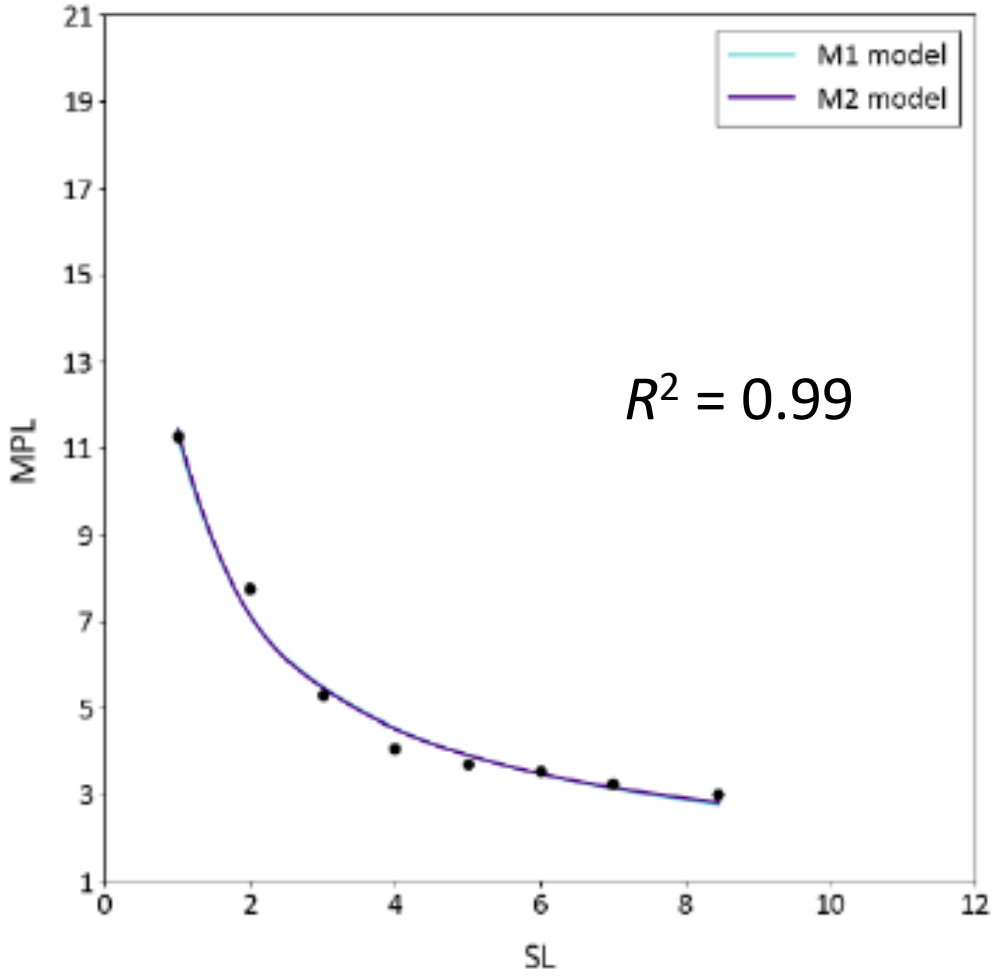
MAL: number of levels to be analyzed

Mačutek, J., Čech, R., Milička, J. (2017). **Menzerath-Altmann Law in Syntactic Dependency Structure.** In Proceedings of the Fourth International Conference on Dependency Linguistics

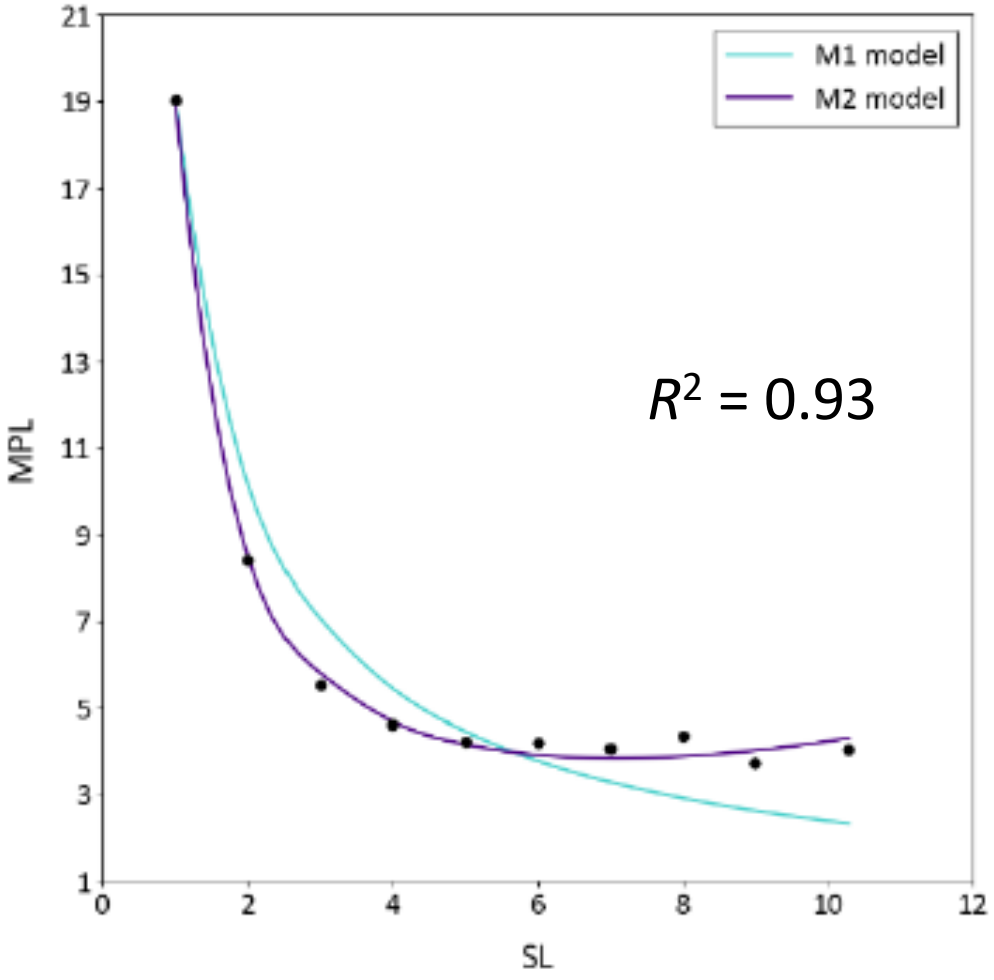


$R^2 = 0.94$

sentence – phrase – word (Chinese)

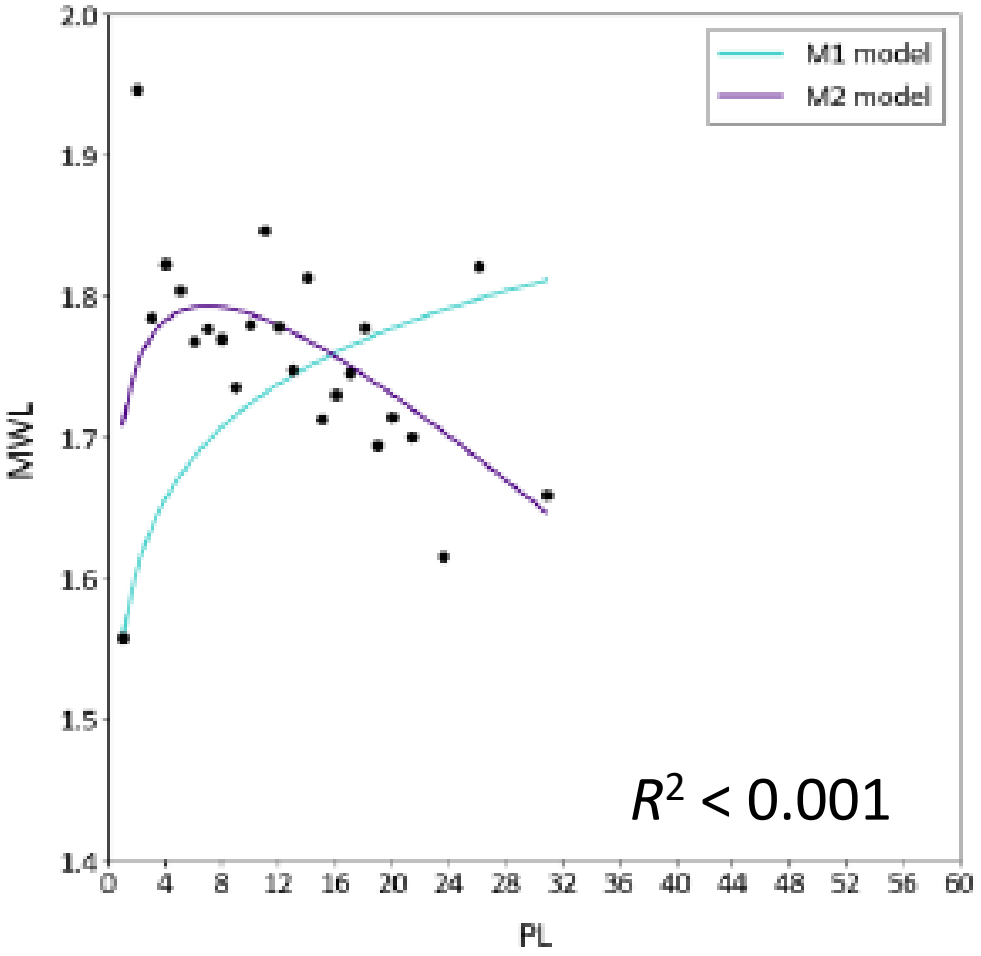


PUD

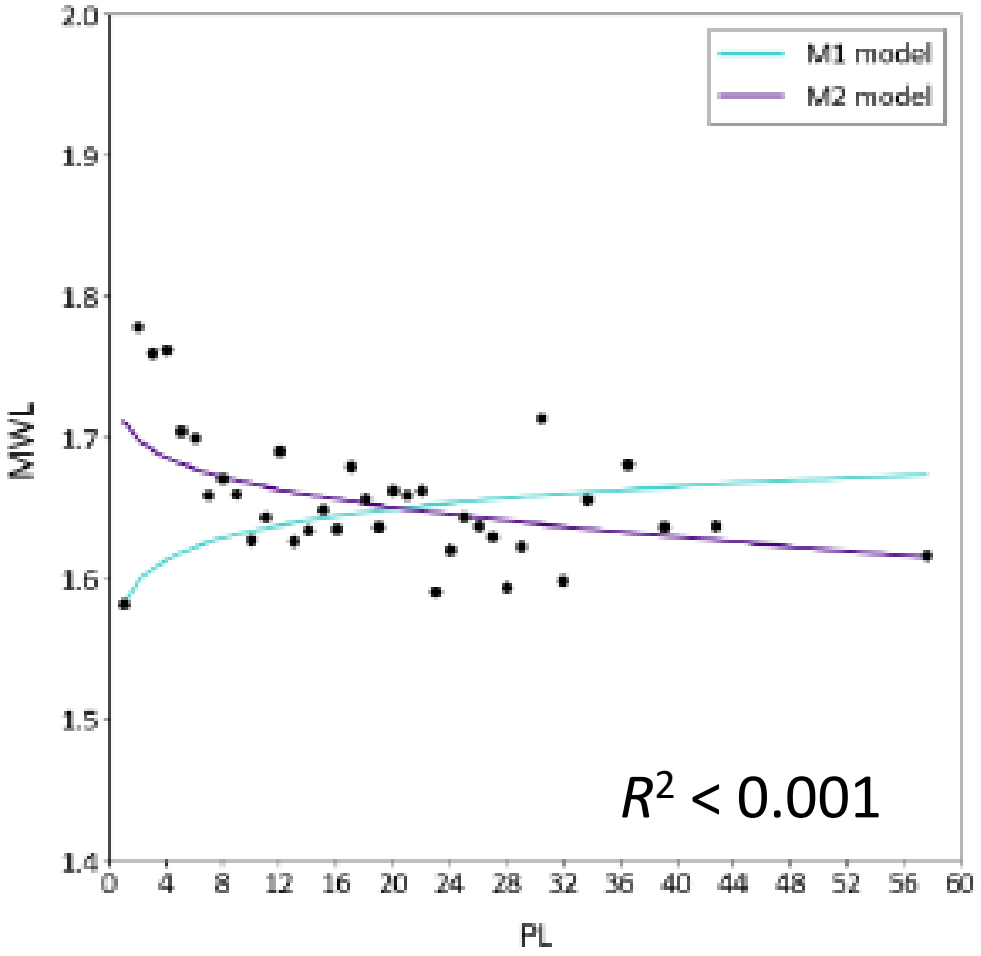


GSD

phrase – word – character (Chinese)

