STaPs'17 — 23 April 2021 (v2)

WS: Corpus analysis

Deriving complex measures from simple annotations

Nils Norman Schiborr
University of Bamberg

Overview

- brief introduction to **multilingual corpora**,
  such as can be used for **corpus-based work in typology**
- closer look at **two of these corpora**
  which feature **specialized annotations**
- **two example analyses:**
  1. *NP length and word order*
  2. *anaphoric distance and subject expression*

Multilingual corpora

- corpora that aggregate data from **multiple languages**,
  usually with the aim of enabling **cross-linguistic comparison**

Multilingual corpora

- some focus on **specific linguistic areas**,
  e.g. *GRALIS* and *ParaSol* on on Slavic languages,
  CorpAfroAs on Afroasiatic languages

- others aim at more **general typological representativity**
  e.g. *Universal Dependencies*, *DoReCo*, etc.

Multilingual corpora

- many focus on languages with **many speakers** (i.e. chiefly European), for which data is relatively easy to come by

- others specifically target **underrepresented and understudied languages**, and often develop out of language documentation projects, e.g. *CorpAfroAs*, *Multi-CAST*, etc.

Multilingual corpora

- many draw mostly from **written sources**, and consequently get quite large,
- others focus exclusively on **spoken languages**, and hence suffer all the concomitant limitations, e.g. *DoReCo*, *SCOPIC*, *Multi-CAST*, etc.

Multilingual corpora

- some aim for **direct comparability** of texts,
  e.g. via use of parallel texts (e.g. Bible and UDHR translations)
  or parallax texts (Pear Film/Frog story recordings)
- others for see corpora as **random samples** from a larger population of texts
  e.g. most original text corpora, web-scraped corpora, etc.

Other projects

- a non-exhaustive list of multilingual corpus projects:
  - *Universal Dependencies* (Zeman et al. 2020)
  - *SCOPIC* (Barth & Evans 2017)
  - *DoReCo* (Paschen et al. 2020)
  - *CorpAfroAs* (Mettouchi et al. 2015)
  - *Multi-CAST* (Haig & Schnell 2021)
  - *GRALIS* (Tošović 2008)
  - *ParaSol* (von Waldenfels & Meyer 2006)
  - *CHILDES* (MacWhinney 1991)
  - *Parallel Bible Corpus* (Mayer & Cysouw 2014)
  - *UDHR corpus* (Cysouw & Wälchli 2007)
  - *EuroParl* (Koehn 2005)
  - *OpenSubtitle2016* (Tiedemann 2012)

  etc.

Corpus-based typology

- **usage-based approaches to linguistic variation**
- **corpus-based** (vs. grammar-based) **typology**
  (Haig et al. 2011; Schnell & Barth to appear: Ch. 11; Schnell & Schiborr submitted)
- *token-based (vs. type-based) typology*
  (Levshina 2019, to appear)
- *typometrics*
  (Gerdes et al. 2021)

Corpus-based typology

- **applies the toolset of corpus linguistics to typological research**
- focus on variation within languages
- seeks to determine **conditioned structural probabilities** of features across languages (vs. *data reduction approaches*, cf. Wälchli 2009)

Late aggregation

- ties in with

  distributional typology: *what is where why?* (Bickel 2007)

  multivariate typology: multiple variants per language (Bickel 2011)
  vs. one value per language (as in WALS)

- Bickel et al. 2016:
  locus of complexity should lie in the analysis, not the data

This workshop

- *here:*
  take a closer look at **two multilingual corpora**,
  one large and one small
- these corpora feature **specialized annotations**
  that go beyond the usual part-of-speech tagging:
  1. a treebank:
     Universal Dependencies (Zeman et al. 2020)
  2. a corpus with co-reference annotations:
     Multi-CAST (Haig & Schnell 2021)

Example analyses

- use these corpora for two **example analyses**

- **the twist**:
  required information cannot be read off directly from the corpus annotations
  (i.e. not simply a matter of searching for an expression
  and then counting the results)

Example analyses

- **need to combine multiple layers of annotation in clever ways**
  to draw out the desired bits of information

- and then **implement** this in code (here in R),
  which brings to notice additional pitfalls and complications,
  but also offers opportunities for further refinement

Procedure

- after introducing each data set, we will
  1. formulate a *research question*,
  2. select the *data*,
  3. design an *algorithm*,
  4. *implement* it in R, and
  5. *evaluate* the results

Universal Dependencies

- **Universal Dependencies 2.0** (Zeman et al. 2020)
  multilingual treebank
  in development since 2014
  published by the University of Prague,
  with contributors from around the world
- mostly **CC-BY** (check individual corpora)
- website:
  `universaldependencies.org/`
- annotation guidelines:
  `universaldependencies.org/guidelines.html`

UD languages

- as of November 2020 (version 2.7):
  183 corpora from **104 languages**, c. **24 million words**
- *Afrikaans, Akkadian, Akuntsu, Albanian, Amharic, Ancient Greek, Apurina, Arabic, Armenian, Assyrian, Bambara, Basque, Belarusian, Bhojpuri, Breton, Bulgarian, Buryat, Cantonese, Catalan, Chinese, Chukchi, Classical Chinese, Coptic, Croatian, Czech, Danish, Dutch, English, Erzya, Estonian, Faroese, Finnish, French, Galician, German, Gothic, Greek, Hebrew, Hindi, Hindi English, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Karelian, Kazakh, Khunsari, Komi Permyak, Komi Zyrian, Korean, Kurmanji, Latin, Latvian, Lithuanian, Livvi, Maltese, Manx, Marathi, Mbya Guarani, Moksha, Munduruku, Naija, Nayini, North Sami, Norwegian, Old Church Slavonic, Old French, Old Russian, Old Turkish, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Skolt Sami, Slovak, Slovenian, Soi, South Levantine Arabic, Spanish, Swedish, Swedish Sign Language, Swiss German, Tagalog, Tamil, Telugu, Thai, Tupinamba, Turkish, Turkish German, Ukrainian, Upper Sorbian, Urdu, Uyghur, Vietnamese, Warlpiri, Welsh, Wolof, Yoruba*

