Cross-linguistic variations in syntactic complexity: insights from a multilingual parallel corpus

Olga Nádvorníková Alexandr Rosen

Faculty of Arts
Charles University
Prague, Czech Republic

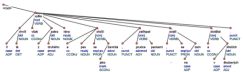
15th International Conference of the Association for Linguistic Typology
Workshop on Dependency Grammar for Typology
Nanyang Technological University
Singapore, December 4-6, 2024

Motivation

Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts.
(A. Camus, *La Peste*)

[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs,

and floundered on to its side, its body writhing in long convulsions.
(transl. S. Gilbert)



V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami,

svalil se na zem a dodělal v škubavých křečích.
(transl. M. Tomášková)

Motivation

Subordination ratio = $\frac{no.\ of\ T-units\ +\ no.\ of\ clauses}{no.\ of\ T-units}$

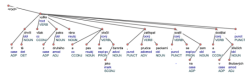
Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts.
(A. Camus, *La Peste*)

Sub.ratio = 2.5 ((2+3)/2)
Max.Tree.Depth = 3

[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs,

and floundered on to its side, its body writhing in long convulsions.
(transl. S. Gilbert)

(SPLIT) Sub.ratio = 1.33 (3+1)/3)
Max.Tree.Depth = 1



V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami,

svalil se na zem a dodělal v škubavých křečích.
(transl. M. Tomášková)

Sub.ratio = 1 (5/5)
Max.Tree.Depth = 0

Goals

- Introduce a new resource allowing for the analysis of syntactic complexity measures on large multilingual data including various genres (the InterCorp parallel corpus)
- Present results of the first pilot study conducted on the new corpus

Goals

- Introduce a new resource allowing for the analysis of syntactic complexity measures on large multilingual data including various genres (the InterCorp parallel corpus)
- Present results of the first pilot study conducted on the new corpus

Research questions

1. What are the cross-linguistic differences in syntactic complexity between a set of 12 languages? Do these languages vary also intra-linguistically, in different genres within the same language? And which factor is more important -– the language, or the genre?
2. What are the linguistic (or other) features lying behind the variation in syntactic complexity (both intra-linguistically and cross-linguistically)?
3. What are the correlations between the different syntactic complexity measures?
4. An additional technical question: Are the variations also due to divergences in the UD annotation?

Ingredients

1. Reliable set of syntactic complexity measures
2. Large multilingual data
3. with consistent syntactic annotation as a basis for calculating SCMs

Ingredients

1. Reliable set of syntactic complexity measures
2. Large multilingual data
3. with consistent syntactic annotation as a basis for calculating SCMs

$\longrightarrow$

InterCorp

- InterCorp multilingual corpus v16ud
- Annotated according to the Universal Dependencies scheme
- with 6 different SCMs for each sentence and text

Overview

1. *InterCorp* – a multilingual parallel corpus
2. Measuring syntactic complexity
   - What is syntactic complexity?
   - Syntactic complexity measures in InterCorp
3. Results
   - Languages vs. text types
   - Metrics in more detail within a single text type
   - Correlation
4. Conclusion

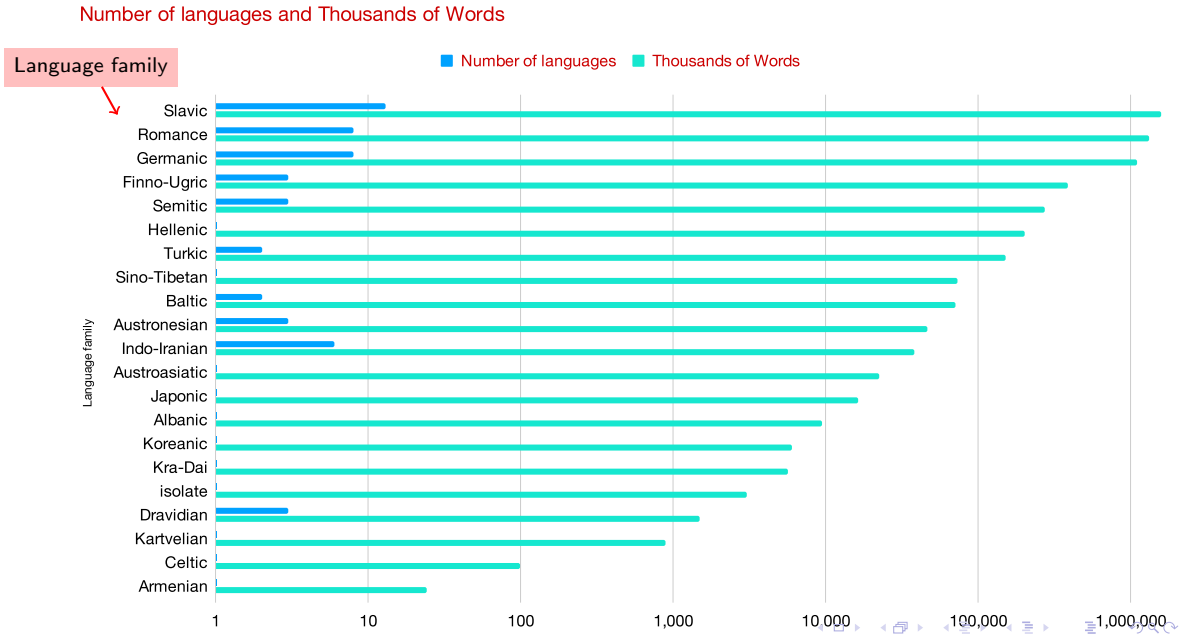
Overview

InterCorp – a multilingual parallel corpus [https://intercorp.korpus.cz/]

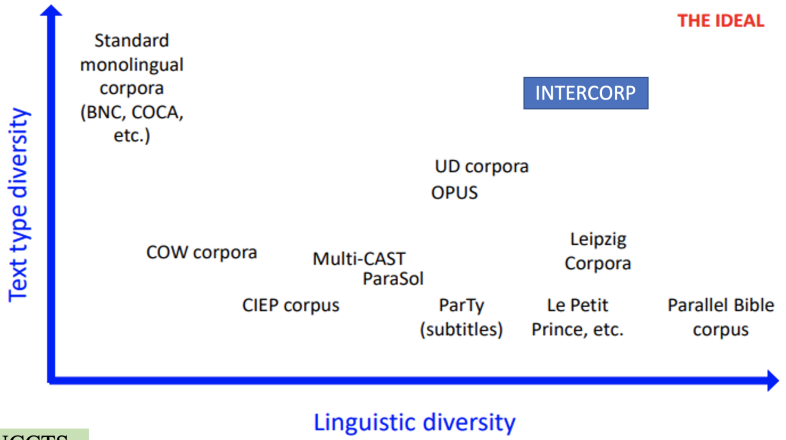
- Part of the Czech National Corpus [https://wiki.korpus.cz/]
- 2008: v0 (first online release)
- 2024 March: v16ud – fiction with UD-based linguistic annotation and complexity metrics
  [https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze16ud]
- 2024 September: v16ud – final release with all text types
- Searchable on line at [https://kontext.korpus.cz]
- Jumpstart with en-fr: [https://www.korpus.cz/kontext/query?corpname=intercorp_v16ud_en&align=intercorp_v16ud_fr]

InterCorp – a multilingual parallel corpus [https://intercorp.korpus.cz/]

- 62 languages, including 49 UD-annotated
- 5.4 billion words
- 880 million sentences
- 2.8 million texts
- Every text in Czech and at least one other language
- Also as monolingual subcorpora
- In most languages: a mix of translated and non-translated texts


Number of languages and Thousands of Words
Language family
Number of languages
Thousands of Words
Slavic
Romance
Germanic
Finno-Ugric
Semitic
Hellenic
Turkic
Sino-Tibetan
Baltic
Austronesian
Indo-Iranian
Austroasiatic
Japonic
Albanic
Koreanic
Kra-Dai
isolate
Dravidian
Kartvelian
Celtic
Armenian
Language family
1
10
100
1,000
10,000
100,000
1,000,000

InterCorp among other corpora



Natalia Levshina, UCCTS conference 2021

Universal Dependencies and syntactic complexity metrics in InterCorp

- The CONLL-U format modified
  for a concordancer with a single level of tokenization
  - pointers to syntactic context (head, function words)
    → additional attributes
  - contractions (fr: *aux*, es: *hacerlo*, en: *isn't*)
    → single tokens with multi-valued attributes
- SCMs as metadata on each sentence and text