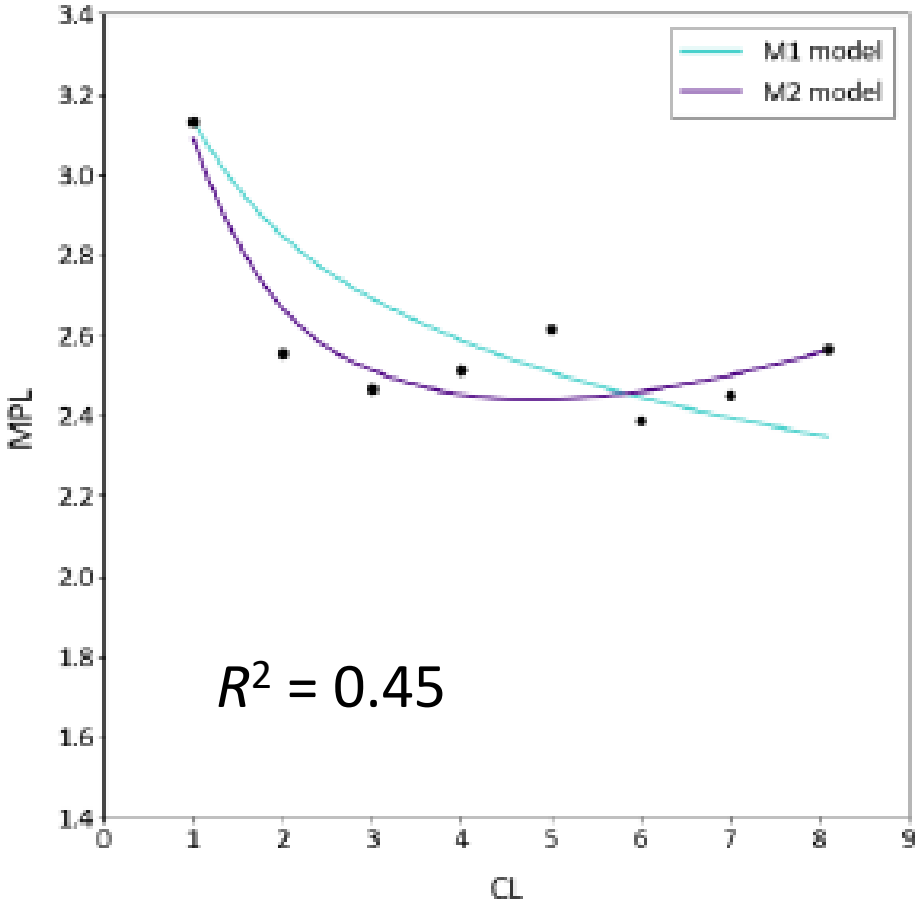


PUD

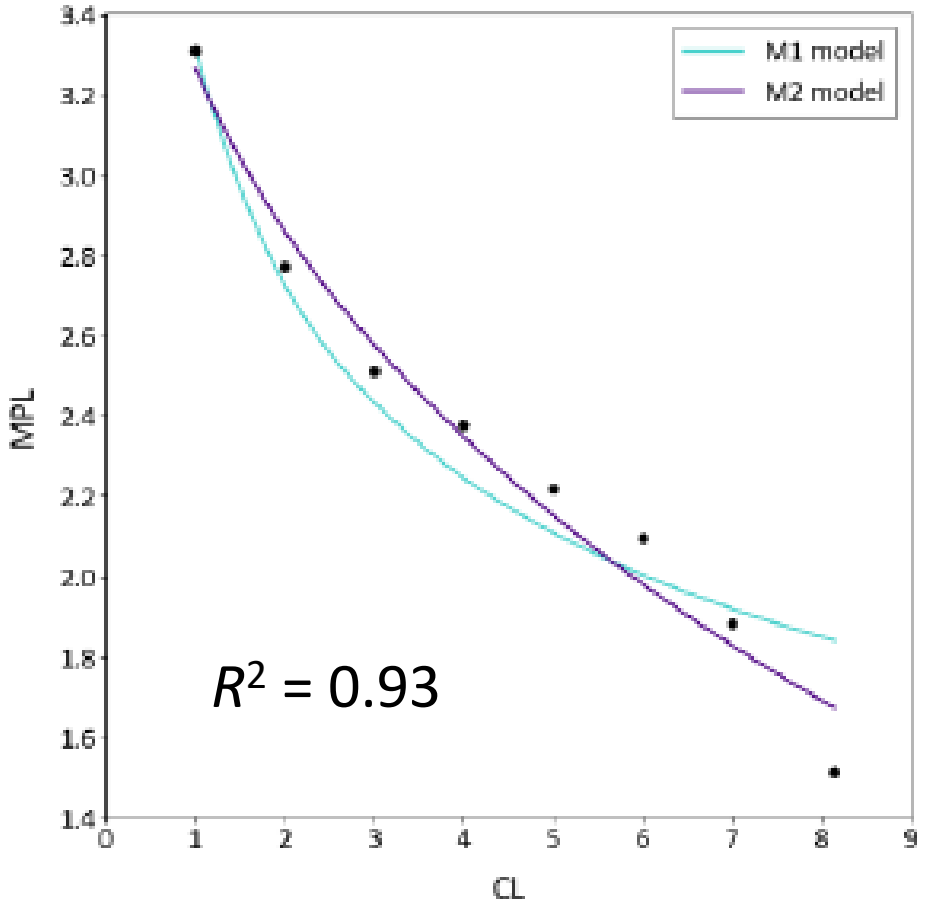


GSD

clause – phrase – word (Chinese)

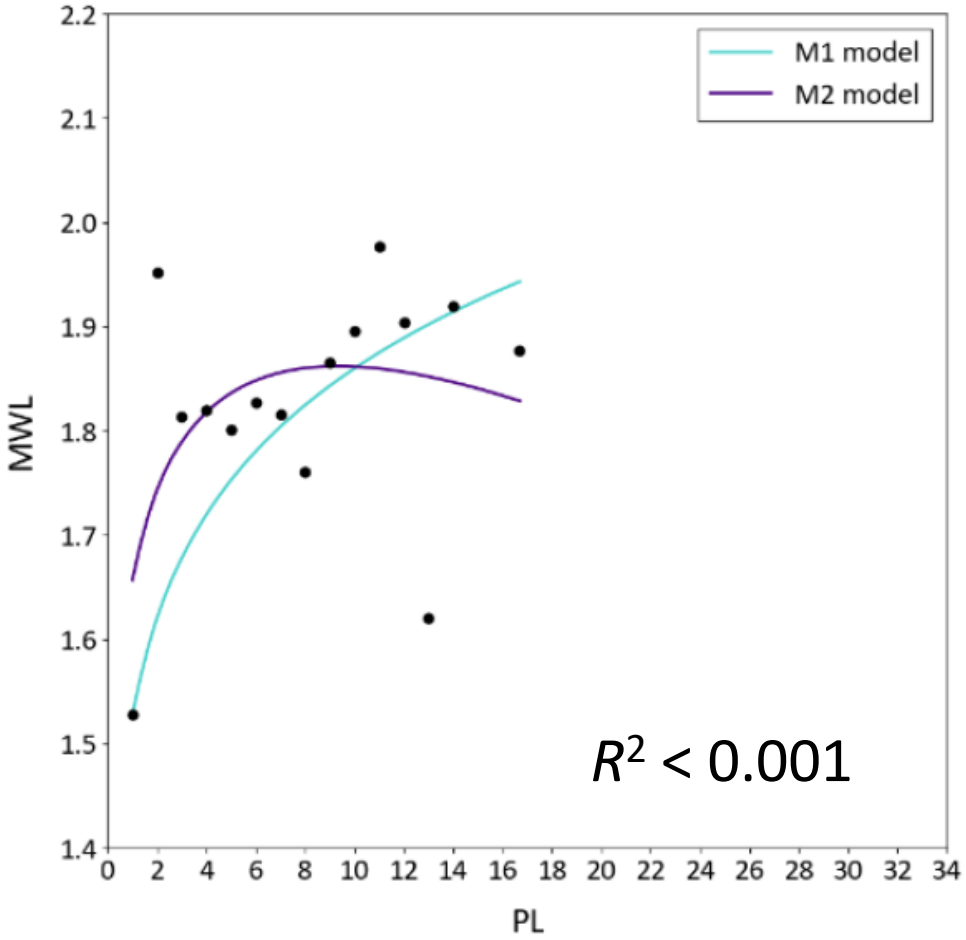


PUD

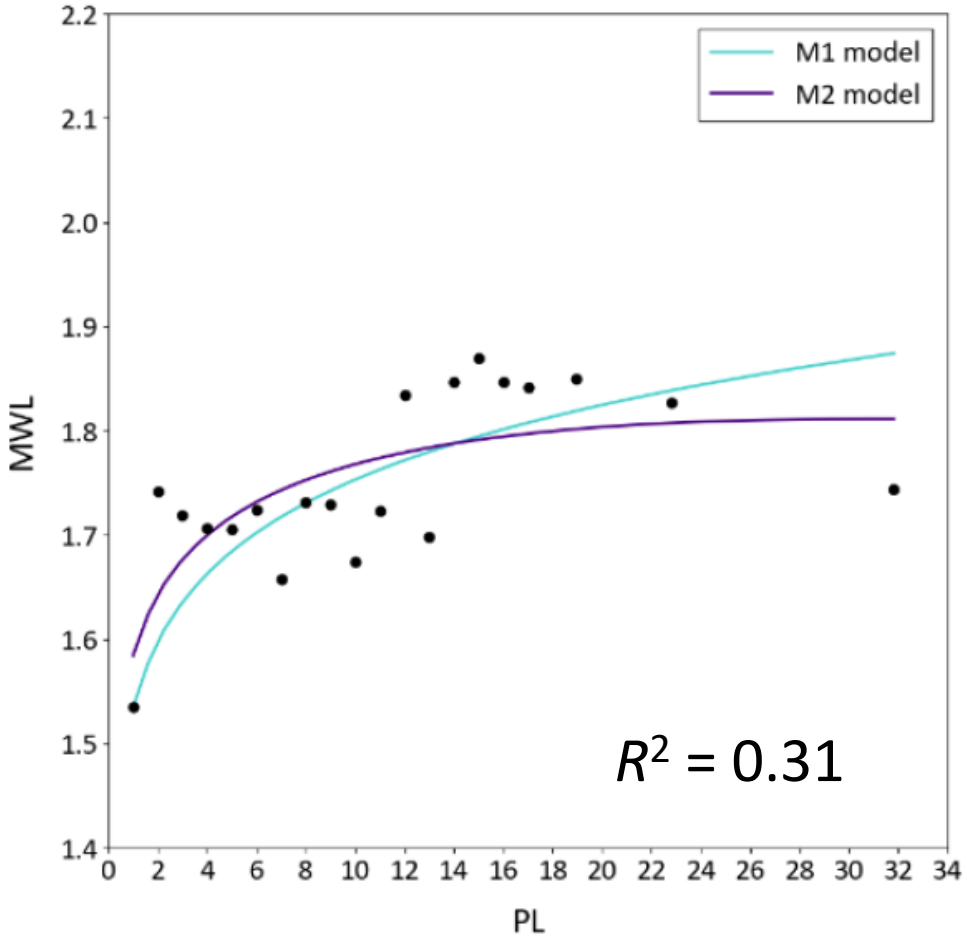


GSD

phrase – word – charachter (Chinese)