UD languages

- as of November 2020 (version 2.7):
  183 corpora from **104 languages**, c. **24 million words**
- *Afrikaans, Akkadian, Akuntsu, Albanian, Amharic, Ancient Greek, Apurina, Arabic, Armenian, Assyrian, Bambara, Basque, Belarusian, Bhojpuri, Breton, Bulgarian, Buryat, Cantonese, Catalan, Chinese, Chukchi, Classical Chinese, Coptic, Croatian, Czech, Danish, Dutch, English, Erzya, Estonian, Faroese, Finnish, French, Galician, German, Gothic, Greek, Hebrew, Hindi, Hindi English, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Karelian, Kazakh, Khunsari, Komi Permyak, Komi Zyrian, Korean, Kurmanji, Latin, Latvian, Lithuanian, Livvi, Maltese, Manx, Marathi, Mbya Guarani, Moksha, Munduruku, Naija, Nayini, North Sami, Norwegian, Old Church Slavonic, Old French, Old Russian, Old Turkish, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Skolt Sami, Slovak, Slovenian, Soi, South Levantine Arabic, Spanish, Swedish, Swedish Sign Language, Swiss German, Tagalog, Tamil, Telugu, Thai, Tupinamba, Turkish, Turkish German, Ukrainian, Upper Sorbian, Urdu, Uyghur, Vietnamese, Warlpiri, Welsh, Wolof, Yoruba*

Representativity

- **but its large size comes at a price:**
  well over half of the languages are Indo-European!
- **huge disparity in corpus sizes:**
  the largest (from German) has *3 million words*,
  the smallest (from Soi, C. Iranian) has *47*
  (mean size = 129 435 words, SD = 285 659 words)
- the largest corpora are from what Dahl (2015) calls **"LOL" languages**:
  "Literate, Official, with Lots of users"

Corpus composition

- **UD corpora consist largely of written texts,**
  only some corpora include spoken sections,
  most are entirely written

- *problem*:
  in corpora composed of multiple "sections" from different text types,
  a lack of low-level metadata makes identifying the text type of
  specific segments impossible (?)

Corpus composition

- while these issues limit UDs' usability for *certain types of inquiries* and overall *cross-linguistic comparability,*

- **UDs are nevertheless a hugely valuable resource with enormous potential!**
  (and hence quite popular at the moment)

Caveats

- **careful pre-selection of corpus data is crucial!**
- *how much is there?* *(i.e. is there enough for my purposes?)*
  what's in there? *(i.e. modes, text types, etc.)*
  how is it structured? *(i.e. file formats, annotations, etc.)*
  what metadata is there?
  etc.

Caveats

- **careful pre-selection of corpus data is crucial!**
- *how much is there?* *(i.e. is there enough for my purposes?)*
  what's in there? *(i.e. modes, text types, etc.)*
  how is it structured? *(i.e. file formats, annotations, etc.)*
  what metadata is there?
  etc.

- *here:*
  throw it all in and hope for the best!

File formats and tools

- annotation data are provided as *CoNLL-U files*
  (i.e. a text file with specific formatting)
- for metadata, *see the UD website*
- also: numerous *tools for analysis and data visualization*,
  either standalone software or (mostly) Python and Perl libraries

File formats and tools

- annotation data are provided as *CoNLL-U files*
  (i.e. a text file with specific formatting)
- for metadata, *see the UD website*
- also: numerous *tools for analysis and data visualization*,
  either standalone software or (mostly) Python and Perl libraries
- *here:*
  use R with a custom import script

UD annotations

1. (*transcription*)
2. *lemmatization*
3. **part-of-speech tagging** (with custom tags)
4. *morphological features* (not LGR)
5. **syntactic relations** (as per dependency grammar)
6. *additional comments*

- uses **unified symbol sets** for each level,
  mostly consistently applied to all corpora
  (though especially with morphological features there can be differences)

Dependency relations

- **treebank**
- **dependency relations** (de Marneffe et al. 2006, 2008, 2014):
  - every "word" is dependent on one other word (its "*head*");
  - exactly one word is the head of the sentence (the "*root*"),
  - a word can have multiple dependents, but only one head;
  - content words are preferred as heads over function words (for better parallelism between languages)
- (not all grammatical relations can be reduced to binary head–dependent pairs, so these are by necessity simplifications)
- (there's also a set of enhanced dependencies in UD, but we'll ignore those here)

Dependencies

Jane was reading a new book on linguistics.

Dependencies

words	Jane	was	reading	a	new	book	on	linguistics.

Dependencies

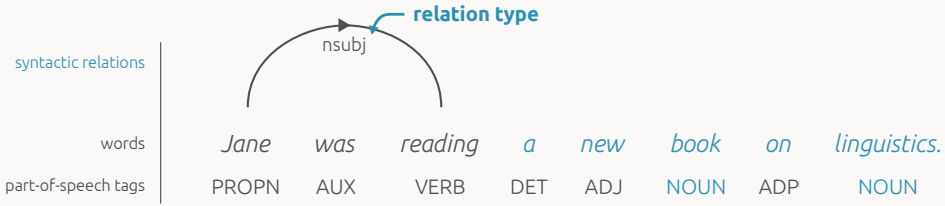
words	Jane	was	reading	a	new	book	on	linguistics.