Overview

Overview

Syntactic complexity

Complexity of a system in general:

the number and variety of elements and the elaborateness of their interrelational structure (Rescher 1998:1, Hübler 2007:10; cited by (Álvarez González et al., 2019))

Syntactic complexity

Complexity of a system in general:

the number and variety of elements and the elaborateness of their interrelational structure (Rescher 1998:1, Hübler 2007:10; cited by (Álvarez González et al., 2019))

Syntactic complexity:

Syntactic complexity in language is related to the number, type, and depth of embedding in a text. Syntactically simple authors use short, single clause sentences and rely more heavily on coordinated structures [...]. Syntactically complex authors [...] use longer sentences and more subordinate clauses that reveal more complex structural relationships. (Beaman, 1984; De Clercq, 2016)

Syntactic complexity

Complexity of a system in general:

the number and variety of elements and the elaborateness of their interrelational structure (Rescher 1998:1, Hübler 2007:10; cited by (Álvarez González et al., 2019))

Syntactic complexity:

Syntactic complexity in language is related to the number, type, and depth of embedding in a text. Syntactically simple authors use short, single clause sentences and rely more heavily on coordinated structures [...]. Syntactically complex authors [...] use longer sentences and more subordinate clauses that reveal more complex structural relationships. (Beaman, 1984; De Clercq, 2016)

→ Syntactic complexity of a sentence can be determined by:

- the number and type of syntactic elements
- their hierarchy within the sentence

Simplifying complexity

- Figuring out syntactic complexity always means simplification :)
- Complexity is multi-dimensional (registers!)
  → more metrics should be combined (Biber et al., 2024)
- Metrics are specific to genre and language (Biber et al., 2024)

Simplifying complexity

- Figuring out syntactic complexity always means simplification :)
- Complexity is multi-dimensional (registers!)
  → more metrics should be combined (Biber et al., 2024)
- Metrics are specific to genre and language (Biber et al., 2024)

Two types of (syntactic) complexity

1. relative (subjective, cognitive) complexity
   – reader-oriented, measuring processing load, readability (difficulty)
2. absolute (structural) complexity
   – measurable linguistic features in two approaches:
   1. grammatical (system) complexity – linguistic system
   2. usage-based complexity (text complexity) – actual language use

(Brunato et al., 2022; Szmrecsanyi & Kortmann, 2012; Miestamo, 2009; Biber et al., 2023; De Clercq, 2016)

What can be done with syntactic complexity

- **Language development** (Givón 2009:4)
- **Monolingual studies** (Mačutek, Čech & Milička 2019; Hudelot 198; (Biber et al., 2023))
- **Translation studies**: Izquierdo & Marco (2000), Canavese & Mori (2021); comparable or parallel corpora (translation universals – simplification, normalisation, etc.)
- **Contrastive studies**: clause-linking (Lehmann, 1988), clause-combining (Cosme 2006, etc.), information packaging (Solfjeld 1996, Fabricius-Hansen 1999), UD shared task (Berdicevskis et al., 2018), etc.
- **Register variation**: spoken/written (Beaman, 1984) etc., academic: (Biber & Gray, 2016), etc.
- **Typology**: Levshina (2019), (Levshina, 2021) – Leipzig Corpora Collection (comparable, UD)
- **Readability**: Kincaid et al. 1975, Dell'Orletta et al. 2011, Gruszczyński & Ogrodniczuk 2015 (*Jasnopis*).
- **Language acquisition, proficiency assessment** (L1 et L2), (Lu, 2010; Jagaiah et al., 2020) etc.

Many different metrics of syntactic complexity

- 47 different metrics proposed
  during the last 30 years in 130 studies
  (Jagaiah et al., 2020)
- Linguistic Profiling – a web-based tool:
  100 items, most of them related to syntax
  (Brunato et al., 2020)
- 8 morphological and 7 syntactic complexity metrics
  used during the first shared task on measuring linguistic complexity
  (Berdicevskis et al., 2018; Xu & Li, 2021; Biber et al., 2023)

Overview

Our choice of SCMs

... is a compromise to satisfy various users, driven mainly by the goal to annotate the corpus

"Manning's Law for SCM"

SCMs in a multilingual corpus should be:

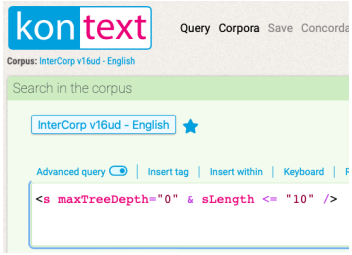
- Implementable via uniform annotation (UD)
- Reliable and comparable across languages and text types
- Well-known and established (to reproduce and compare with existing results)

Dimension	**Noun phrase**	**Sentence**
Horizontal	*maxNPLength*	*sLength*
Vertical	*maxNPDepth*	*subRatio* *maxTreeDepth*
Combined		*mdd*

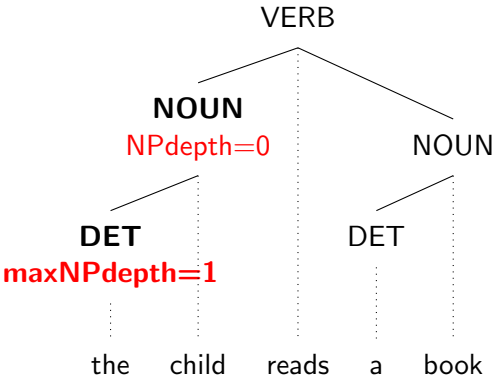
Implemented as metadata

- for each sentence
- for each text (as weighted averages)

SCMs in the corpus search interface



Nominal depth and width



MaxNPDepth

Maximum no. of embeddings in any NP

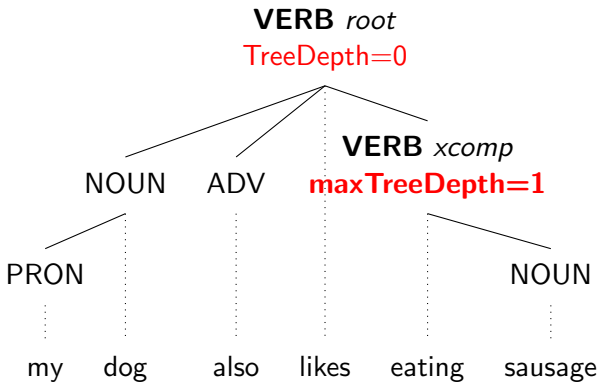
MaxNPLength

No. of words in the longest NP

What counts as an NP

- Subtree with NOUN, PNOM, PRON as the head
- Every conjunct separately
- Ignore punctuation
- Nominal predicate? Ignore subject, copula, adverbials

Clausal depth and width



MaxTreeDepth

Max. no. of clausal embeddings, skipping coordination

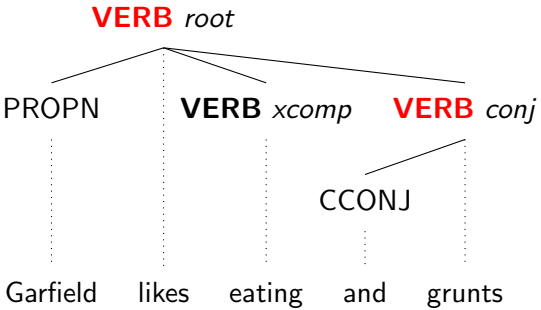
sLength

No. of words, ignoring punctuation

What counts as a clause

- *csubj* – subject clause
- *ccomp* – complement clause
- *xcomp* – open predicate (predicative complement)
- *advcl* – adverbial clause
- *acl* – attribute clause

Clausal depth: subordination ratio



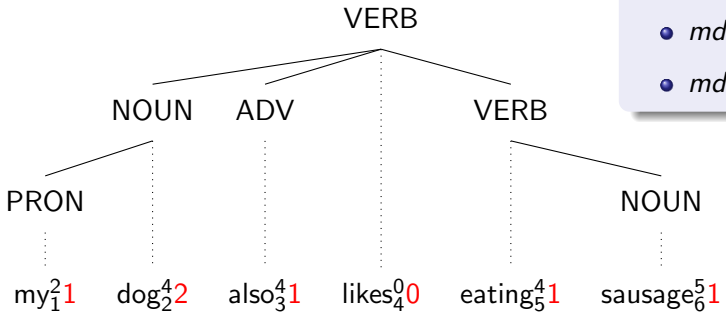
subRatio

- $\frac{no.\ of\ T-units\ +\ no.\ of\ clauses}{no.\ of\ T-units}$
- T-unit: main clause (conjunct) including all dependents