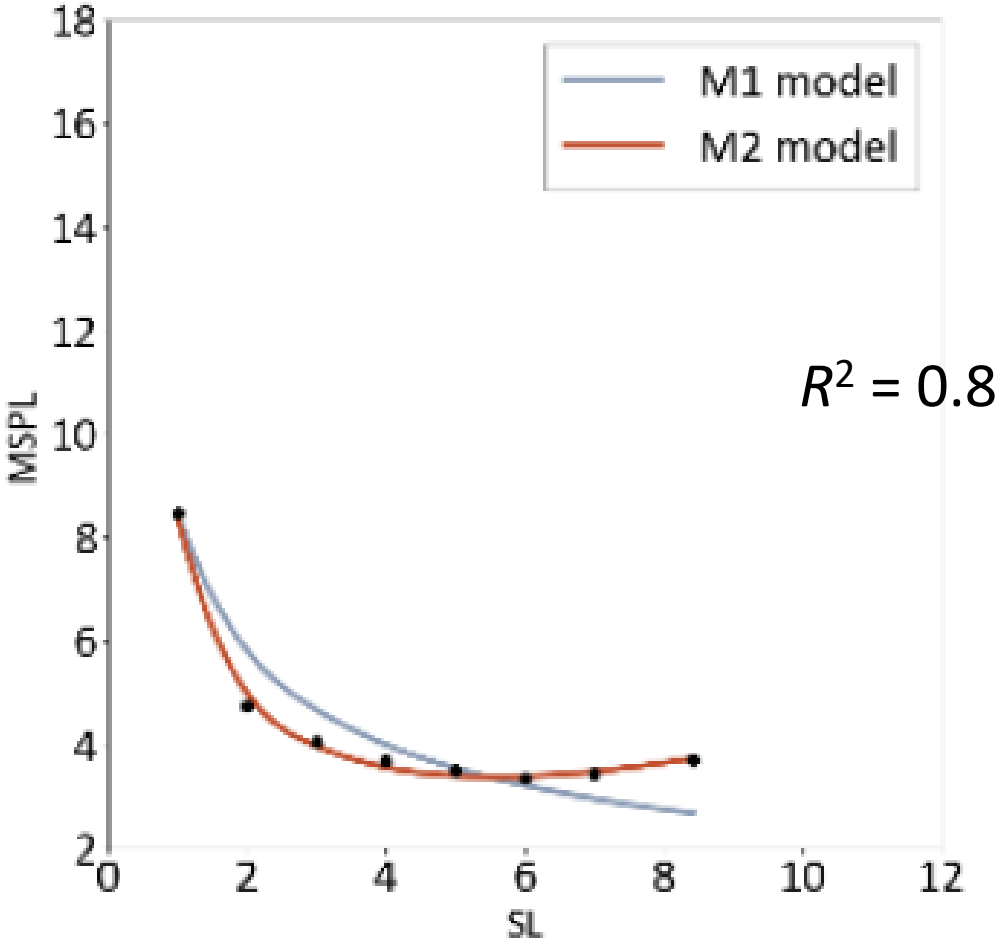


PUD



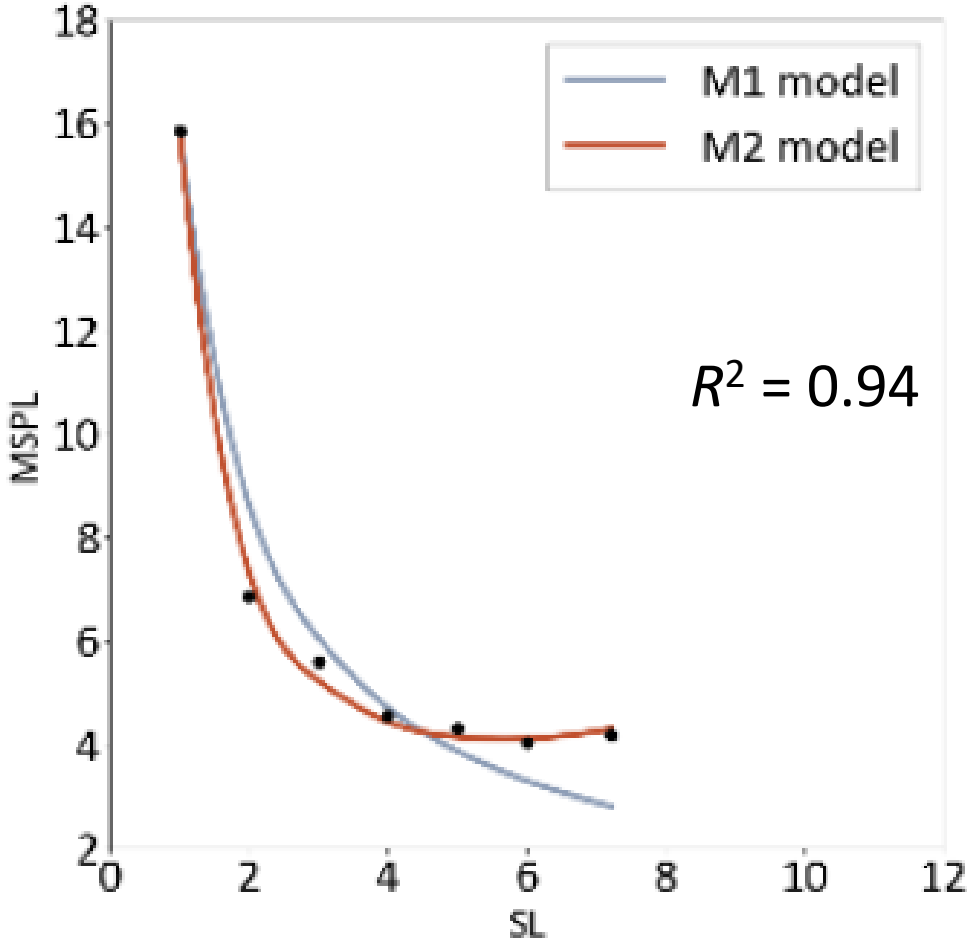
GSD

sentence – phrase – word (Czech)



$R^2 = 0.86$

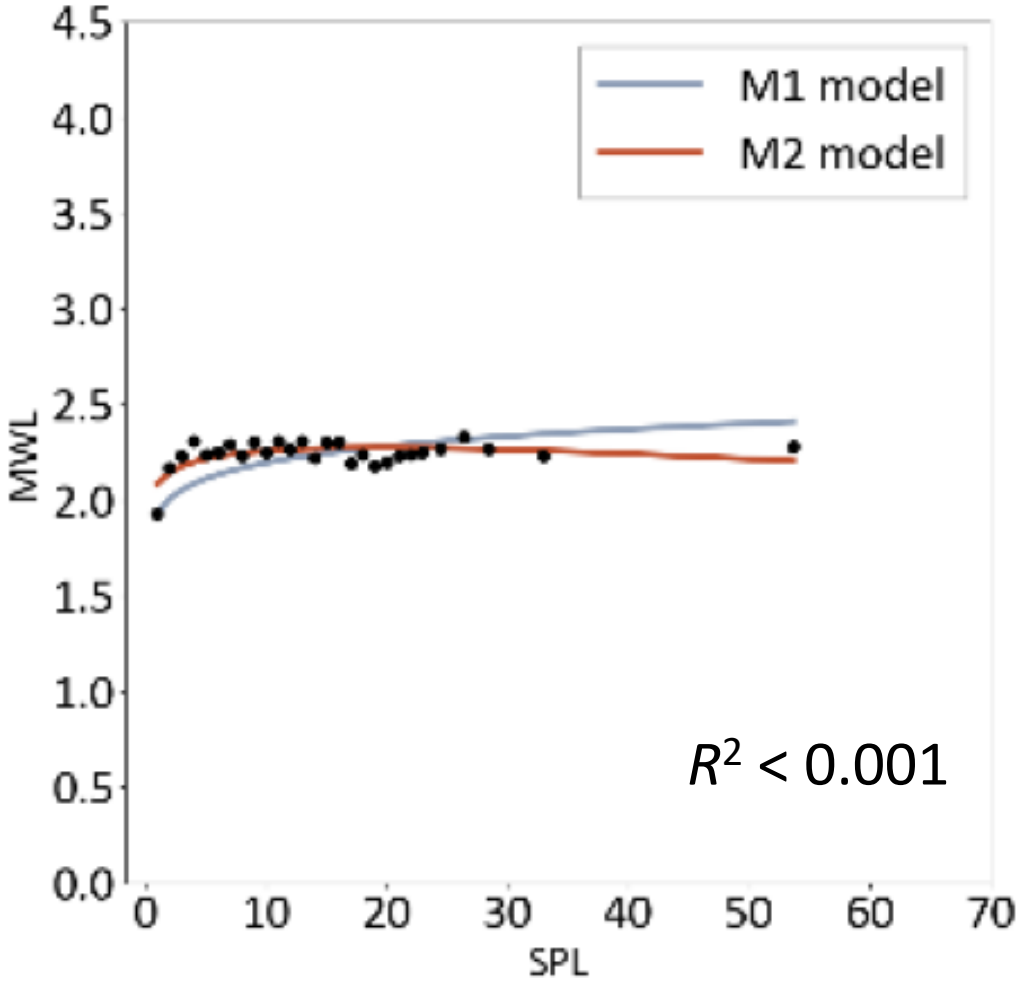
essays



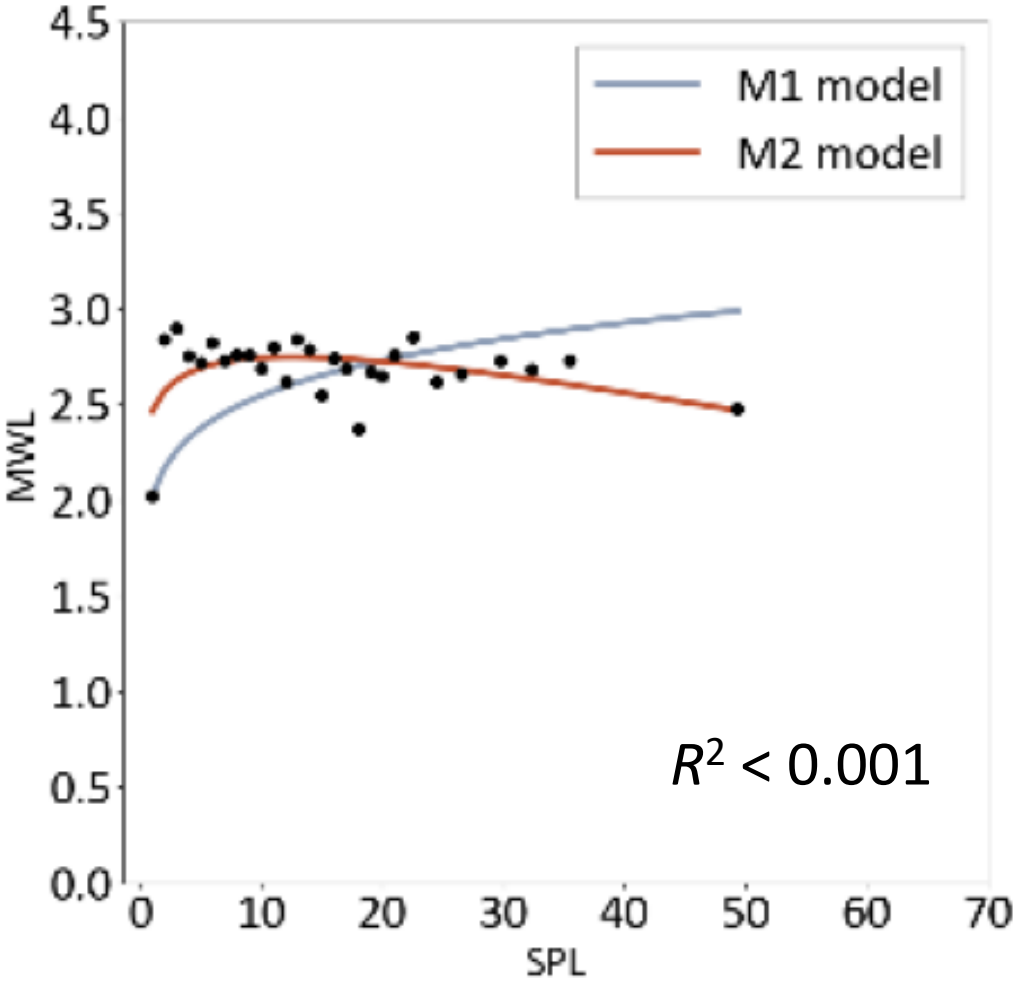
$R^2 = 0.94$

sci texts

phrase – word – morpheme (Czech)



essays



sci texts

MAL: how many levels to be analyzed?

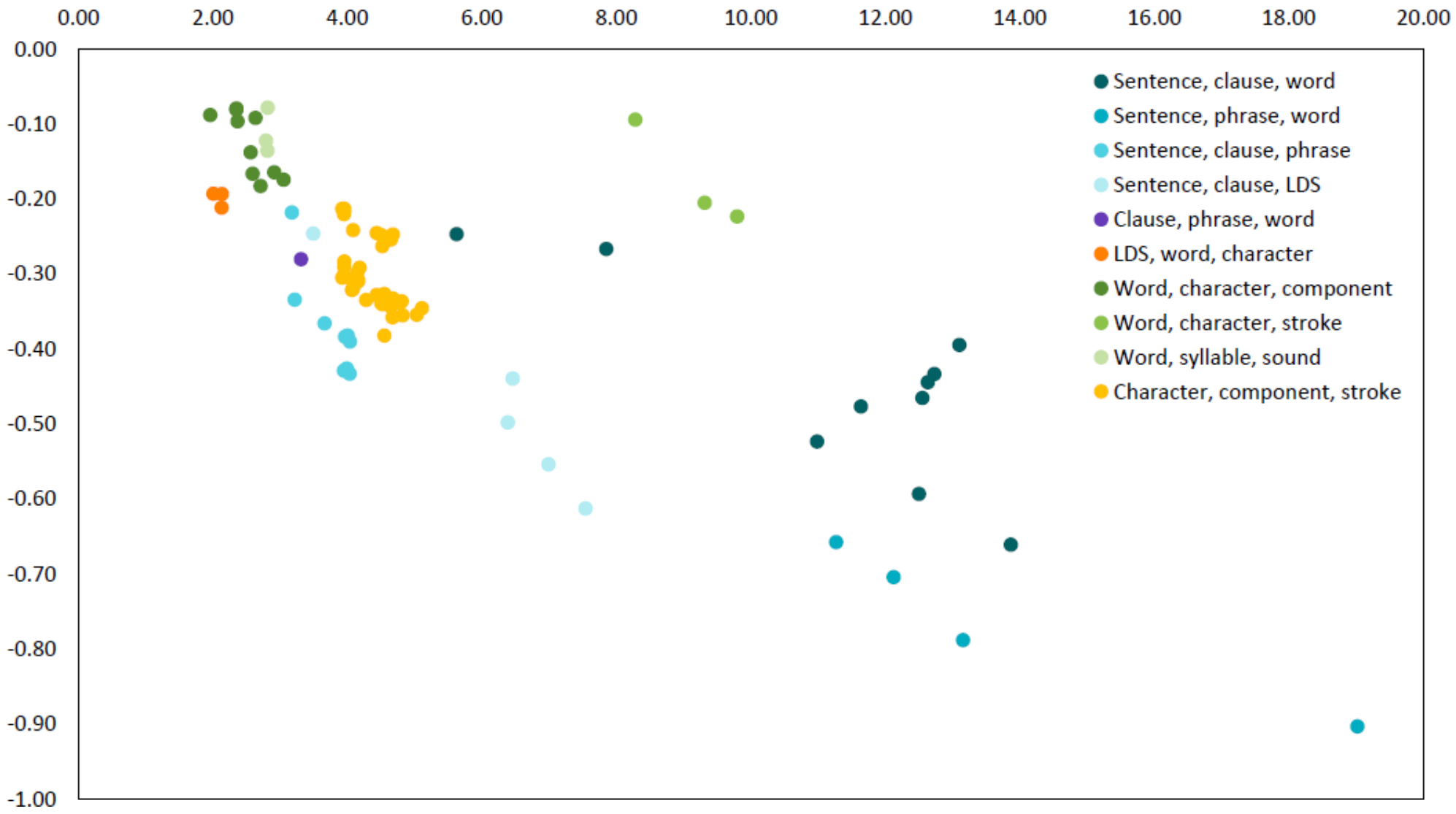
- as much as possible…

MAL: how many levels to be analyzed?