Dependencies

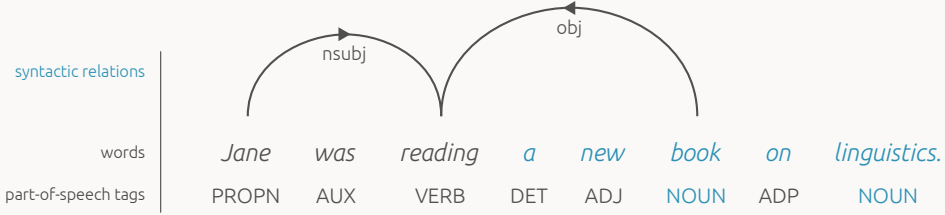
words	*Jane*	*was*	*reading*	*a*	*new*	*book*	*on*	*linguistics.*
part-of-speech tags	PROPN	AUX	VERB	DET	ADJ	NOUN	ADP	NOUN

PoS tag

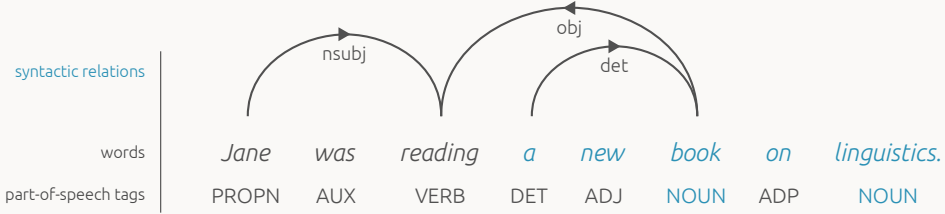
Dependencies



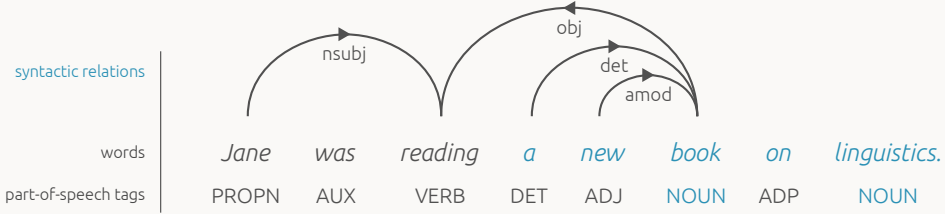
Dependencies



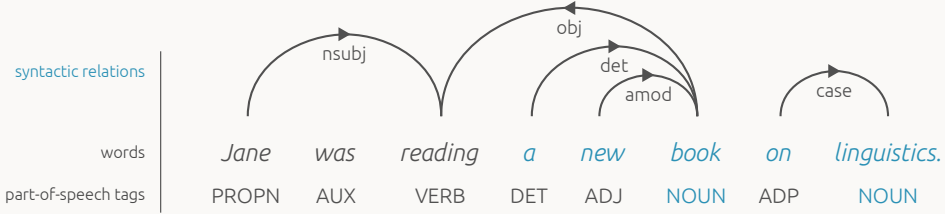
Dependencies



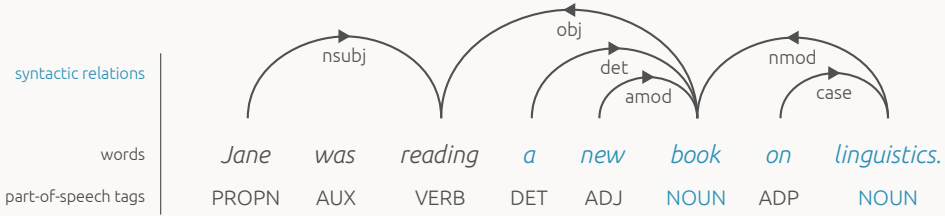
Dependencies



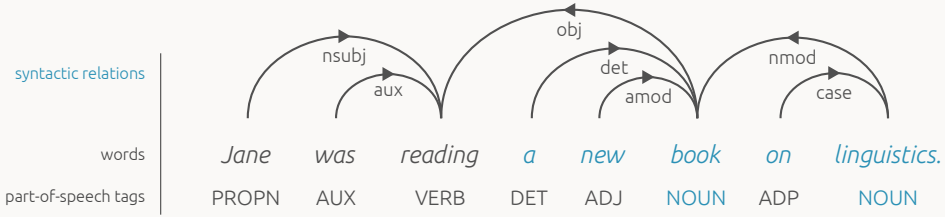
Dependencies



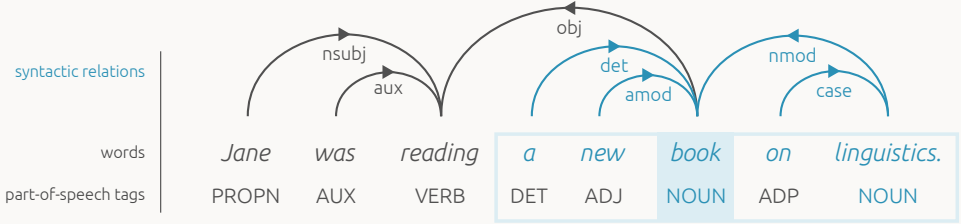
Dependencies



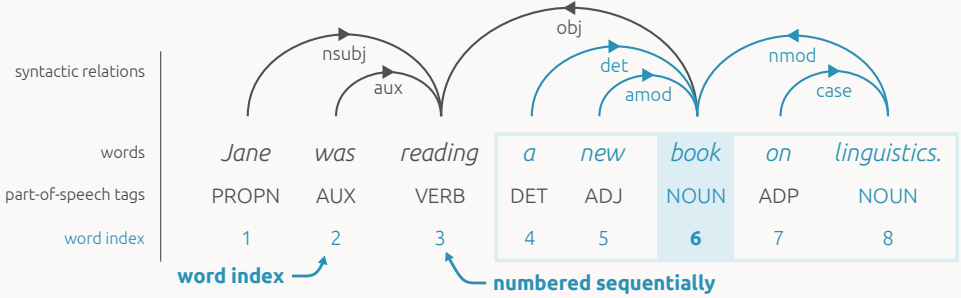
Dependencies



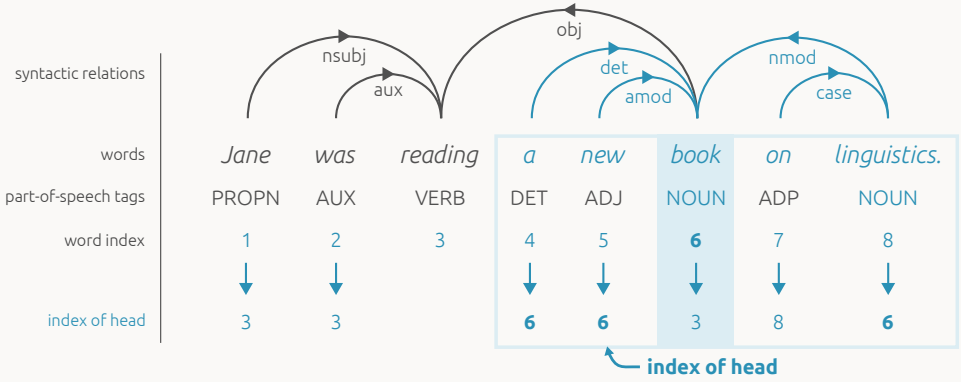
Dependencies



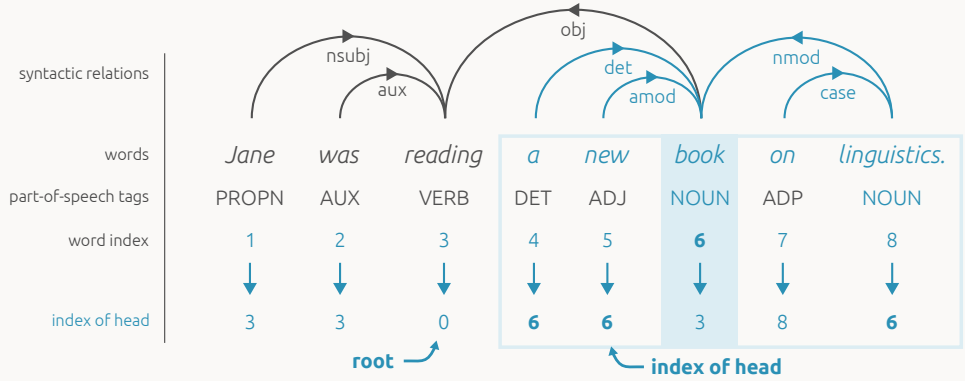
Dependencies



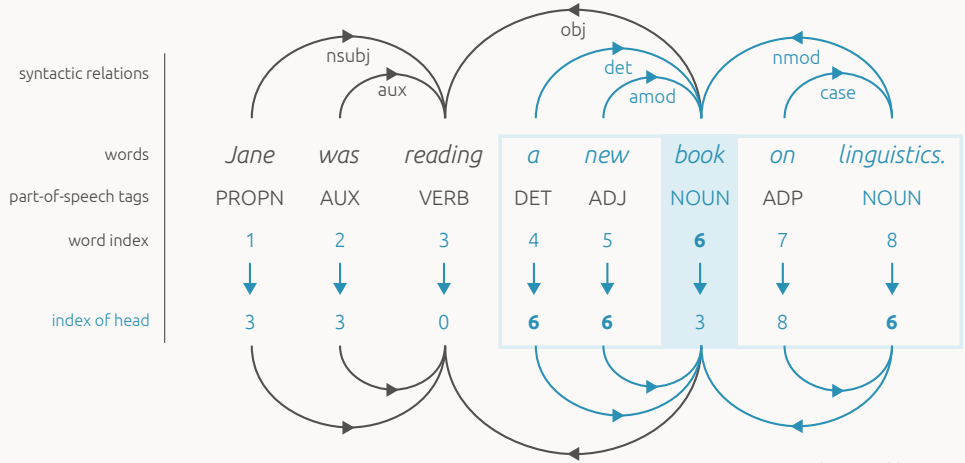
Dependencies



Dependencies



Dependencies



Example A: NP lengths

- **NP length** as a measure of *complexity/information content* (Wasow 1997)
- **word order**: *heavier constituents* tend to be placed *later* in the clause,
  e.g. dative alternation, heavy NP shift, etc. (Arnold et al. 2000)
- cf. also Futrell et al. (2020) on dependency distances and word order

- *here:*
  evaluate position *relative to the predicate* (i.e. the dependency head)
  (an admittedly fairly crude measure, chosen for simplicity's sake)

Analysis design

- **three basic issues:**
  1. *which expressions to consider?*
  2. *how to measure the length of expressions?*
  3. *how to identify their position in the clause?*

Which expressions?

- only core argument NPs, i.e. **subjects** and (in)direct **objects**
  (with dependency relation types ⟨nsubj⟩, ⟨obj⟩, or ⟨iobj⟩)

- only NPs with **common nouns as heads** (i.e. common "lexical" NPs)
  (with the part-of-speech tag ⟨NOUN⟩)

Only lexical

- *why not include all types of expressions (i.e. pronouns)?*

Only lexical

- *why not include all types of expressions (i.e. pronouns)?*

- **inclusion of pronouns would also require the consideration of elliptical arguments (i.e. zero)**,
  otherwise NP lengths would be skewed upwards in languages that strongly prefer zero (e.g. Japanese, Mandarin Chinese, etc.), since they would not be counted as length $l = 0$, but be entirely missing from the sample
  (also: what's the position of a zero argument?)