(Hunt, 1965)

- T-units = 2
- clauses = 3
- $subRatio = \frac{2+3}{2} = 2.5$

Mean Dependency Distance



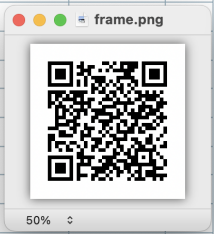
MDD

- Average head-daughter distance
- Punctuation is ignored
- $mdd = \frac{\Sigma_{i=0\ldots n} \, |ID_i - head_i|}{(n-1)}$
- $mdd = \frac{6}{5} = \mathbf{1.2}$

(Yan & Li, 2019; Mačutek et al., 2021; Alemany-Puig & Ferrer-i Cancho, 2024; Futrell et al., 2020)

Metrics for texts in InterCorp [https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze16ud#detailed_statistics]

Lang	Collection	Number of		Thousands of			Lexical diversity		Syntactic complexity (average)					
		docs	texts	sentences	words	tokens	lexDivWord	lexDivLemma	sLength	subRatio	maxTreeDepth	maxNPLength	maxNPDepth	mdd
af	Subtitles	1	24	23.0	134.6	161.7	406.4	347.2	5.887	1.093	0.095	2.377	0.811	2.251
ar	Core-fiction	2	2	2.1	28.8	35.6	620.3	576.6	13.830	2.712	1.310	5.293	2.016	2.817
	Core-misc	1	1	1.3	5.5	7.4	451.4	421.4	4.150	1.330	0.290	1.870	0.840	2.010
	Subtitles	1	34 193	28 726.4	126 195.5	157 188.9	592.8	557.3	4.421	1.338	0.336	2.216	0.986	1.678
	Syndicate	3	433	19.0	384.5	439.0	622.7	560.3	20.513	2.485	1.312	11.036	3.940	2.405
be	Core-fiction	104	104	625.1	7 068.7	8 978.9	615.4	492.7	11.583	1.865	0.804	4.122	1.436	2.316
	Core-misc	4	4	7.6	57.7	76.0	556.2	425.6	7.608	1.672	0.605			
bg	Core-fiction	87	87	559.6	7 067.3	8 597.7	548.3	439.5	13.125	1.728	0.732			
	Acquis	1	10 846	862.3	13 582.3	16 991.2	392.4	306.3	18.073	1.801	0.514			
	Europarl	1	45 271	408.3	9 082.0	10 379.8	498.4	386.3	23.014	2.538	1.263			
	Subtitles	1	40 986	32 591.1	164 644.1	214 988.4	518.2	384.6	5.089	1.336	0.322			
bn	Subtitles	1	252	363.8	1 517.7	2 072.1	419.4	–	–	–	–			
br	Subtitles	1	27	19.7	97.4	145.2	363.5	–	–	–	–			
bs	Subtitles	1	14 208	12 165.3	56 465.9	75 945.3	450.2	–	–	–	–			
ca	Core-fiction	91	91	678.0	9 951.3	11 363.4	471.6	375.2	15.579	2.140	0.962	6.099	1.920	2.551
	Core-misc	1	1	0.7	9.7	11.2	463.7	362.5	14.300	2.040	0.930	5.850	1.880	2.520
	Bible	2	66	50.3	728.2	839.4	405.3	308.0	15.729	2.056	0.912	6.460	2.103	2.602



Overview

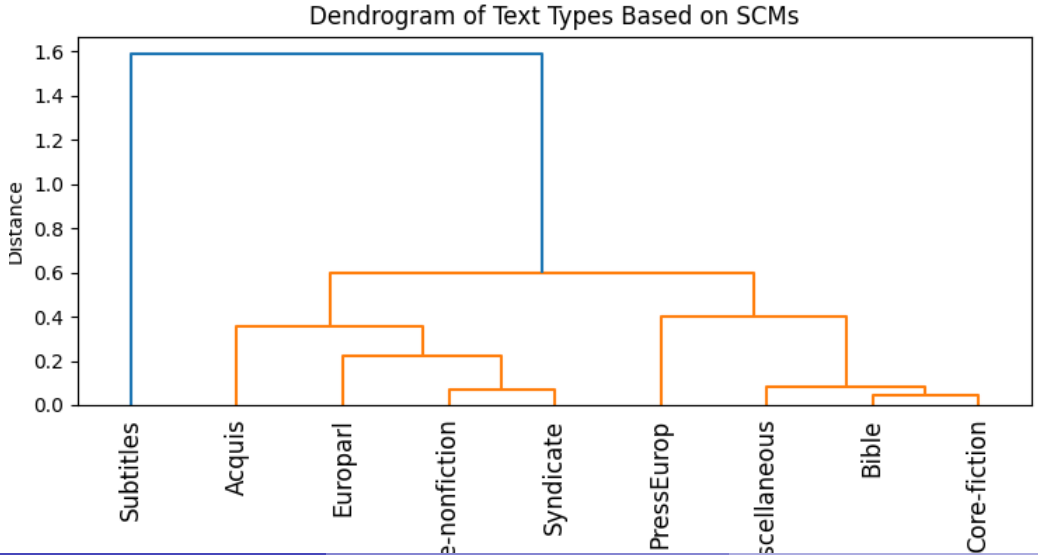
Text types, samples

Text type	Number of texts	Thousands of words
Fiction	5 879	473 208
Misc	226	7 853
Nonfiction	350	29 450
Acquis	380 049	424 874
Bible	1 252	12 050
Europarl	1 369 378	276 543
PressEurop	69 894	26 964
Subtitles	965 557	3 970 273
Syndicate	39 158	35 385
TOTAL	2 831 743	5 256 601

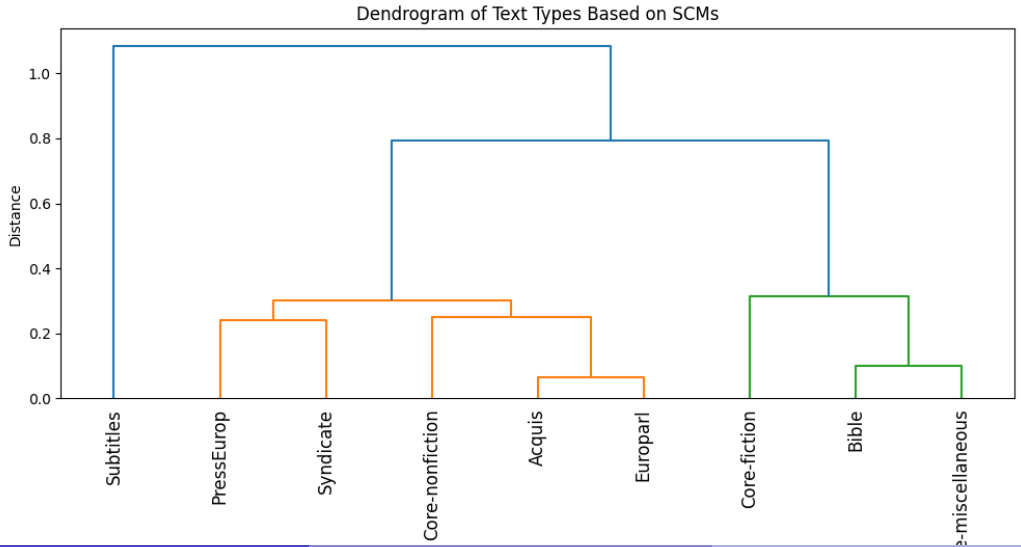
Sample ID	12x6	12xF	12xAll	49xAll
Languages	12	12	12	49
Text types	fiction	fiction	all	all
Texts (cs)	6	1 629	164 th.	164 th.
Texts (all)	72	4 133	1 197 th.	2 796 th.
M words (cs)	0.4	114	397	397
M words (all)	6.8	333	2 427	5 132

Overview

Czech only, sLength



English only, maxNPLength



Larger differences across text types or languages?