- as much as possible…

- analysis of more leveles → comparison and interpretaion of parameters

Parameters (Chinese)



MAL – challenges/problems

- what should be analyzed?
  - tokens / types / lemmas
- how the law should be analyzed?
  - look to whole structure
- do we know/analyze proper levels?
  - clause - ??? - word

MAL: the levels to be analyzed

SENTENCE

CLAUSE

WORD

SYLLABLE MORPHEME

PHONEME

MAL: the levels to be analyzed

(SEMANTIC AGREGGATE / HREB)

SENTENCE

CLAUSE

WORD

SYLLABLE MORPHEME

PHONEME

(PHONEME/SOUND DURATION)

MAL: the levels to be analyzed

SENTENCE

CLAUSE / ???

???

WORD

MORPHEME

PHONEME

Syntactic segmentation

sentence

↓

???

↓

???

Syntactic segmentation

sentence

↓

clause

↓

???

Syntactic segmentation

sentence

↓

clause

↓

word

Syntactic segmentation

sentence

↓

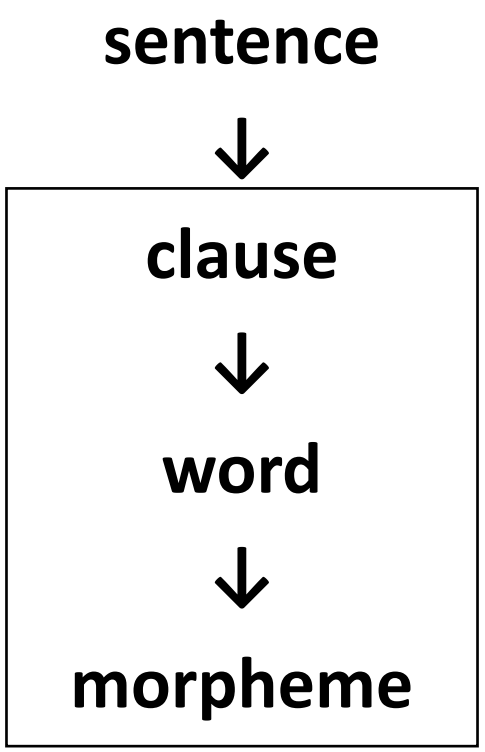
clause

↓

word

ambiguous results

Syntactic segmentation

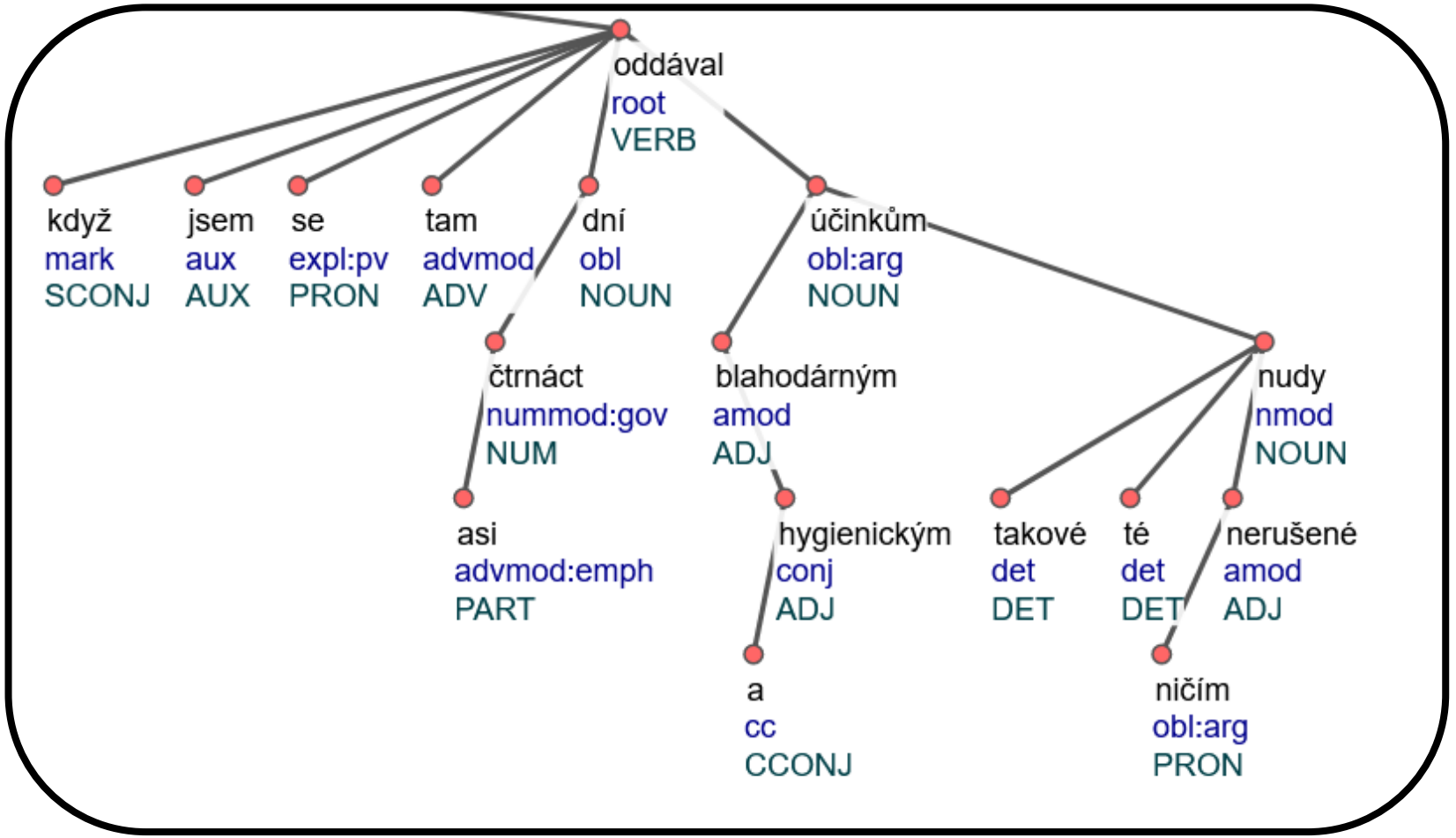


ambiguous results

working memory → length of the construct

1 clause

17 words



Syntactic segmentation

sentence

↓

clause

↓

word

sentence

↓

???

↓

???

Syntactic segmentation

sentence

↓

clause

↓

word

sentence

↓

syntactic phrase

↓

???

Syntactic segmentation

sentence

↓

clause

↓

word

sentence

↓

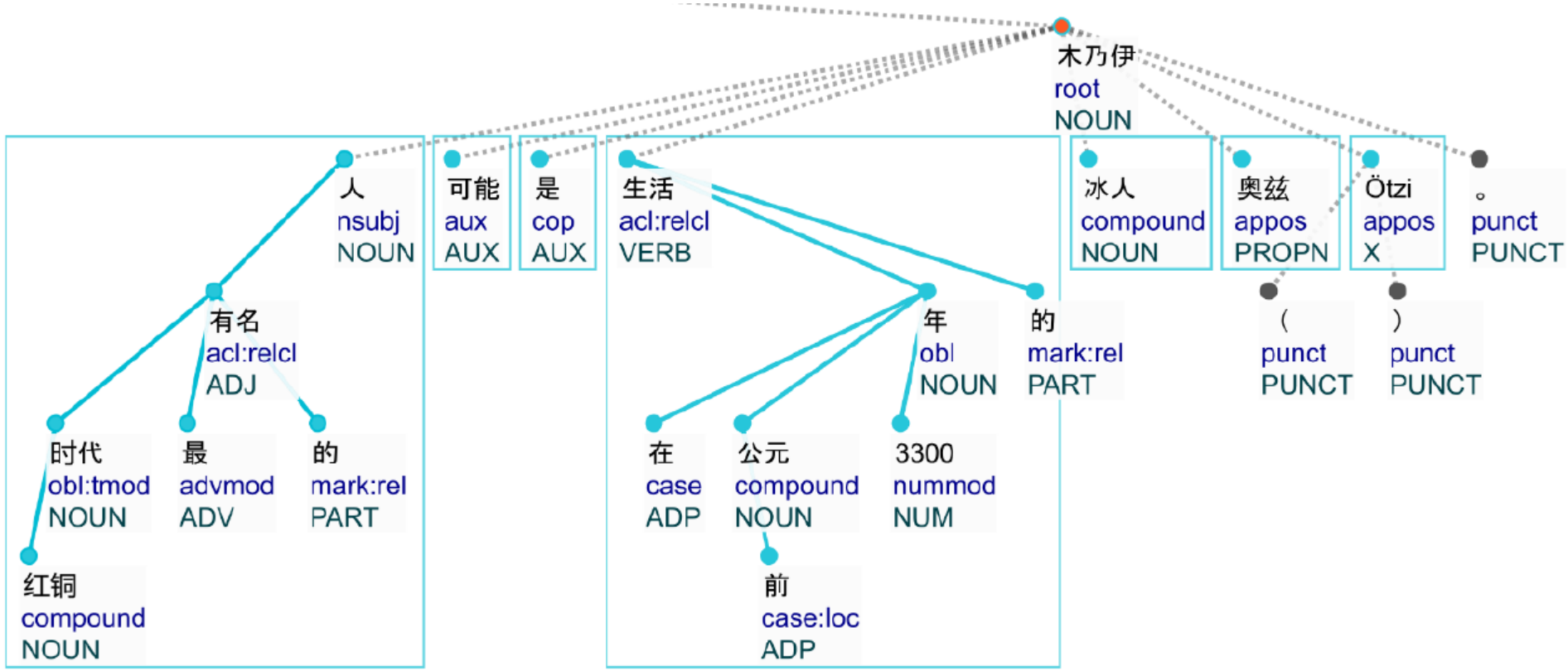
syntactic phrase

↓

word

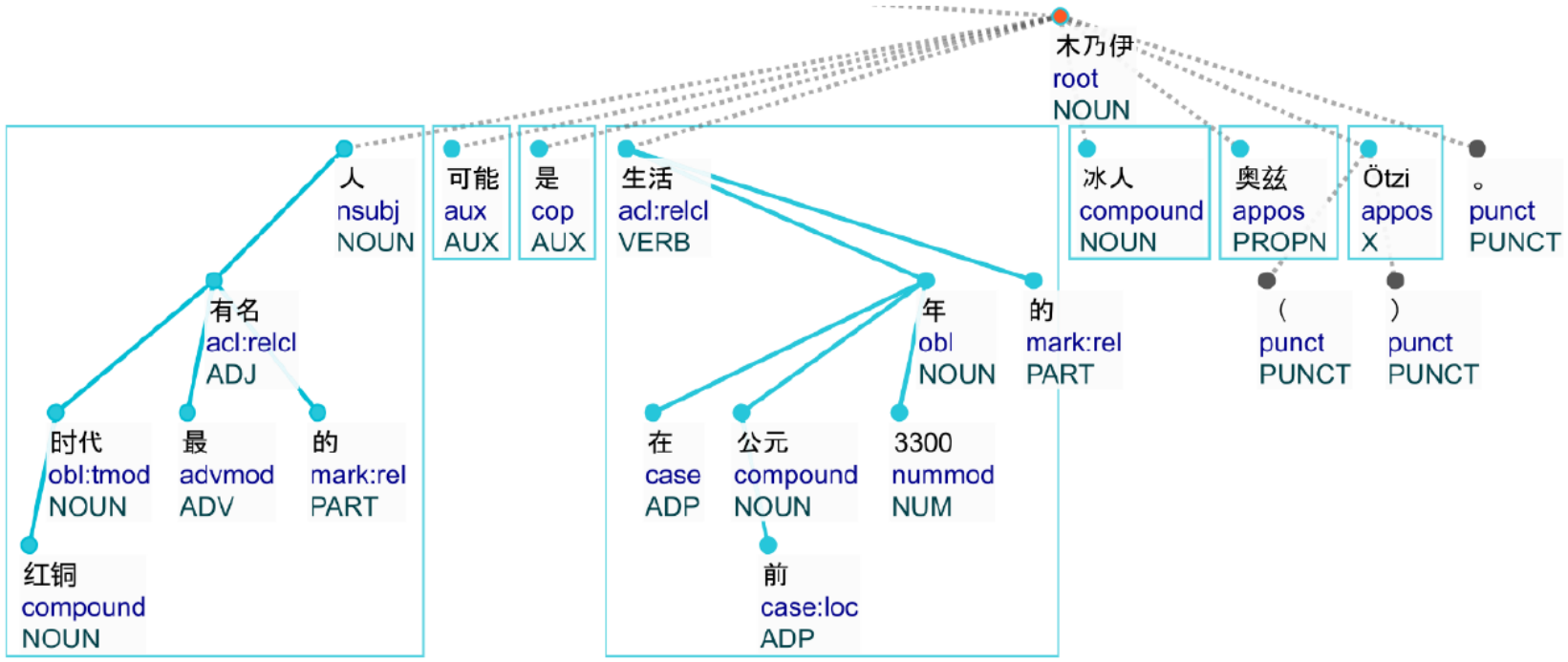
Syntactic phrase

- a subtree that is immediately dependent on the predicate of the main clause



Syntactic phrase

- problems
  - predicate is not counted
  - what about one-word sentences?
    - size = 0?