Only lexical

- *why not include all types of expressions (i.e. pronouns)?*

- **inclusion of pronouns would also require the consideration of elliptical arguments (i.e. zero)**,
  otherwise NP lengths would be skewed upwards in languages that strongly prefer zero (e.g. Japanese, Mandarin Chinese, etc.), since they would not be counted as length l=0, but be entirely missing from the sample
  (also: what's the position of a zero argument?)

- **but in any case:**
  there's no (straightforward) way of capturing zero arguments in UD corpora!

No reverse dependencies

- UD annotations do not mark **"reverse dependencies"**,
  i.e. while the identity of the head can be directly read off a dependent,
  we cannot easily find all dependents of a head

No reverse dependencies

- *so how do we go about doing this?*

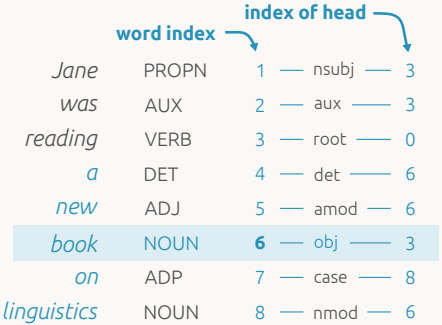
No reverse dependencies

- *so how do we go about doing this?*
- **though reverse dependencies are not explicitly marked, all the information required for identifying them is there**

Calculating NP lengths

		word index
Jane	PROPN	1
was	AUX	2
reading	VERB	3
a	DET	4
new	ADJ	5
book	NOUN	**6**
on	ADP	7
linguistics	NOUN	8

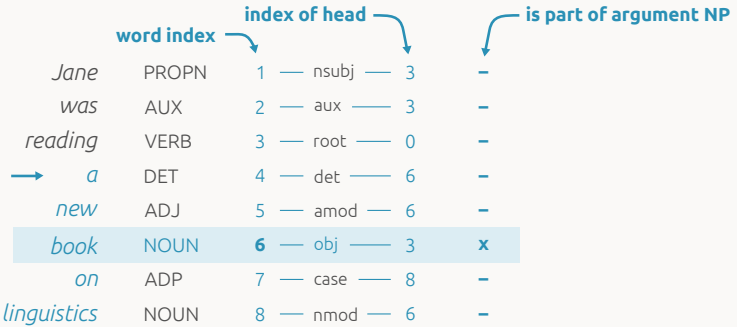
Calculating NP lengths



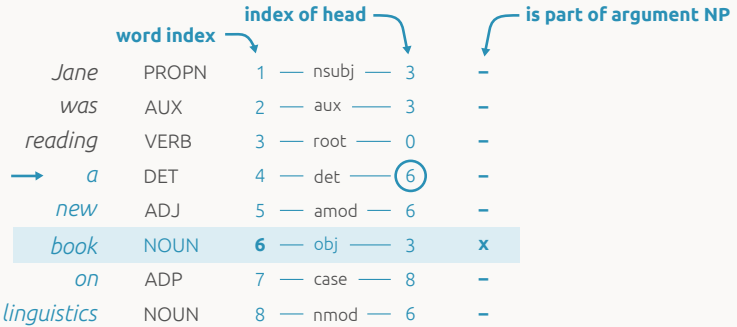
Calculating NP lengths

		word index		index of head		is part of argument NP
Jane	PROPN	1	nsubj	3		–
was	AUX	2	aux	3		–
reading	VERB	3	root	0		–
a	DET	4	det	6		–
new	ADJ	5	amod	6		–
book	NOUN	**6**	obj	3		**x**
on	ADP	7	case	8		–
linguistics	NOUN	8	nmod	6		–