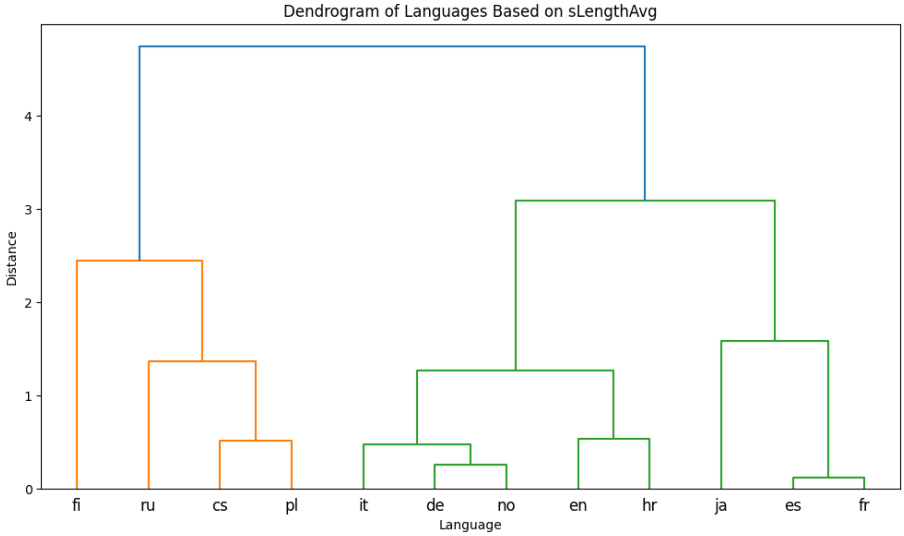
	Text types (8)		Languages (12)	
SCM	**Effect size η^2**	**Evaluation**	**Effect size η^2**	**Evaluation**
subRatio	0.23	High	0.12	Medium
maxTreeDepth	0.08	Low	0.15	High
sLength	0.18	High	0.10	Medium
mdd	0.16	High	0.08	Low
maxNPLength	0.14	High	0.09	Low
maxNPDepth	0.08	Low	0.13	Medium
Average	**0.14**	**High**	**0.11**	**Medium**

Overview

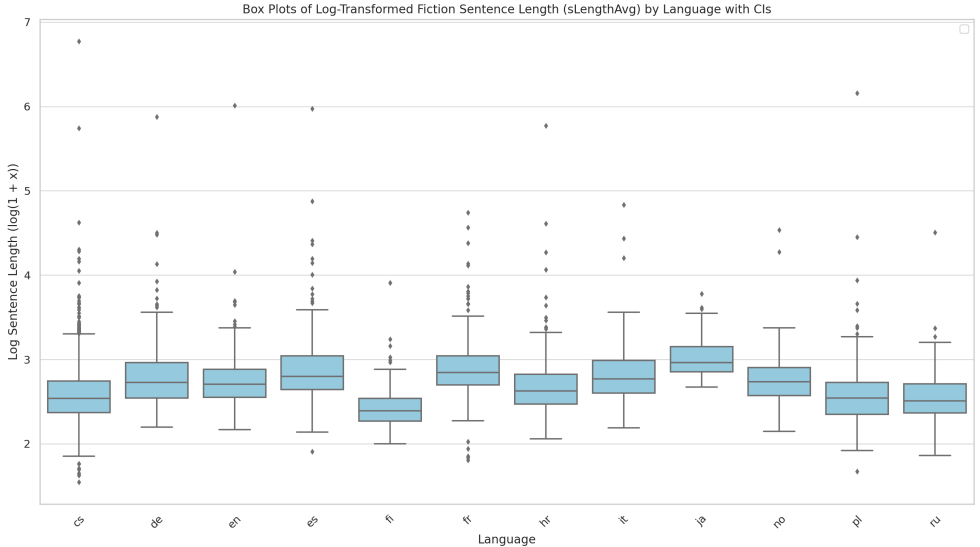
Data in this subsection

- Statistics about texts in 12 languages
- Text type: fiction
- Extent: all available texts (about 5900)
- Metrics: sLength and subRatio

sLength


Dendrogram of Languages Based on sLengthAvg
Distance
0
1
2
3
4
fi
ru
cs
pl
it
de
no
en
hr
ja
es
fr
Language

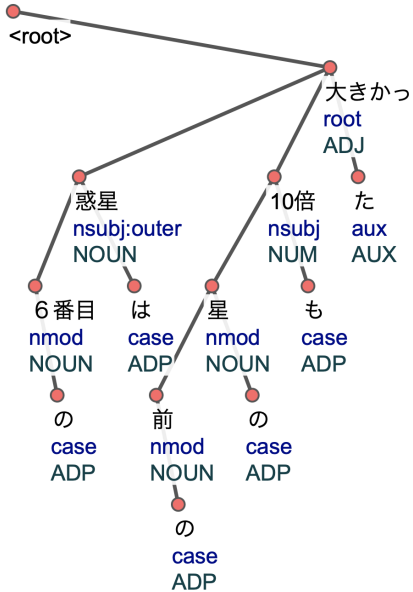
sLength


Box Plots of Log-Transformed Fiction Sentence Length (sLengthAvg) by Language with CIs
Log Sentence Length (log(1 + x))
7
6
5
4
3
2
cs
de
en
es
fi
fr
hr
it
ja
no
pl
ru
Language

sLength values for a sample sentence in a few languages

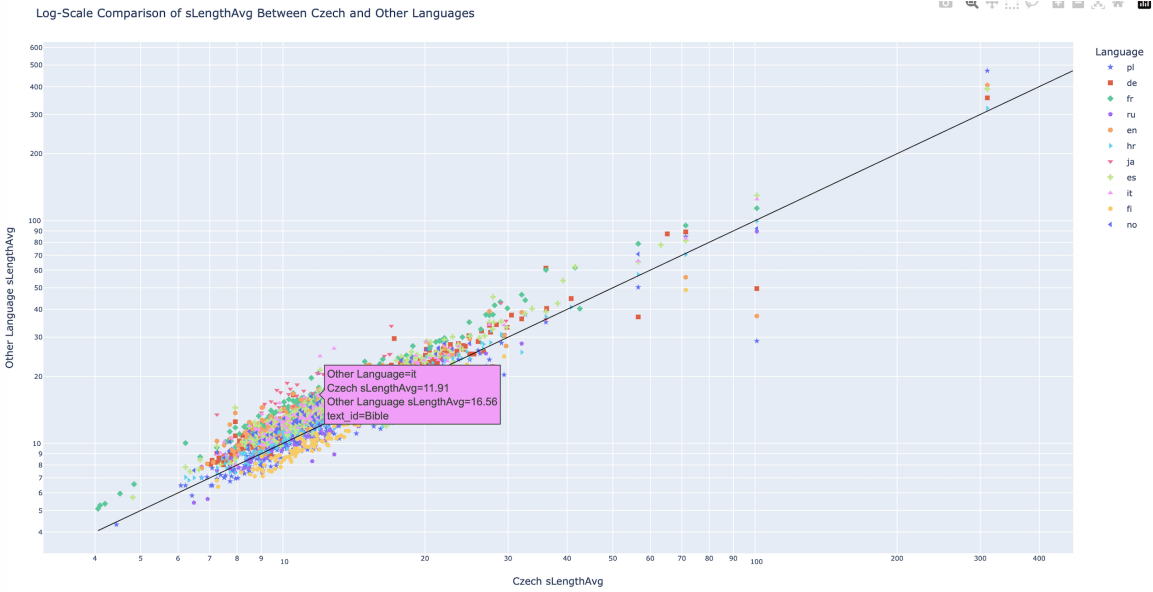
(1)
a. *Šestá planeta byla desetkrát větší.* (cs, 5)
b. *Kuudes kiertotähti oli kymmenen kertaa suurempi.* (fi, 6)
c. *La sixième planète était une planète dix fois plus vaste.* (FR, 10)
d. *The sixth planet was ten times larger than the last one.* (en, 11)
e. *roku ban me no wakusei wa mae no hoshi no 10 bai mo ōki kat ta* (ja, 16)