Syntactic segmentation

sentence

↓

clause

↓

clause phrase

↓

word

sentence

↓

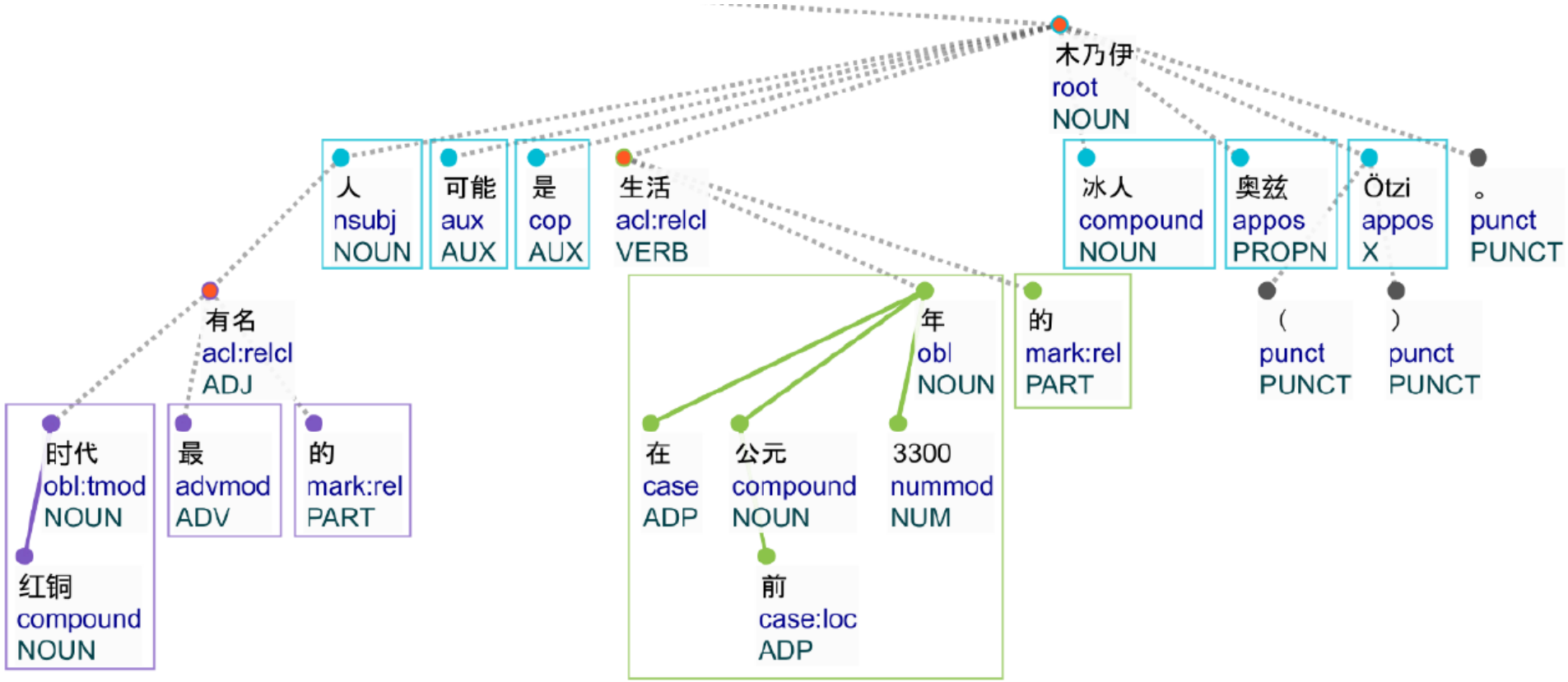
syntactic phrase

↓

word

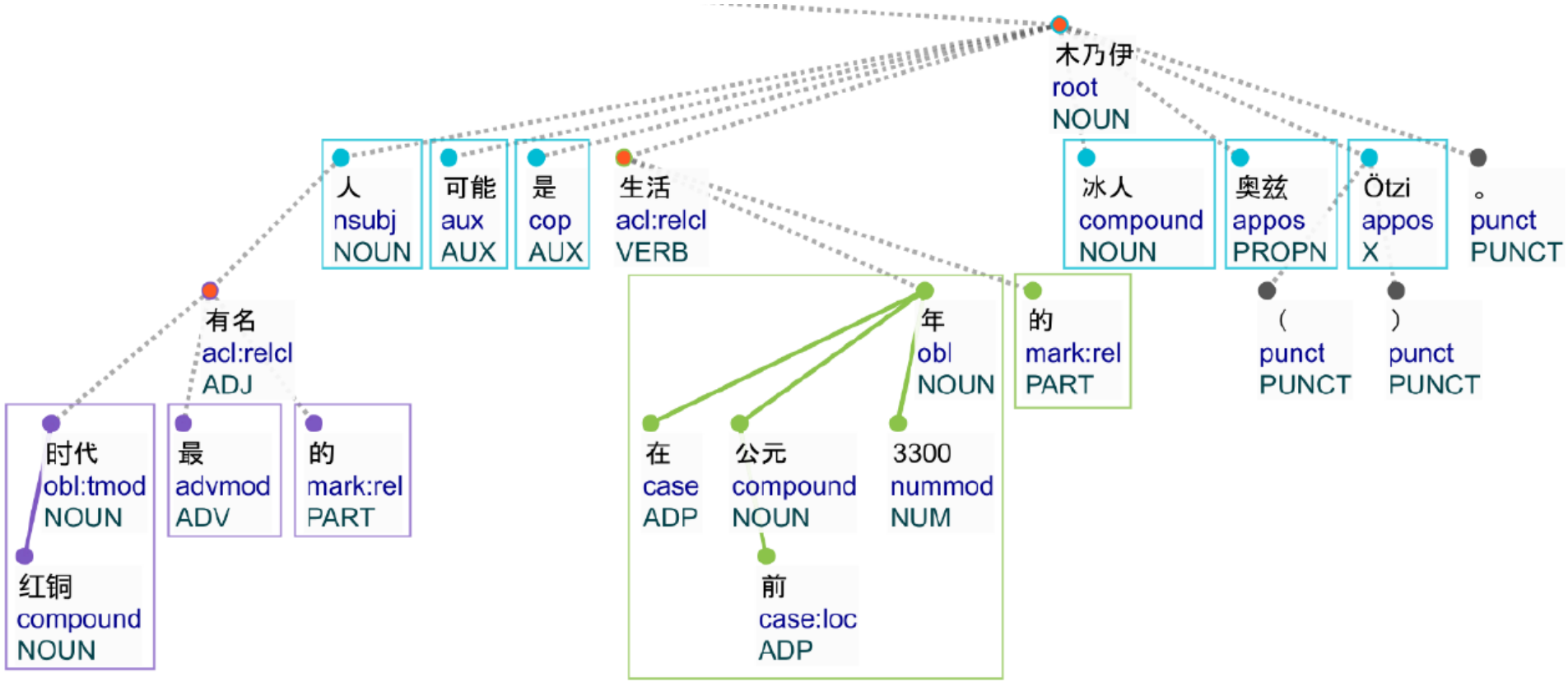
Clause phrase

- a subtree that is immediately dependent on the predicate of a clause

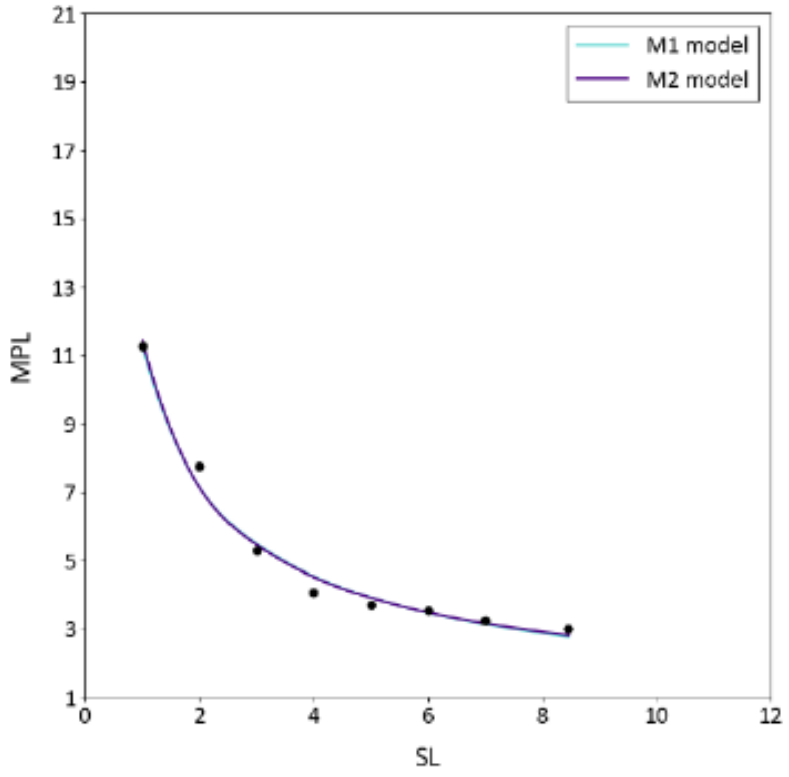


Clause phrase

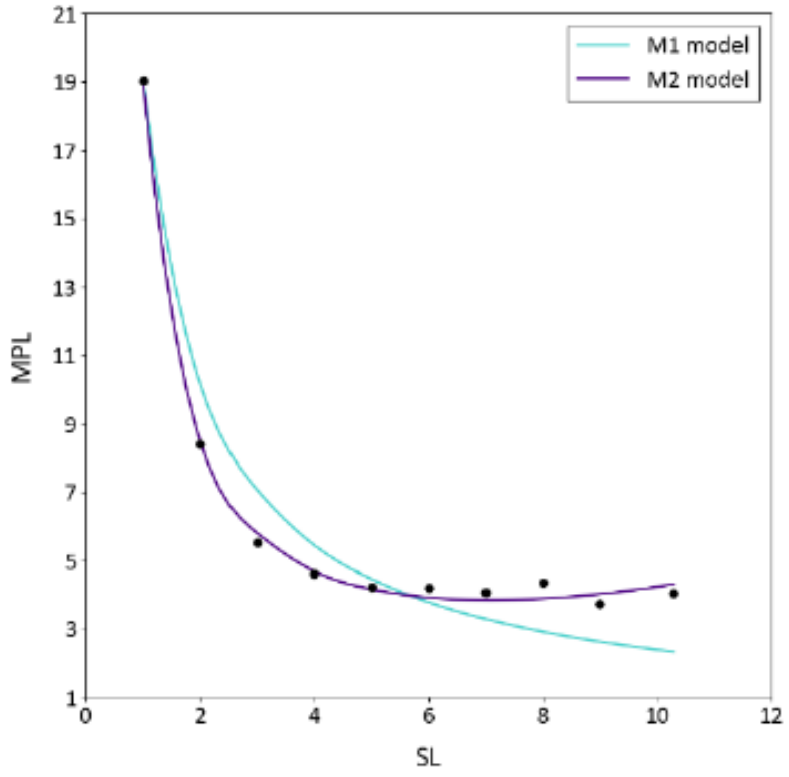
- the same problems as the sentence



sentence – phrase – word (Chinese)