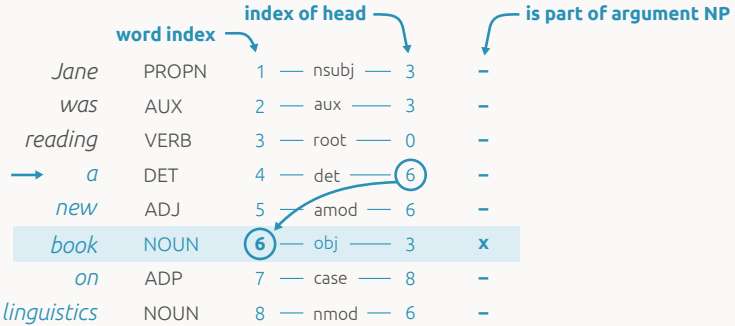
Calculating NP lengths



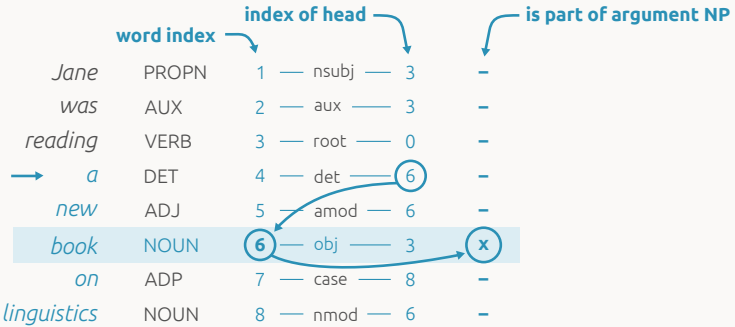
Calculating NP lengths



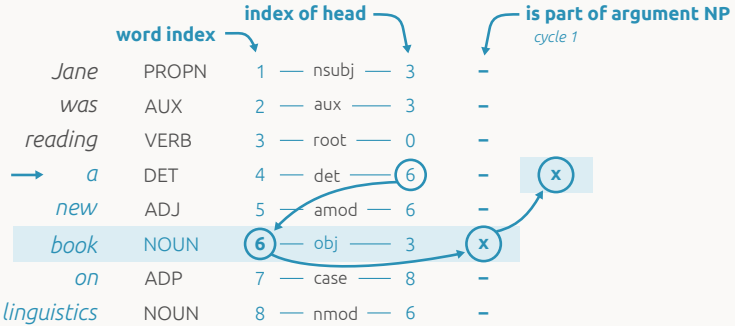
Calculating NP lengths



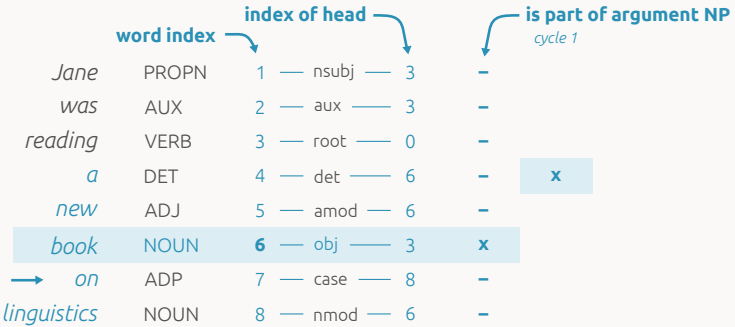
Calculating NP lengths



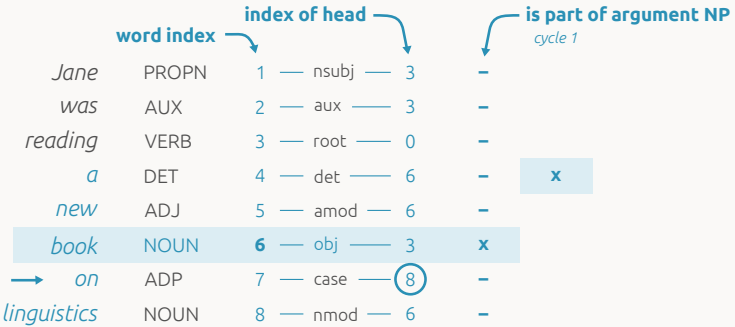
Calculating NP lengths



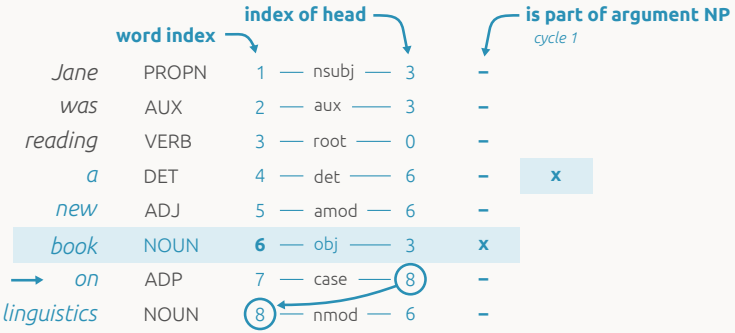
Calculating NP lengths



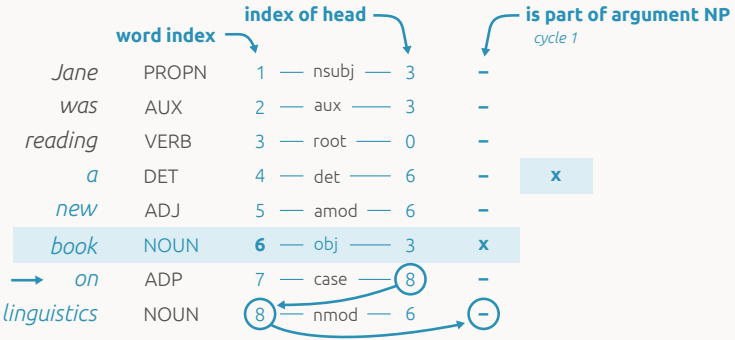
Calculating NP lengths



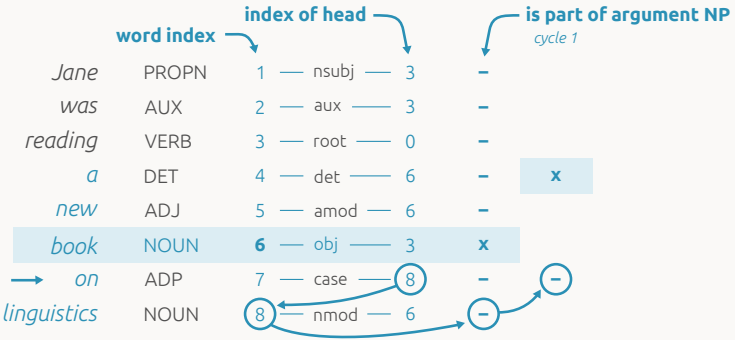
Calculating NP lengths



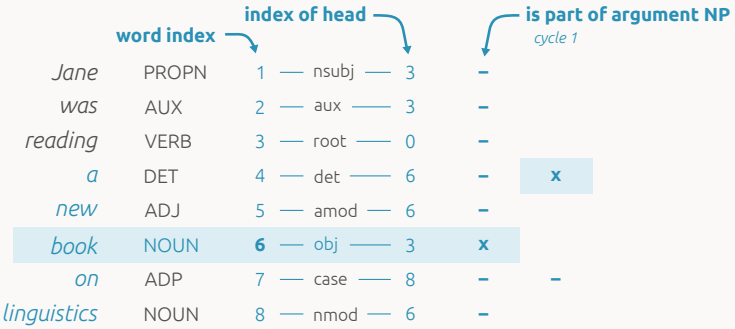
Calculating NP lengths



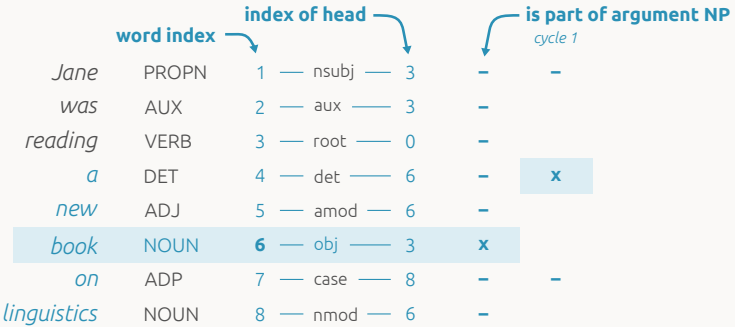
Calculating NP lengths



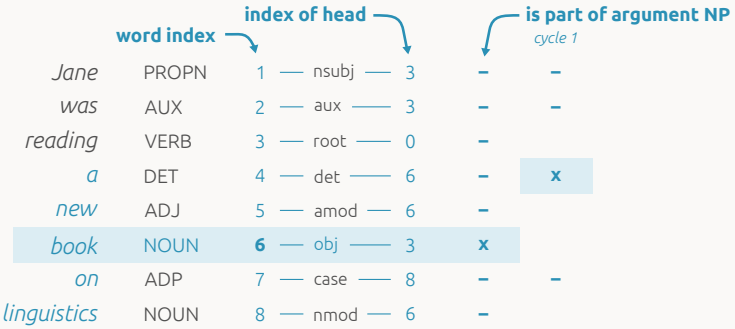
Calculating NP lengths



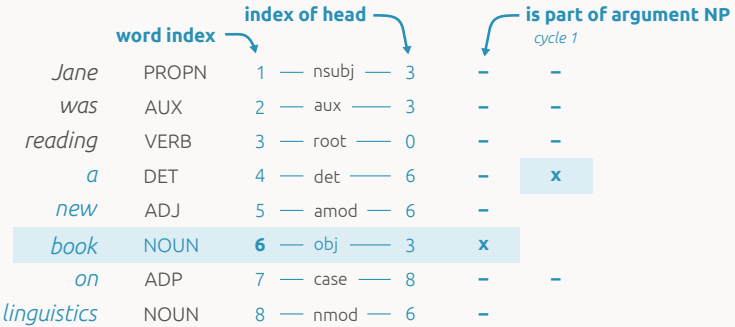
Calculating NP lengths



Calculating NP lengths



Calculating NP lengths



Calculating NP lengths

word index		index of head			is part of argument NP (cycle 1)	
Jane	PROPN	1	nsubj	3	–	–
was	AUX	2	aux	3	–	–
reading	VERB	3	root	0	–	–
a	DET	4	det	6	–	x