6番目 の 惑星 は 前 の 星 の 10倍 も 大きかっ た



sLength, Fiction + Bible

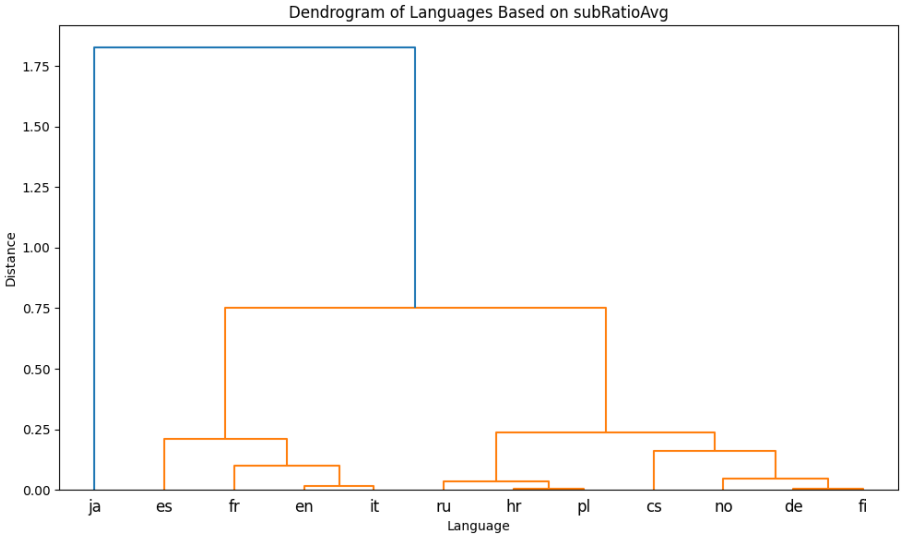
[https://jakobson.korpus.cz/~rosen/public/COMPLEXITY/sLength_scatter.html]


Log-Scale Comparison of sLengthAvg Between Czech and Other Languages
Other Language sLengthAvg
Czech sLengthAvg
Language
pl
de
fr
ru
en
hr
ja
es
it
fi
no
Other Language=it
Czech sLengthAvg=11.91
Other Language sLengthAvg=16.56
text_id=Bible

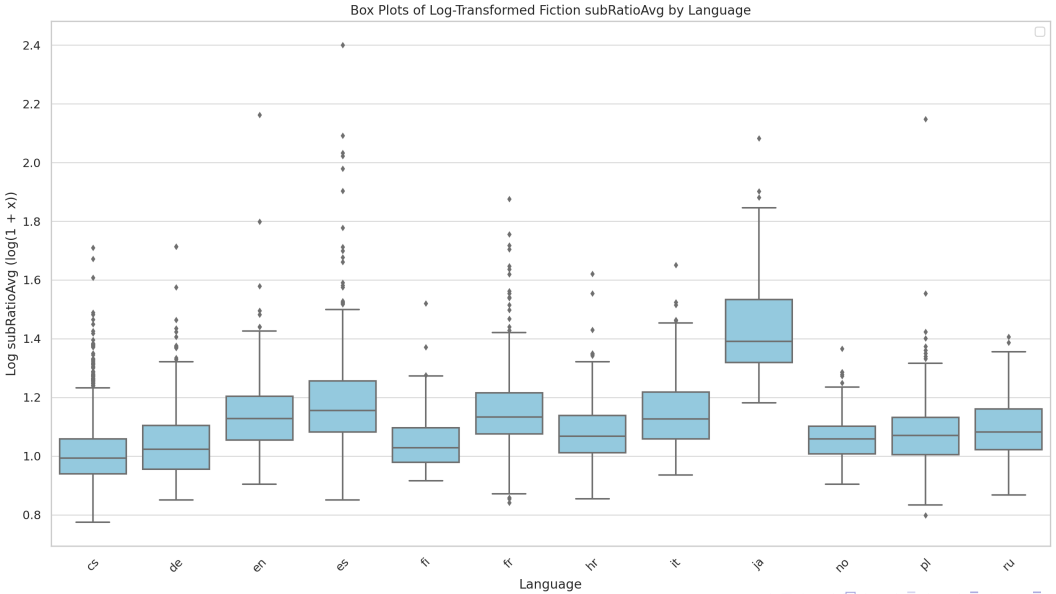
SCMs for the Bible in various languages

lang	sLength	subRatio	maxTreeDepth	maxNPLength	maxNPDepth	mdd
cs	11.907	1.603	0.635	4.125	1.590	2.451
de	15.637	1.648	0.657	5.263	1.737	2.998
en	17.458	2.166	1.051	6.271	2.125	2.608
fi	13.324	1.911	0.871	4.231	1.534	2.511
fr	17.822	2.060	0.893	6.743	2.171	2.758
hr	12.989	1.855	0.773	4.359	1.599	2.504
it	16.561	1.969	0.881	6.739	2.168	2.723
no	13.099	1.573	0.620	4.645	1.713	2.447
pl	12.695	1.724	0.727	4.479	1.725	2.397
ru	20.730	2.746	1.302	6.198	2.121	2.828

subRatio


Dendrogram of Languages Based on subRatioAvg
Distance
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
ja
es
fr
en
it
ru
hr
pl
cs
no
de
fi
Language

subRatio


Box Plots of Log-Transformed Fiction subRatioAvg by Language
Log subRatioAvg (log(1 + x))
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
cs
de
en
es
fi
fr
hr
it
ja
no
pl
ru
Language

Motivation

Subordination ratio = $\frac{no.\ of\ T-units\ +\ no.\ of\ clauses}{no.\ of\ T-units}$

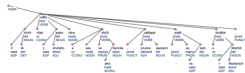
Au même moment, un coup de revolver partit du second et le chien se retourna comme une crêpe, agitant violemment ses pattes pour se renverser enfin sur le flanc, secoué par de longs soubresauts.
(A. Camus, *La Peste*)

Sub.ratio = 2.5 ((2+3)/2)
Max.Tree.Depth = 3

[…] when a revolver barked from the third-floor window. // The dog did a somersault like a tossed pancake, lashed the air with its legs,

and floundered on to its side, its body writhing in long convulsions.
(transl. S. Gilbert)

(SPLIT) Sub.ratio = 1.33 (3+1)/3)
Max.Tree.Depth = 1



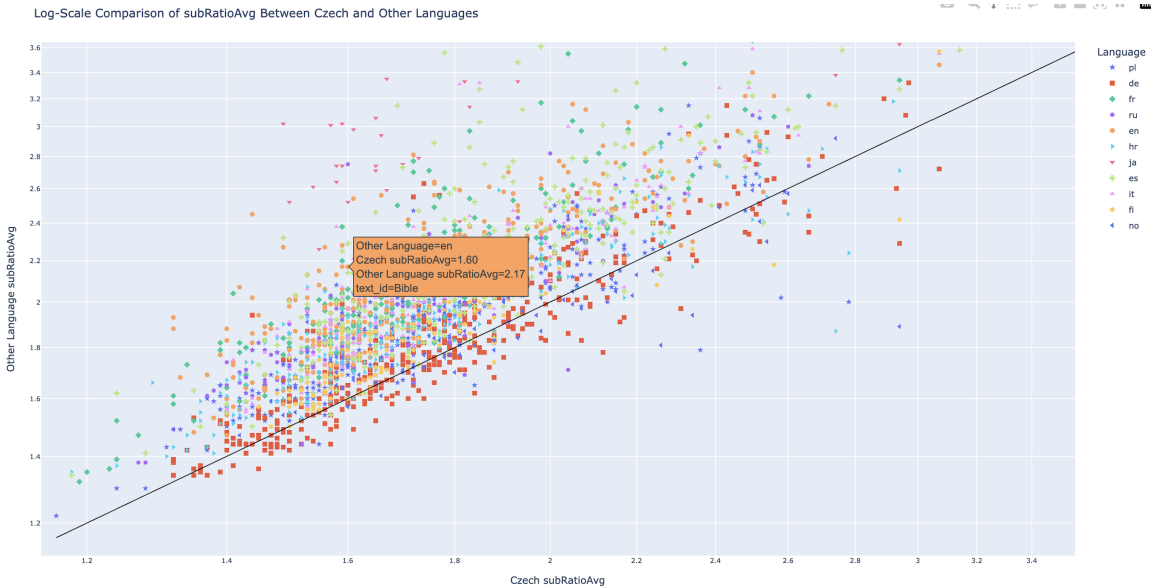
V té chvíli však vyšla z druhého patra rána a pes se otočil jako čamrda, prudce zatřepal packami,

svalil se na zem a dodělal v škubavých křečích.
(transl. M. Tomášková)