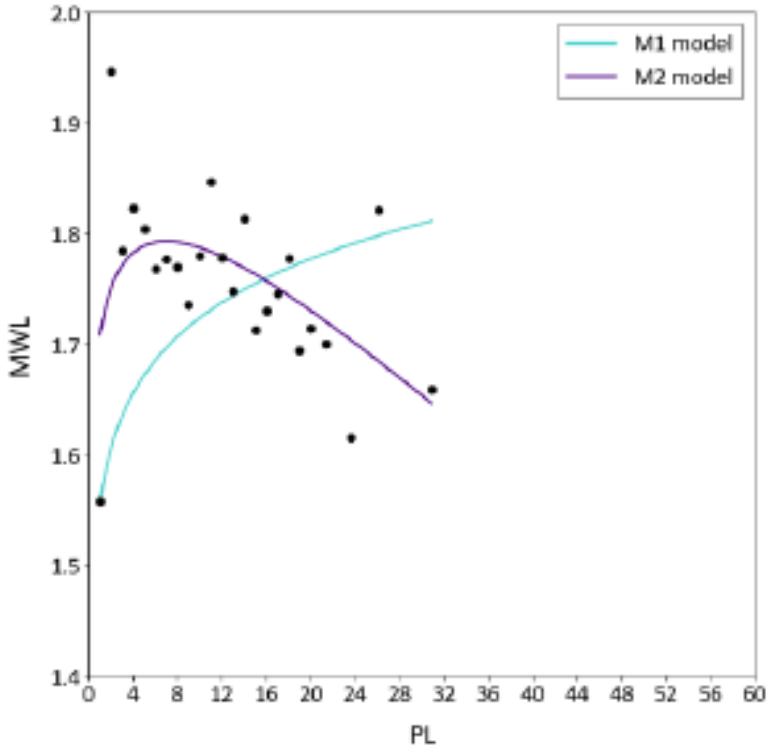


PUD

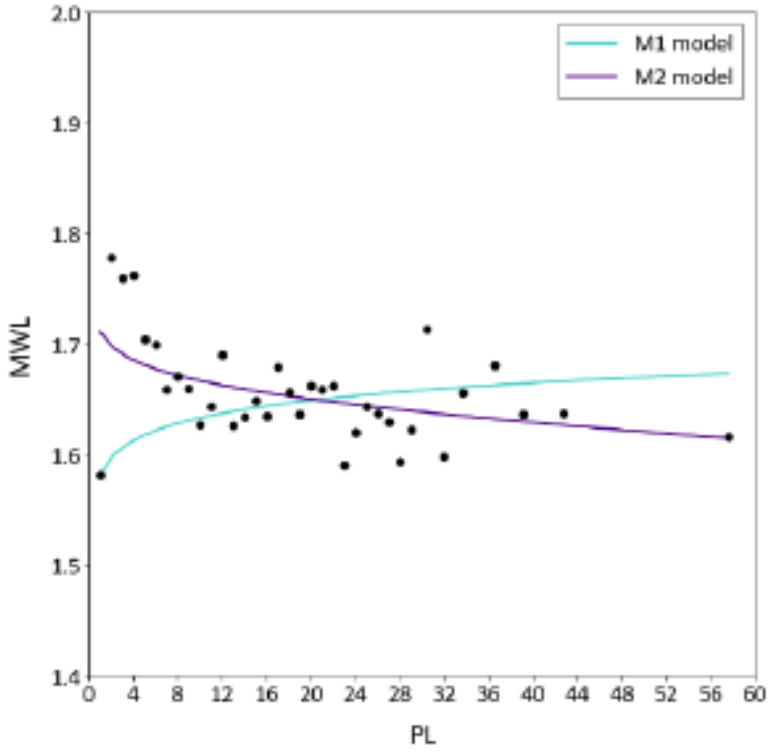


GSD

phrase – word – character (Chinese)



PUD



GSD

MAL & syntax

sentence

↓

clause

↓

word

sentence

↓

clause

↓

clause phrase

↓

word

sentence

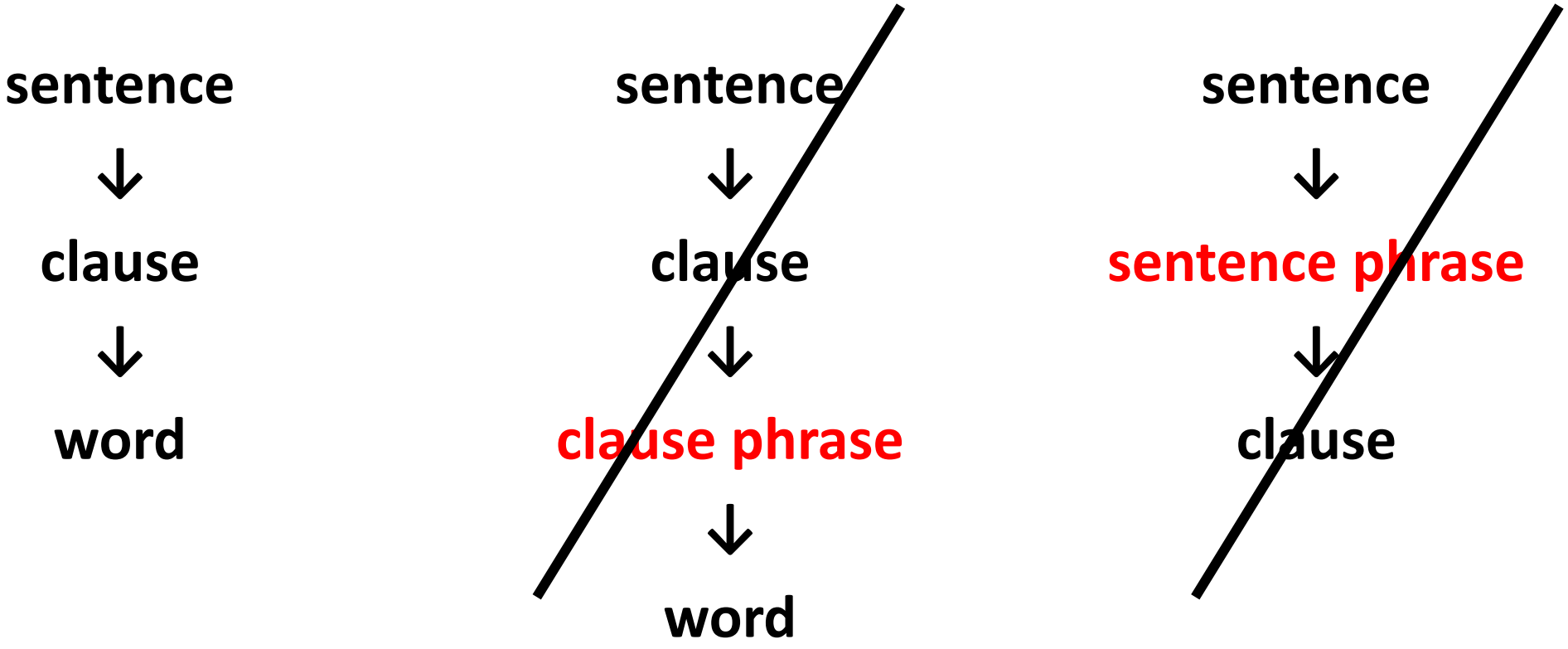
↓

sentence phrase

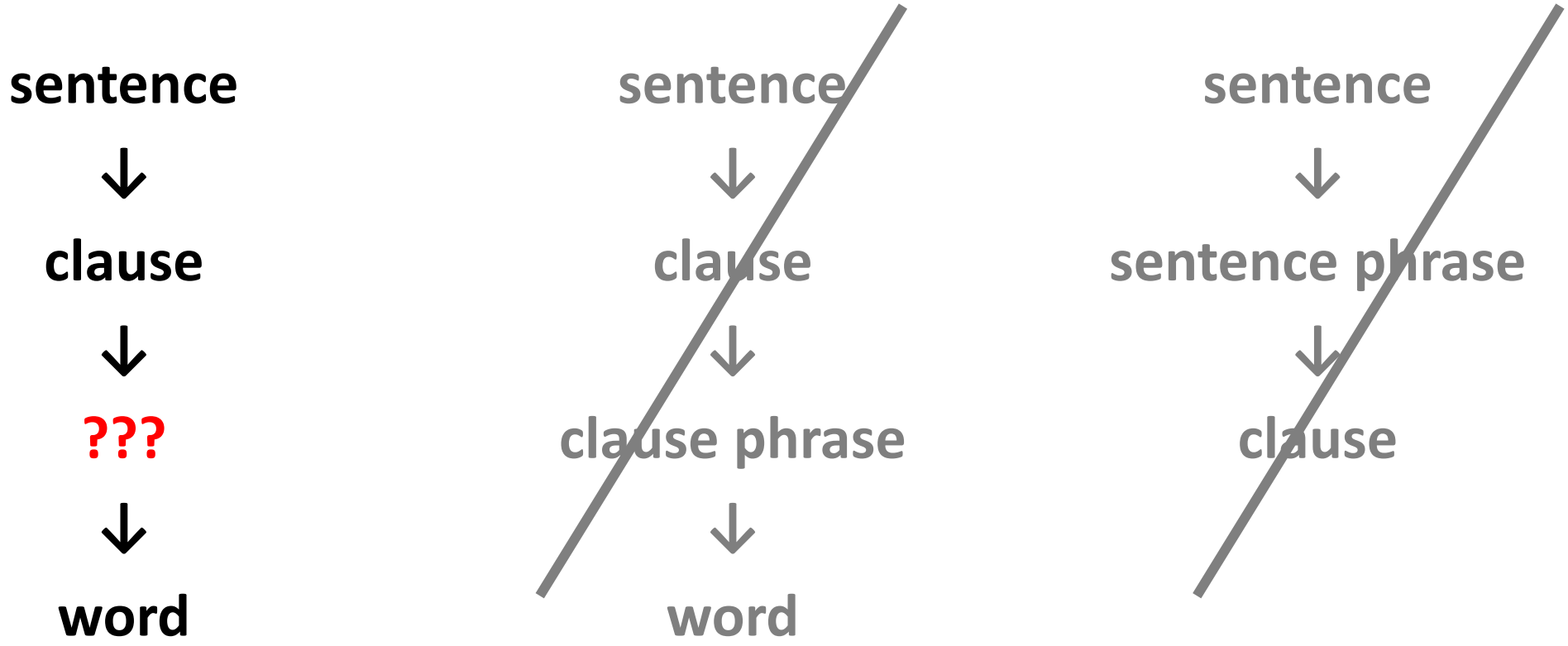
↓

clause

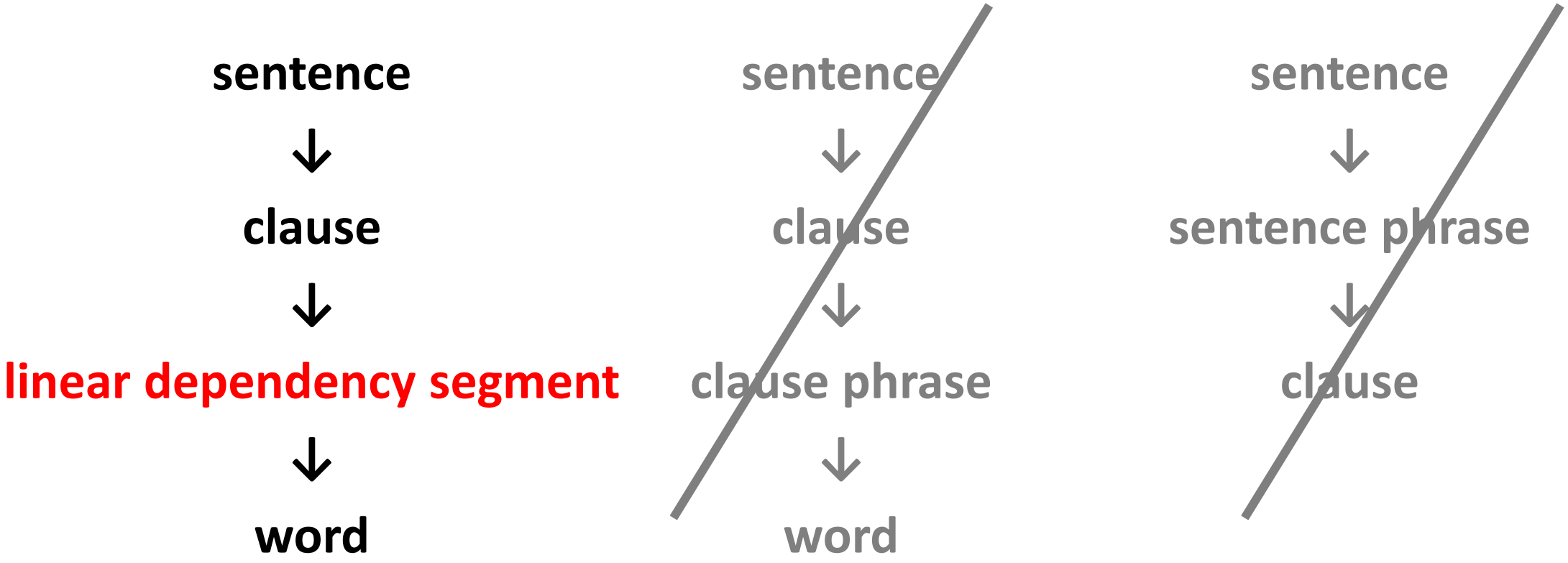
MAL & syntax


sentence
clause
word
sentence
clause
clause phrase
word
sentence
sentence phrase
clause

MAL & syntax


sentence
clause
???
word
sentence
clause
clause phrase
word
sentence
sentence phrase
clause

MAL & syntax


sentence
clause
linear dependency segment
word
sentence
clause
clause phrase
word
sentence
sentence phrase
clause

Linear dependency segment

- **the longest possible sequence of directly dependent words, which are adjoined in the linear sentence ordering**

Linear dependency segment

- **the longest possible sequence of directly dependent words, which are adjoined in the linear sentence ordering**

- if a word isn‘t neighbour of his syntactical parent, this word create LDS all by itself