new	ADJ	5	amod	6	–	x
book	NOUN	**6**	obj	3	x	
on	ADP	7	case	8	–	–
linguistics	NOUN	8	nmod	6	–	

Calculating NP lengths

word index			index of head			is part of argument NP (cycle 1)	
Jane	PROPN	1	nsubj	3	–	–
was	AUX	2	aux	3	–	–
reading	VERB	3	root	0	–	–
a	DET	4	det	6	–	x
new	ADJ	5	amod	6	–	x
book	NOUN	**6**	obj	3	x	
on	ADP	7	case	8	–	–
linguistics	NOUN	8	nmod	6	–	x

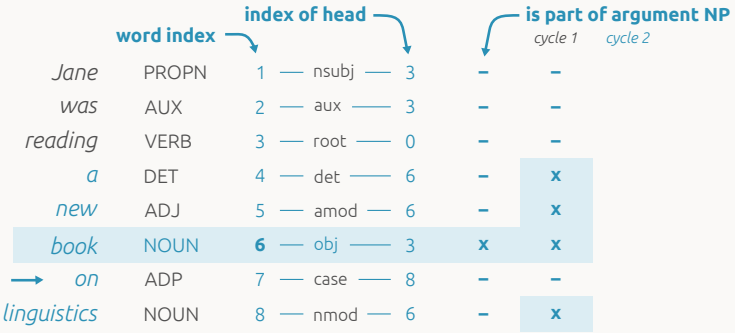
Calculating NP lengths

word index		index of head			is part of argument NP (cycle 1)	
Jane	PROPN	1	nsubj	3	–	–
was	AUX	2	aux	3	–	–
reading	VERB	3	root	0	–	–
a	DET	4	det	6	–	x
new	ADJ	5	amod	6	–	x
book	NOUN	**6**	obj	3	x	x
on	ADP	7	case	8	–	–
linguistics	NOUN	8	nmod	6	–	x

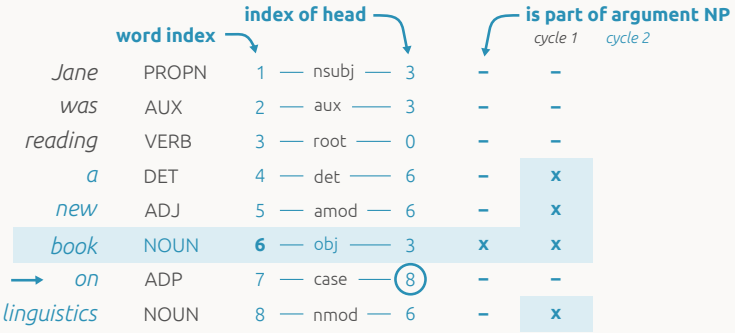
Calculating NP lengths

word index		index of head			is part of argument NP (cycle 1)	is part of argument NP (cycle 2)
Jane	PROPN	1	nsubj	3	–	–
was	AUX	2	aux	3	–	–
reading	VERB	3	root	0	–	–
a	DET	4	det	6	–	x
new	ADJ	5	amod	6	–	x
book	NOUN	**6**	obj	3	x	x
on	ADP	7	case	8	–	–
linguistics	NOUN	8	nmod	6	–	x