Sub.ratio = 1 (5/5)
Max.Tree.Depth = 0

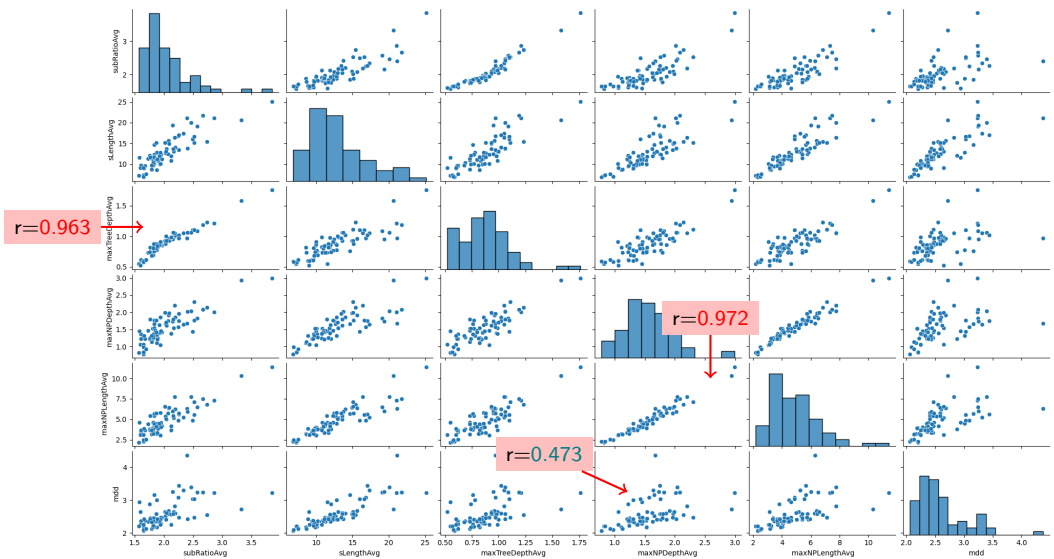
subRatio, Fiction + Bible

[https://jakobson.korpus.cz/~rosen/public/COMPLEXITY/subRatio_scatter.html]


Log-Scale Comparison of subRatioAvg Between Czech and Other Languages
Other Language=en
Czech subRatioAvg=1.60
Other Language subRatioAvg=2.17
text_id=Bible
Other Language subRatioAvg
Czech subRatioAvg
Language
pl
de
fr
ru
en
hr
ja
es
it
fi
no
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
3.2
3.4

Overview

SCM correlation on 6 texts in 12 languages (for texts, not sentences)



SCM correlation matrix on 6x12

SCM	subRatio	sLength	maxTreeDepth	maxNPDepth	maxNPLength	mdd	Correlation
subRatio	1	0.860	0.963	0.828	0.864	0.562	Pearson
sLength	0.860	1	0.841	0.823	0.896	0.834	Pearson
maxTreeDepth	0.963	0.841	1	0.845	0.857	0.500	Pearson
maxNPDepth	0.828	0.823	0.845	1	0.972	0.473	Pearson
maxNPLength	0.864	0.896	0.857	0.972	1	0.599	Pearson
mdd	0.562	0.834	0.500	0.473	0.599	1	Pearson
subRatio	1	0.851	0.974	0.766	0.794	0.649	Spearman
sLength	0.851	1	0.847	0.851	0.911	0.882	Spearman
maxTreeDepth	0.974	0.847	1	0.758	0.768	0.630	Spearman
maxNPDepth	0.766	0.851	0.758	1	0.970	0.638	Spearman
maxNPLength	0.794	0.911	0.768	0.970	1	0.759	Spearman
mdd	0.649	0.882	0.630	0.638	0.759	1	Spearman
subRatio	1	0.673	0.883	0.583	0.612	0.476	Kendall
sLength	0.673	1	0.669	0.669	0.753	0.720	Kendall
maxTreeDepth	0.883	0.669	1	0.577	0.585	0.468	Kendall
maxNPDepth	0.583	0.669	0.577	1	0.861	0.458	Kendall
maxNPLength	0.612	0.753	0.585	0.861	1	0.571	Kendall
mdd	0.476	0.720	0.468	0.458	0.571	1	Kendall

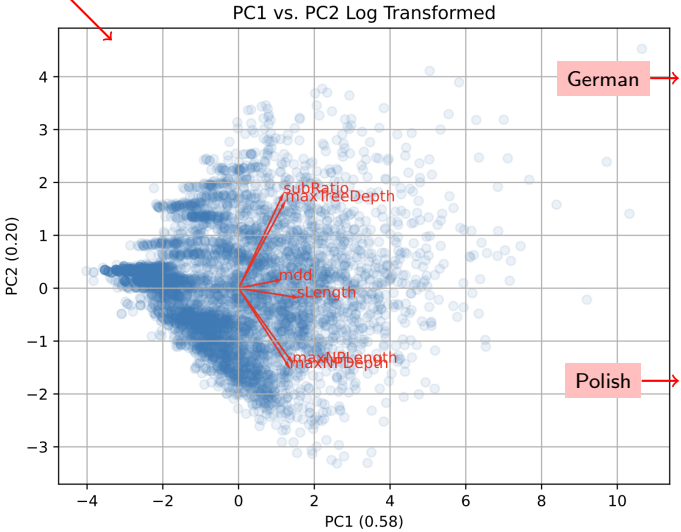
Correlations and PCA

The sample:

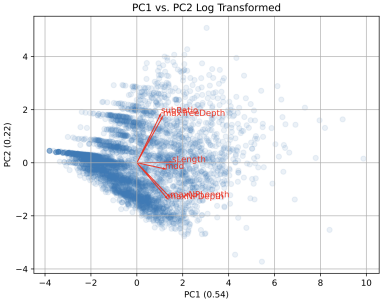
- 12 languages
- 6 texts (super-parallel)
- Fiction
- 10,000 random sentences
- Log transformed