Linear dependency segment

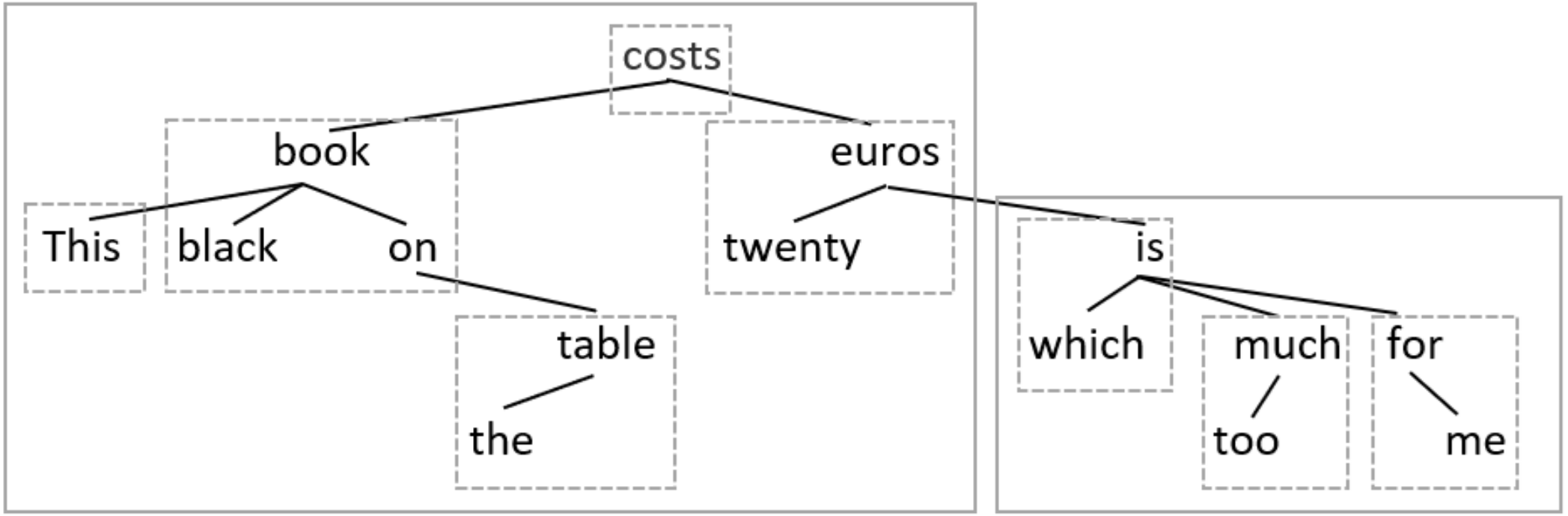
- **the longest possible sequence of directly dependent words, which are adjoined in the linear sentence ordering**

- if a word isn‘t neigbour of his syntactical parent, this word create LDS all by itself

- none of the LDS can exceed the border of the clause

Linear dependency segment

This black book on the table costs twenty euros which is too much for me



LDS analysis

- Czech (PUD treebank)
- surface universal dependency annotation scheme (SUD)

LDS analysis

- Czech (PUD treebank)
- surface universal dependecy annotation scheme (SUD)

SENTENCE

CLAUSE

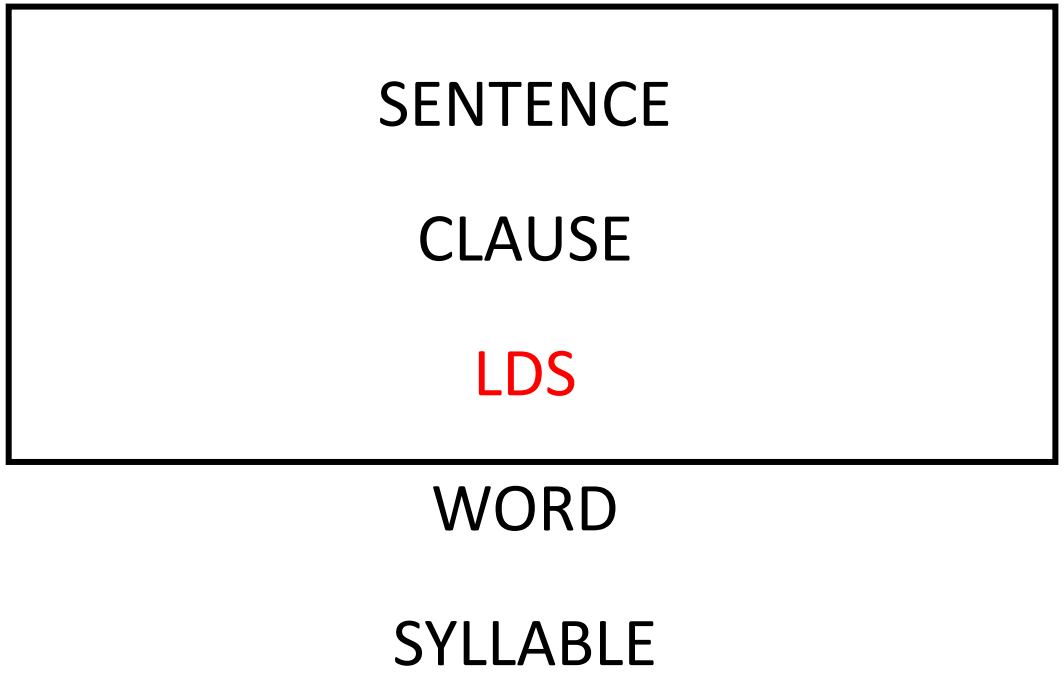
LDS

WORD

SYLLABLE

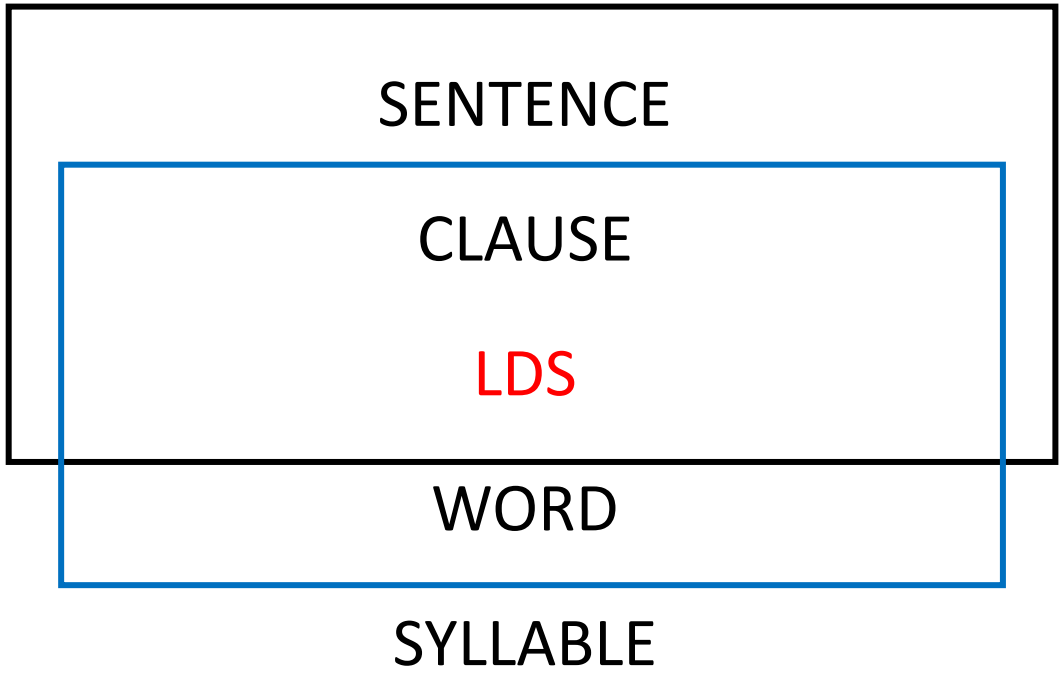
LDS analysis

- Czech (PUD treebank)
- surface universal dependecy annotation scheme (SUD)



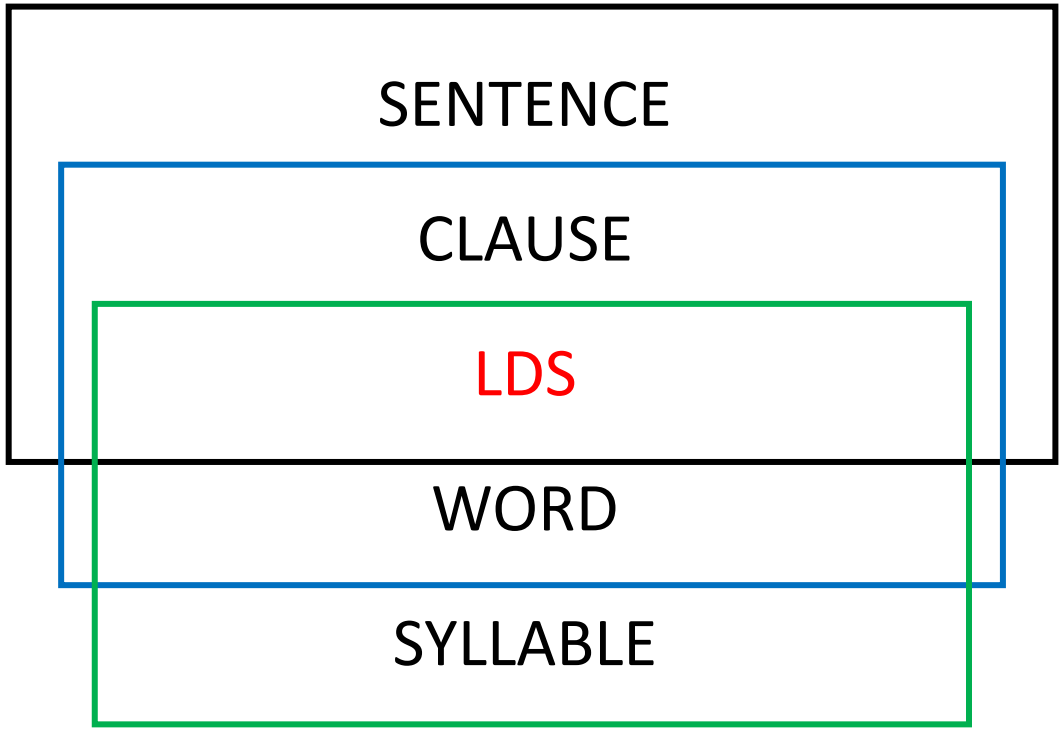
LDS analysis

- Czech (PUD treebank)
- surface universal dependecy annotation scheme (SUD)

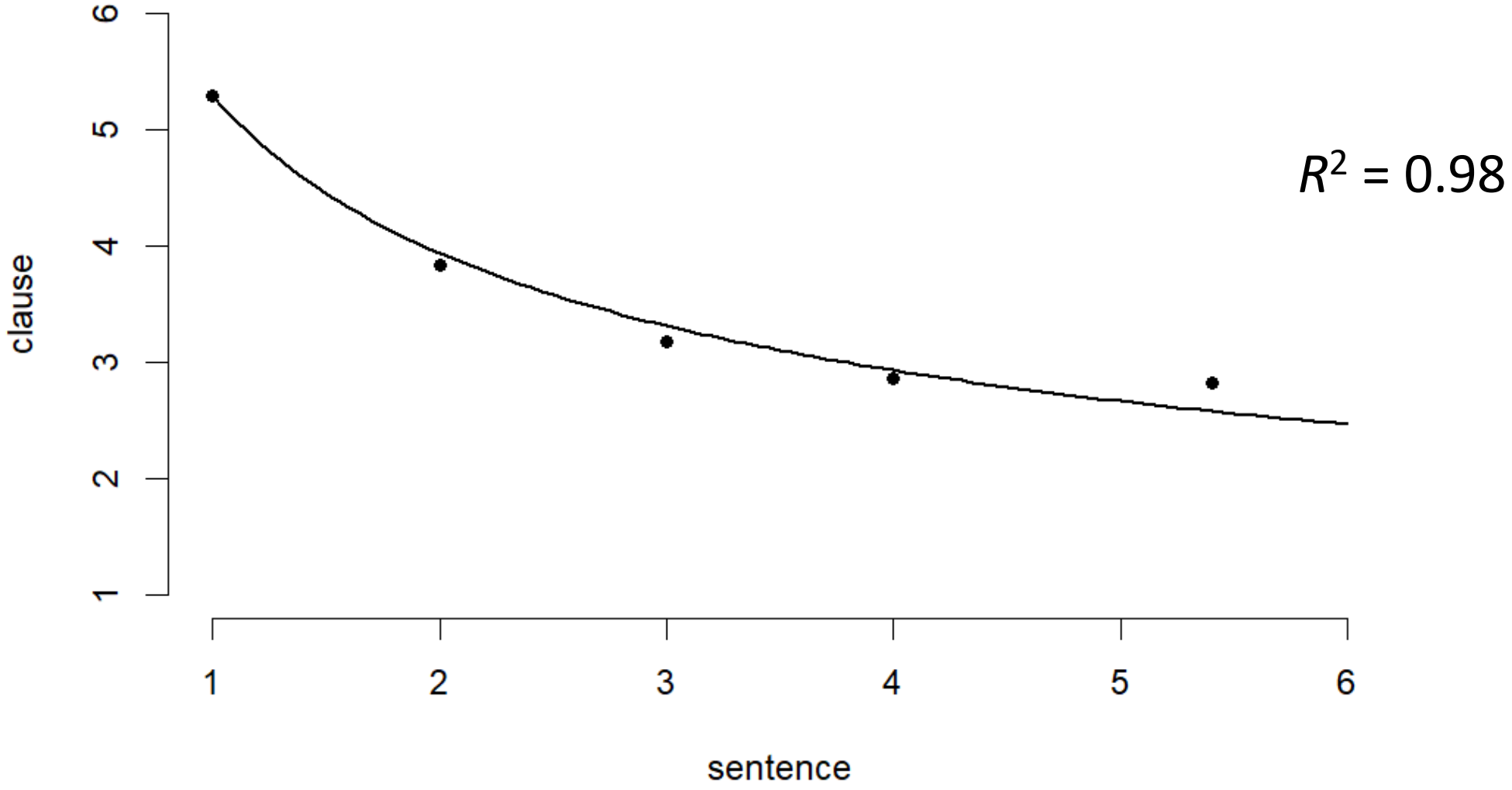


LDS analysis

- Czech (PUD treebank)
- surface universal dependecy annotation scheme (SUD)

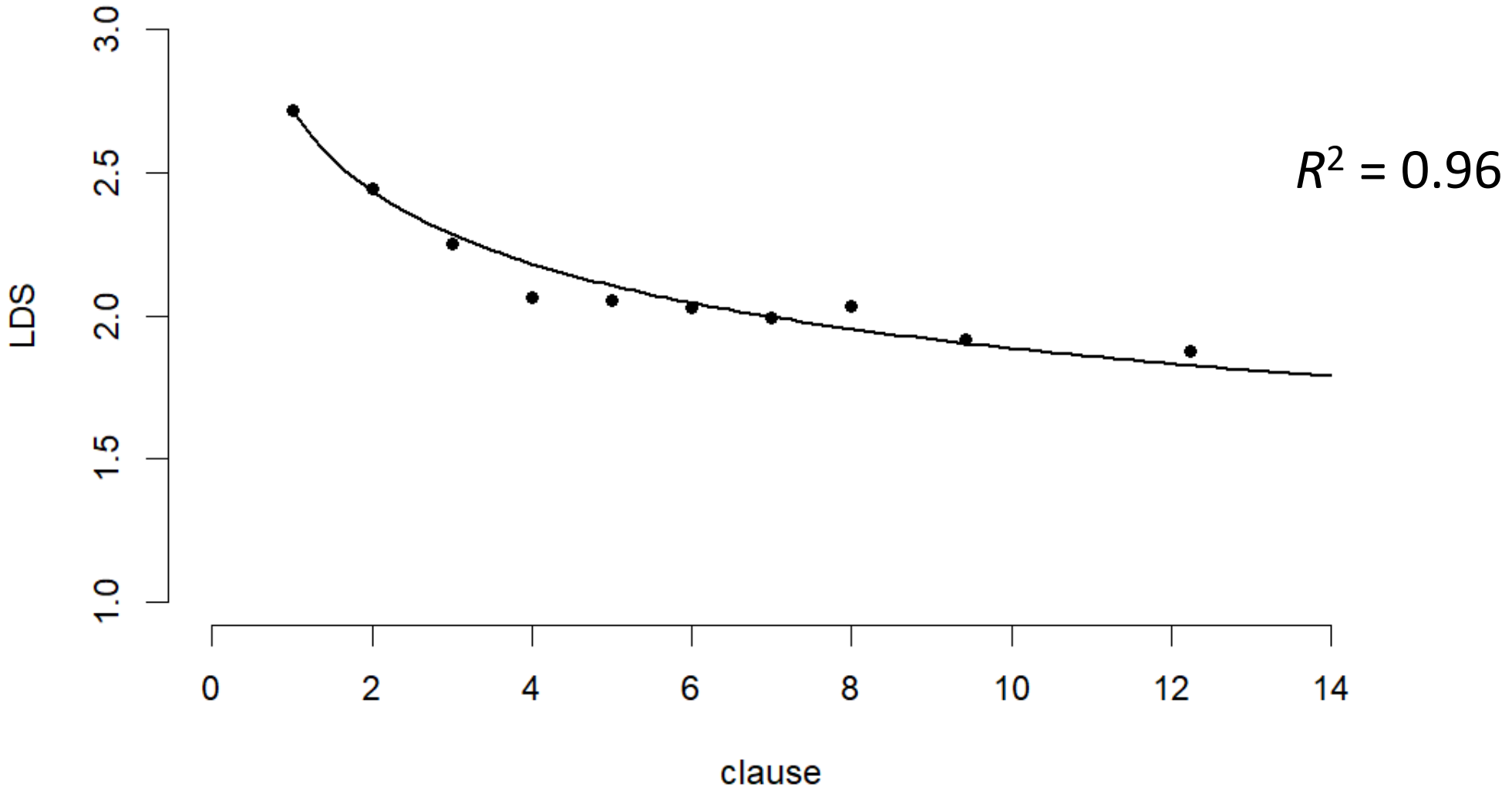


sentence – clause – LDS



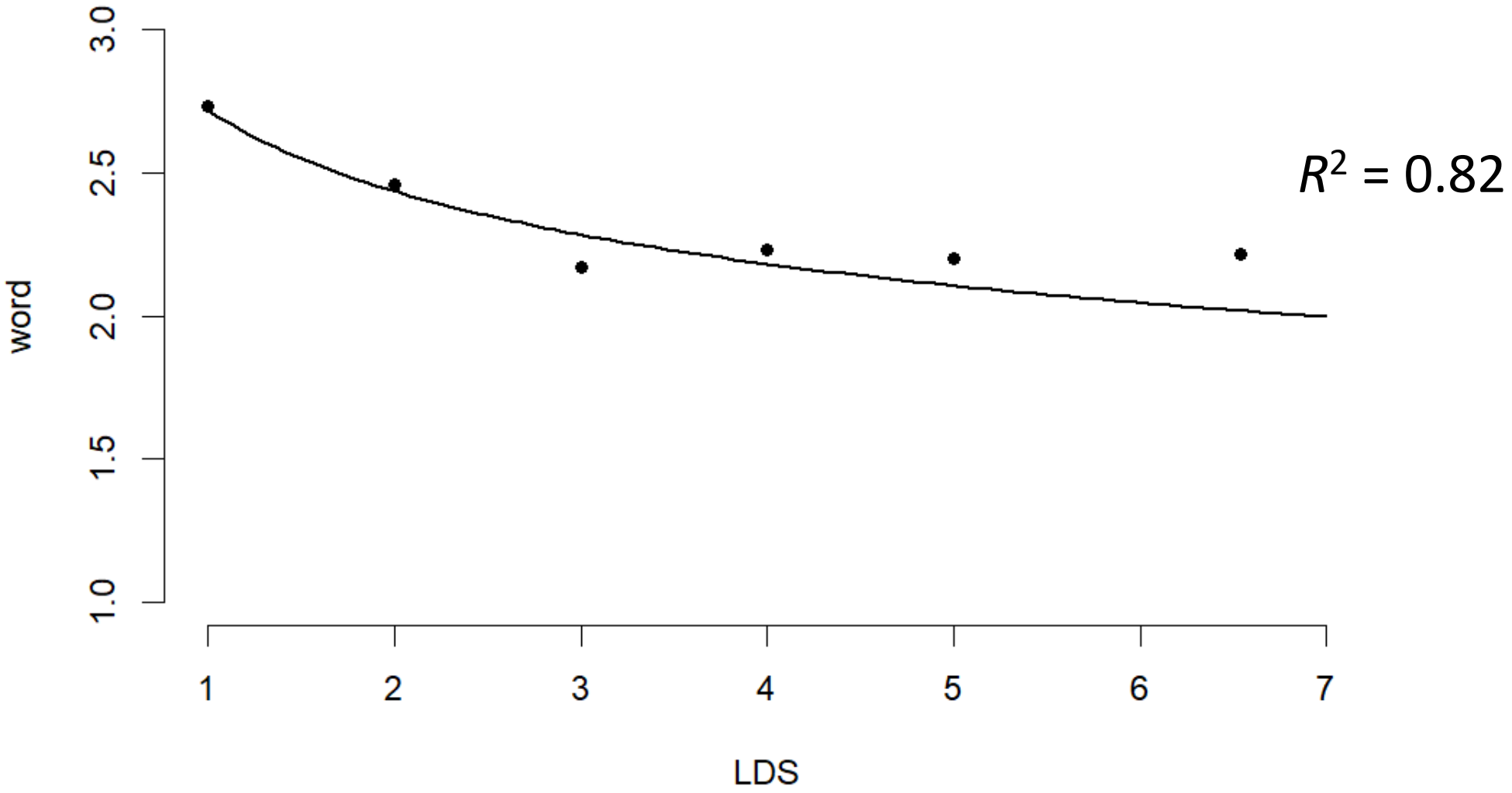
$R^2 = 0.98$

clause – LDS – word



$R^2 = 0.96$

LDS – word – syllable



$R^2 = 0.82$

Further problems

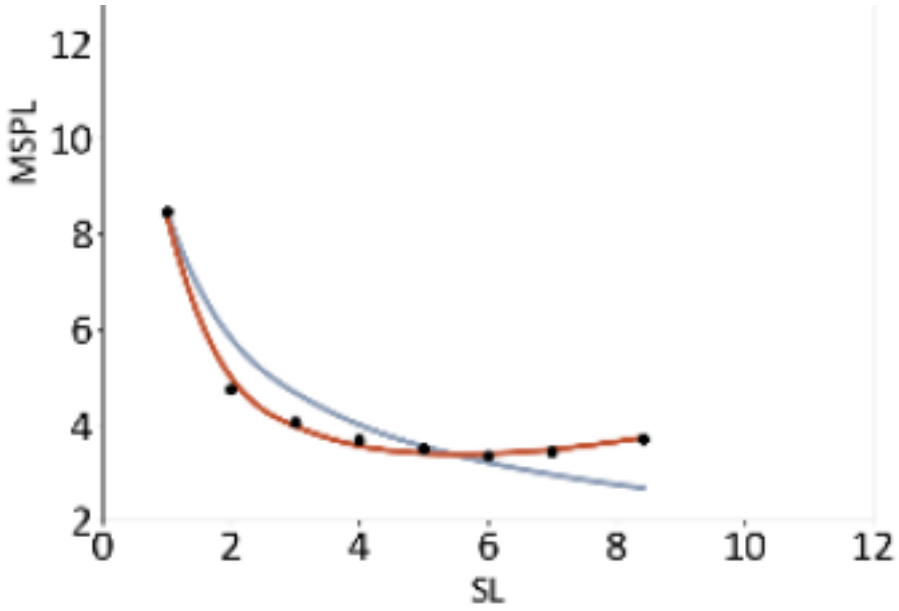
Further problems

- annotation schema
  - universal dependency vs. surface universal dependency
  - segmentation of units (e.g., morphems)
  - …

Further problems

- annotation schema
  - universal dependecy vs. surface universal dependency
  - segmentation of units (e.g., moprhmems)
  - …

- inverse regime



Conclusions / recommendations

- the MAL → properties of a language system, not a language usage

Conclusions / recommendations

- the MAL → properties of a language system, not a language usage

- analyze types (or lemmas)

Conclusions / recommendations

- the MAL → properties of a language system, not a language usage
- analyze types (or lemmas)
- more than one triplet should be analyzed

Conclusions / recommendations

- the MAL → properties of a language system, not a language usage
- analyze types (or lemmas)
- more than one triplet should be analyzed
- syntax is a challenge

Conclusions / recommendations

- the MAL → properties of a language system, not a language usage
- analyze types (or lemmas)
- more than one triplet should be analyzed
- syntax is a challenge
- are there other levels?
  - above sentence
  - under phoneme

Conclusions / recommendations

- the MAL → properties of a language system, not a language usage
- analyze types (or lemmas)
- more than one triplet should be analyzed
- syntax is a challenge
- are there other levels?
  - above sentence
  - under phoneme
- **linguistic interpretation**

Conclusions

- Motalová Tereza (2022): *Menzerath-Altmann Law in Chinese*. PhD thesis
  - https://theses.cz/id/vqk2ml/?zpet=%2Fvyhledavani%2F%3Fsearch%3DKate%C5%99ina%20PELEGRINOV%C3%81%26amp%3Bstart%3D1

- Pelegrinová Kateřina (2023): *Menzerath-Altmann Law in Czech*. PhD thesis
  - https://theses.cz/id/qsskml/?zpet=%2Fvyhledavani%2F%3Fsearch%3DKate%C5%99ina%20PELEGRINOV%C3%81%26start%3D1;isshlret=Kate%C5%99ina%3BPELEGRINOV%C3%81%3B

Conclusions

- Motalová Tereza (2022): *Menzerath-Altmann Law in Chinese*. PhD thesis
  - https://theses.cz/id/vqk2ml/?zpet=%2Fvyhledavani%2F%3Fsearch%3DKate%C5%99ina%20PELEGRINOV%C3%81%26amp%3Bstart%3D1

- Pelegrinová Kateřina (2023): *Menzerath-Altmann Law in Czech*. PhD thesis
  - https://theses.cz/id/qsskml/?zpet=%2Fvyhledavani%2F%3Fsearch%3DKate%C5%99ina%20PELEGRINOV%C3%81%26start%3D1;isshlret=Kate%C5%99ina%3BPELEGRINOV%C3%81%3B

- https://cechradek.cz/

Thank you for your attention!