Pattern 1: clausal vs. NP measures

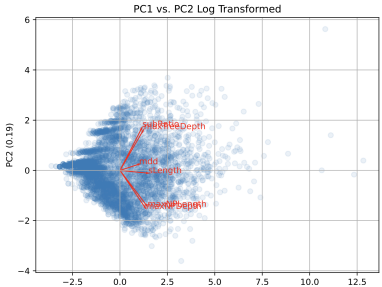
Spanish



German

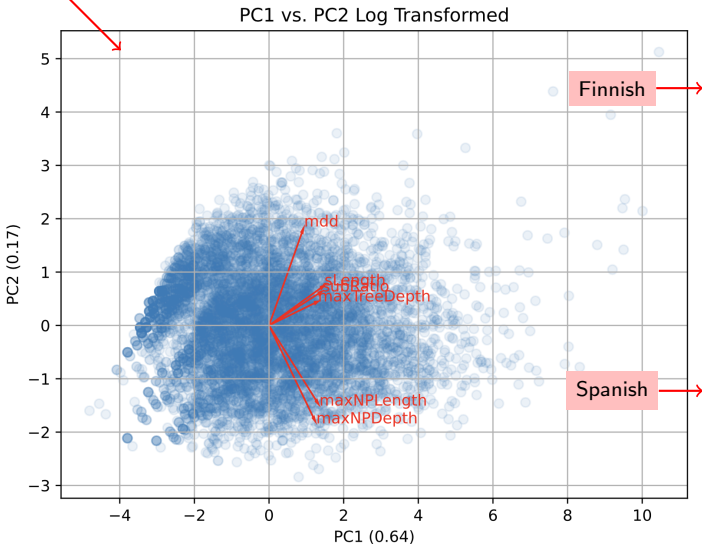


Polish

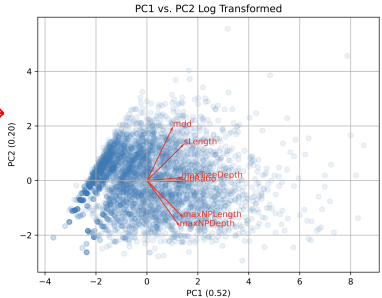


Pattern 2: MDD vs. NP measures

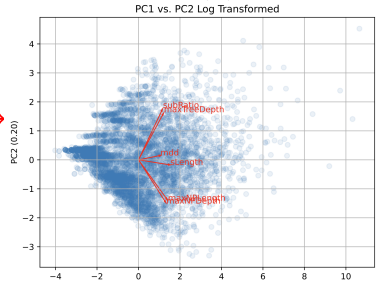
Japanese

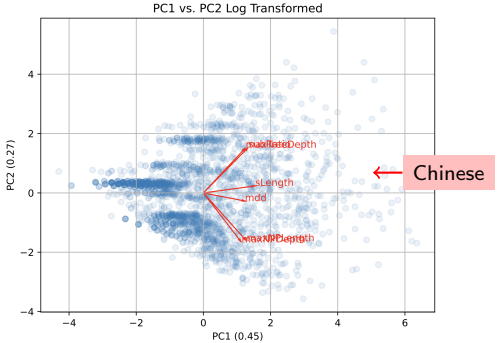


Finnish

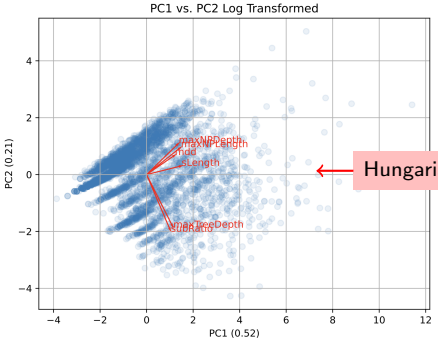


Spanish

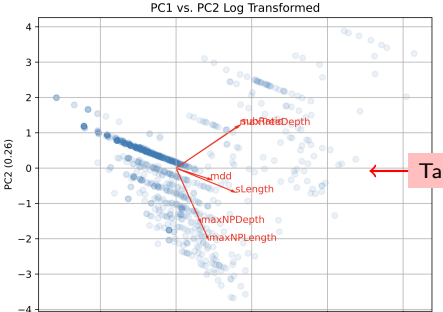




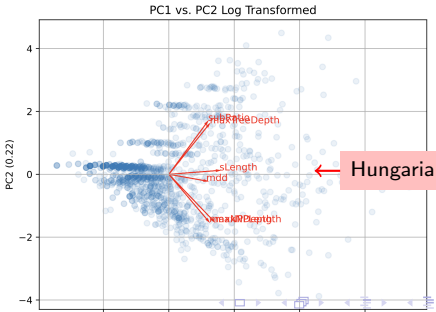
Chinese



Hungarian fiction



Tamil



Hungarian Subtitles

Overview

Goal 1: New multingual resource

The new version of the InterCorp multilingual corpus (16ud) allows for the analysis of 6 syntactic complexity measures both cross-linguistically and intra-linguistically (across text types), on 49 languages annotated by Universal Dependencies.

Goal 2: Pilot study of the syntactic complexity in InterCorp (12 languages)

- RQ1 – Both **language and text type** influence the syntactic complexity of texts and sentences, but text type shows a larger effect size than language (in our sample). This highlights the importance of distinguishing text types in cross-linguistic analyses.
- RQ2 – Within **fiction**, languages cluster differently according to different SCMs (sLength and subRatio). The differences are given by **structural variations** between languages, but also – in Japanese – partly by the **specifics of the UD** annotation (tokenization). In fiction, languages show high degree of variation, due to the **stylistic diversity** of the analyzed texts.
- RQ3 – In our sample, a strong correlation was observed between **the two NP measures** (maxNPDepth and maxNPLength) and between **the two clausal measures** (maxTreeDepth and subRatio). Most of the languages of the sample show a similar **PCA pattern**, but in other languages, patterns vary both cross-linguistically and intra-linguistically (text type variation).

Grazie mille della vostra attenzione.
Labai dėkoju už dėmesį.
Liels paldies par uzmanību.
Dank u zeer voor uw aandacht.
Dziękuję bardzo Państwu za uwagę.
Muito obrigado pela vossa atenção.
感谢观看
Veľmi pekne vám ďakujem za pozornosť.
Najlepša hvala za vašo pozornost.
Tack så mycket för er uppmärksamhet.
Mange tak for Deres opmærksomhed.
Vielen Dank für Ihre Aufmerksamkeit.
Thank you very much for your attention.
Muchísimas gracias por su atención.
Suur tänu tähelepanu eest.
ご清聴ありがとうございました。
Oikein paljon kiitoksia mielenkiinnostanne.
Je vous remercie de votre attention.
Nagyon szépen köszönöm a figyelmüket.
Velice vám děkuji za pozornost.

Bibliography I

Alemany-Puig, L. & Ferrer-i Cancho, R. (2024). The expected sum of edge lengths in planar linearizations of trees. *Journal of Language Modelling*, **12**(1), 1–42.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen and R. O. Freedle, editors, *Coherence in Spoken and Written Discourse*, page 45–80. ABLEX Publishing Corporation, Norwood, New Jersey.

Berdicevskis, A., Çöltekin, Ç., Ehret, K., von Prince, K., Ross, D., Thompson, B., Yan, C., Demberg, V., Lupyan, G., Rama, T., & Bentz, C. (2018). Using Universal Dependencies in cross-linguistic complexity research. In M.-C. de Marneffe, T. Lynn, and S. Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17, Brussels, Belgium. Association for Computational Linguistics.

Biber, D. & Gray, B. (2016). Grammatical complexity in academic English: Linguistic change in writing. *Applied Linguistics*, **37**(6), 887–890.

Biber, D., Larsson, T., & Hancock, G. R. (2023). The linguistic organization of grammatical text complexity: comparing the empirical adequacy of theory-based models. *Corpus Linguistics and Linguistic Theory*.

Biber, D., Larsson, T., & Hancock, G. R. (2024). Dimensions of text complexity in the spoken and written modes: A comparison of theory-based models. *Journal of English Linguistics*, **52**(1), 65–94.

Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2020). Profiling-UD: a tool for linguistic profiling of texts. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Bibliography II

Brunato, D., Dell'Orletta, F., & Venturi, G. (2022). Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on italian. *Frontiers in Psychology*, **13**.

De Clercq, B. (2016). Le développement de la complexité syntaxique en français langue seconde : complexité structurelle et diversité. *SHS Web of Conferences*, **27**, 07006.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, **30**(2), 210–233.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, **44**(3), e12814.

Givón, T. (1991). Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, **15**(2), 335–370.

Hunt, K. W. (1965). A synopsis of clause-to-sentence length factors. *The English Journal*, **54**(4).

Jagaiah, T., Olinghouse, N. G., & Kearns, D. M. (2020). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. *Reading and Writing*, **33**, 2577–2638.

Lehmann, C. (1988). Towards a typology of clause linkage. In J. Haiman and S. A. Thompson, editors, *Clause combining in grammar and discourse*, Typological Studies in Language 18, page 181–225. John Benjamins.

Levshina, N. (2021). Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, **26**(1).

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, **15**(4), 474–496.

Bibliography III

Mačutek, J., Čech, R., & Courtin, M. (2021). The Menzerath-Altmann law in syntactic structure revisited. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 65–73, Sofia, Bulgaria. Association for Computational Linguistics.

Miestamo, M. (2009). Implicational hierarchies and grammatical complexity. In G. Sampson, D. Gil, and P. Trudgill, editors, *Language Complexity as an Evolving Variable*, Oxford Studies in the Evolution of Language. Oxford University Press.

Mondorf, B. (2003). Support for more-support. In G. Rohdenburg and B. Mondorf, editors, *Determinants of Grammatical Variation in English*, pages 251–304. De Gruyter Mouton, Berlin, New York.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in english. *Cognitive Linguistics*, **7**(2), 149–182.

Szmrecsanyi, B. (2004). On operationalizing syntactic complexity. In B. Szmrecsanyi, G. Purnelle, C. Fairon, and A. Dister, editors, *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, volume 2, pages 1032–1039. Presses universitaires de Louvain; Louvain-la-Neuve.

Szmrecsanyi, B. & Kortmann, B. (2012). Introduction: Linguistic complexity: Second language acquisition, indigenization, contact. In B. Kortmann and B. Szmrecsanyi, editors, *Second Language Acquisition, Indigenization, Contact*, pages 6–34. De Gruyter, Berlin, Boston.

Xu, J. & Li, J. (2021). A syntactic complexity analysis of translational English across genres. *Across Languages and Cultures*, **22**(2), 214–232.

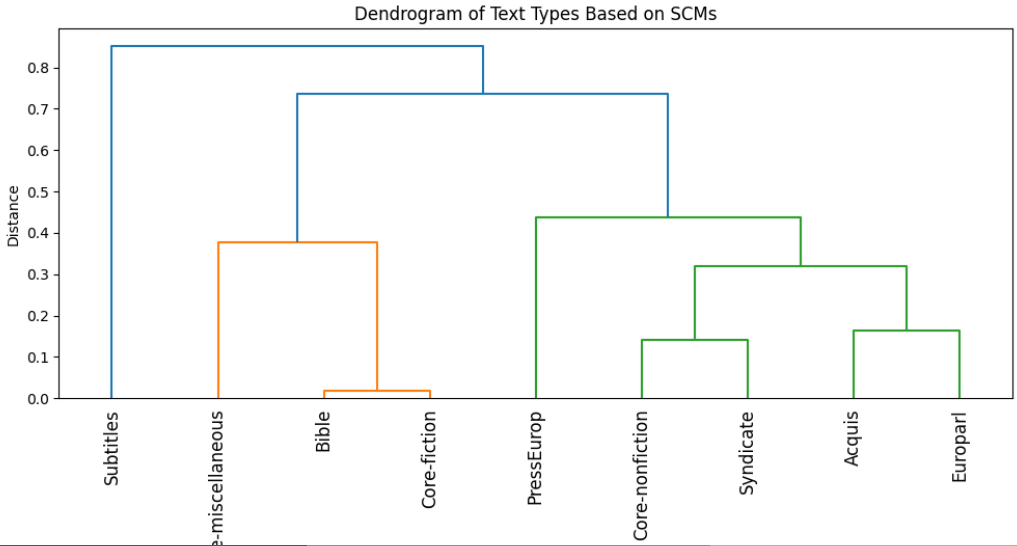
Bibliography IV

Yan, H. & Li, Y. (2019). Beyond length: Investigating dependency distance across L2 modalities and proficiency levels. *Open Linguistics*, **5**(1), 601–614.

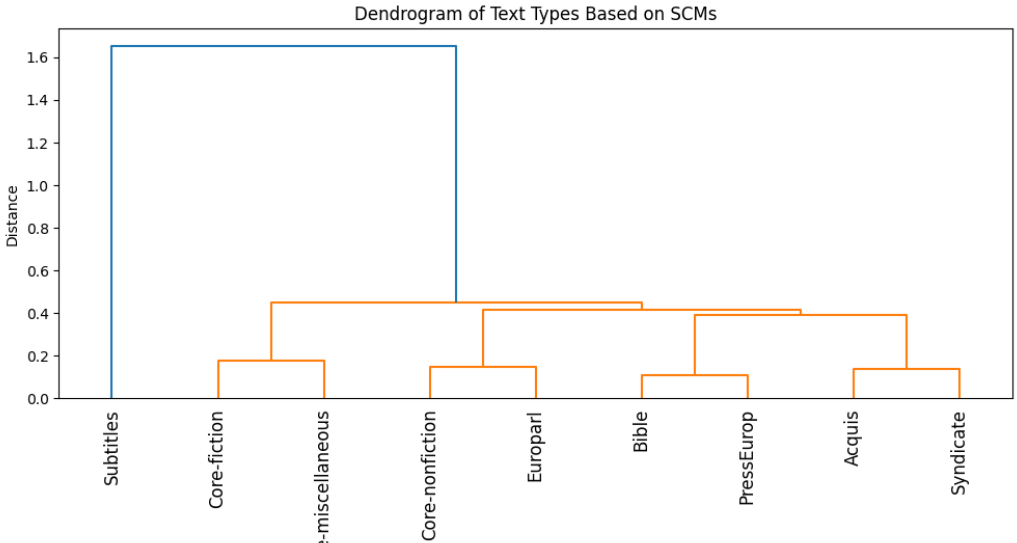
Álvarez González, A., Fernández, Z. E., & Chamoreau, C. (2019). *Diverse scenarios of syntactic complexity*. John Benjamins Publishing Company, Amsterdam.

The End

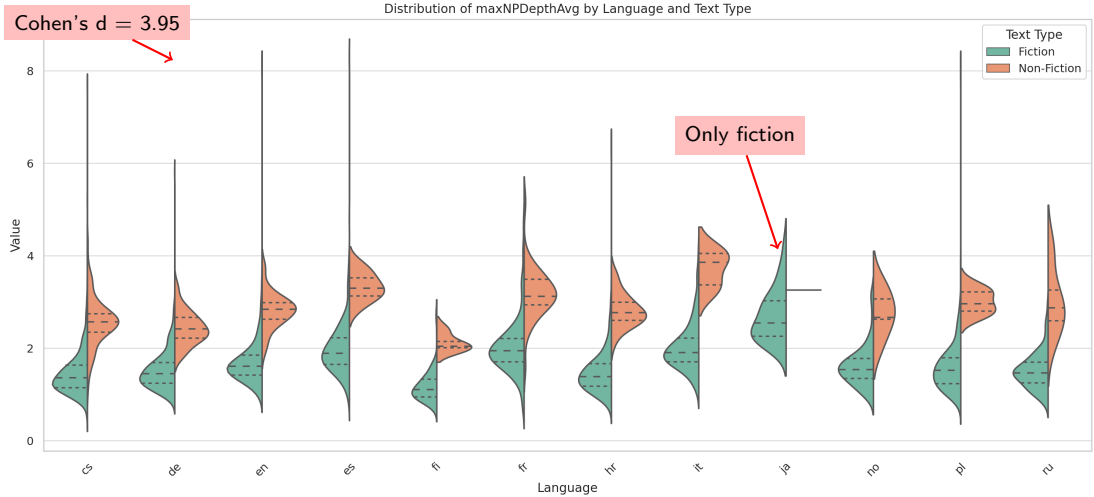
Czech only, maxNPLength



English only, sLength



Fiction vs. non-fiction in maxNPDepth

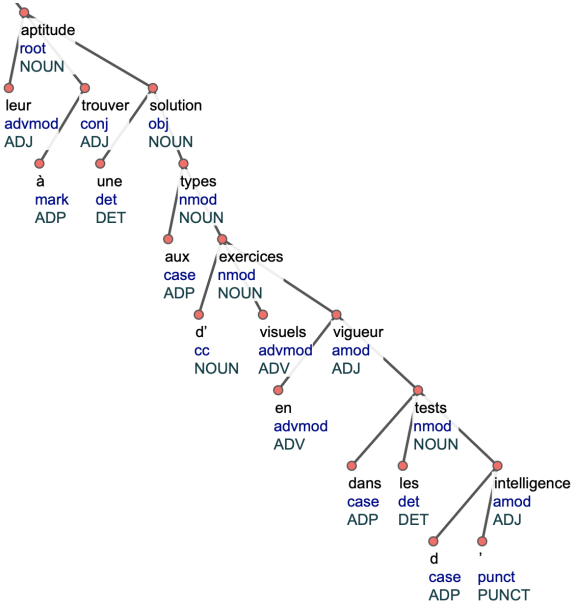


maxNPDepth – French-German example

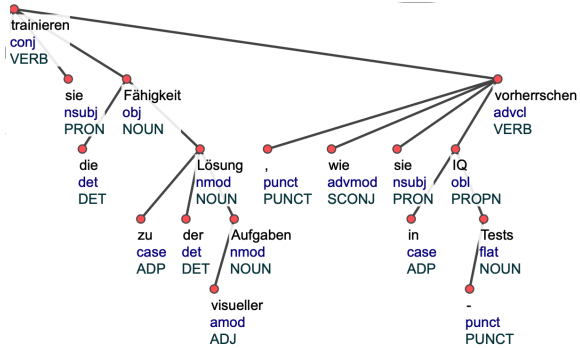
(2) *Les enfants vivent dans un environnement visuel beaucoup plus riche que jadis, ce qui contribue à développer* **leur aptitude à trouver une solution aux types d'exercices visuels en vigueur dans les tests d' intelligence**. (fr, 10)

(3) *Kinder erleben heute ein viel reichhaltigeres visuelles Umfeld als früher, und deshalb trainieren sie* **die Fähigkeit zur Lösung visueller Aufgaben, wie sie in IQ-Tests vorherrschen**. (de, 3)

(4) *Children experience a much richer visual environment than once they did , which helps develop* **their skills in visual puzzles of the kind that dominate IQ tests**. (Matt Ridley: *Genome*, en, 7)



(développer) leur aptitude à trouver une solution aux types d'exercices visuels en vigueur dans les tests d' intelligence



(trainieren sie) die Fähigkeit zur Lösung visueller Aufgaben, wie sie in IQ-Tests vorherrschen

Complexity in a wider context

- **syntactic** complexity (Ferreira, 1991; Givón, 1991; Szmrecsanyi, 2004), *complexité syntaxique* (De Clercq, 2016)
- **cognitive** complexity (Mondorf, 2003; Givón, 1991; Rohdenburg, 1996)
- **clause** complexity (Kuboň: 2001)
- **linguistic** complexity (Schleppegrell: 1992)
- **structural** complexity (Givón: 1991; Arnold et al.: 2000)
- **grammatical / syntactic weight** (Wasow: 1997; Wasow and Arnold: 2003)
- **information density** (Fabricius-Hansen 1999)

All SCMs, averaged normalized (12 languages, all text types)