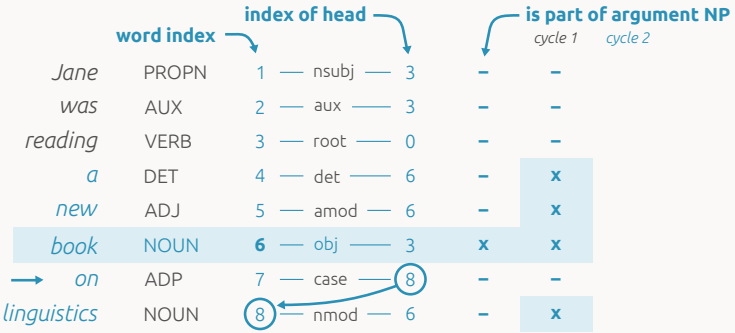
Calculating NP lengths



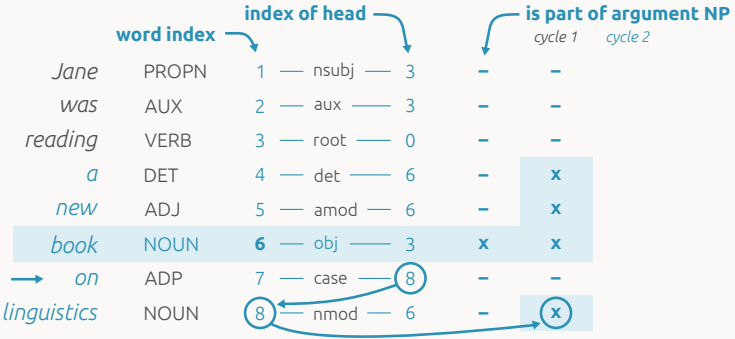
Calculating NP lengths



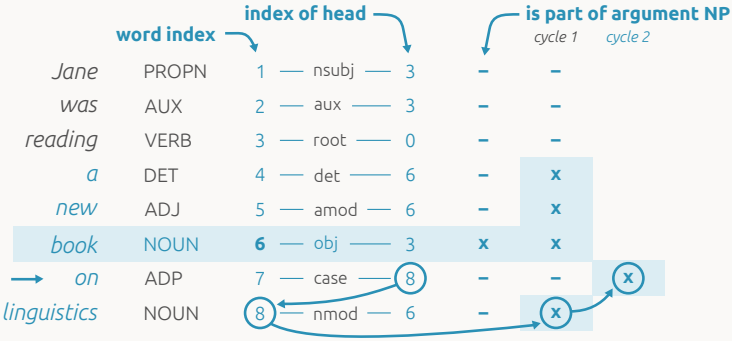
Calculating NP lengths



Calculating NP lengths



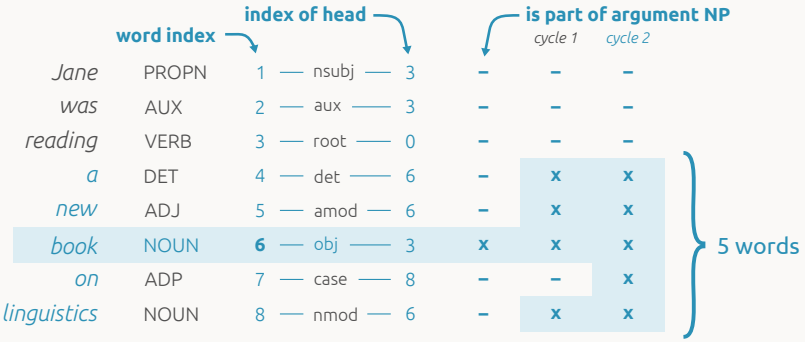
Calculating NP lengths



Calculating NP lengths

word index		index of head			is part of argument NP (cycle 1)	is part of argument NP (cycle 2)	
Jane	PROPN	1	nsubj	3	–	–	–
was	AUX	2	aux	3	–	–	–
reading	VERB	3	root	0	–	–	–
a	DET	4	det	6	–	x	x
new	ADJ	5	amod	6	–	x	x
book	NOUN	**6**	obj	3	x	x	x
on	ADP	7	case	8	–	–	x
linguistics	NOUN	8	nmod	6	–	x	x

Calculating NP lengths



Calculating more NP lengths

		word index		index of head		is part of argument NP
Jane	PROPN	1	nsubj	3		–
was	AUX	2	aux	3		–
reading	VERB	3	root	0		–
a	DET	4	det	6		–
new	ADJ	5	amod	6		–
book	NOUN	**6**	obj	3		**x**
on	ADP	7	case	8		–
linguistics	NOUN	8	nmod	6		–
that	PRON	9	obj	13		–
her	DET	10	det	11		–
professor	NOUN	11	nsubj	13		–
had	AUX	12	aux	13		–
recommended	VERB	13	acl	6		–

Calculating more NP lengths

		word index		index of head		is part of argument NP (cycle 1)	
Jane	PROPN	1	nsubj	3	–	–
was	AUX	2	aux	3	–	–
reading	VERB	3	root	0	–	–
a	DET	4	det	6	–	x
new	ADJ	5	amod	6	–	x
book	NOUN	**6**	obj	3	x	x
on	ADP	7	case	8	–	–
linguistics	NOUN	8	nmod	6	–	x
that	PRON	9	obj	13	–	–
her	DET	10	det	11	–	–
professor	NOUN	11	nsubj	13	–	–
had	AUX	12	aux	13	–	–
recommended	VERB	13	acl	6	–	x

Calculating more NP lengths

word index		index of head			is part of argument NP	cycle 1	cycle 2
Jane	PROPN	1	nsubj	3	–	–	–
was	AUX	2	aux	3	–	–	–
reading	VERB	3	root	0	–	–	–
a	DET	4	det	6	–	x	x
new	ADJ	5	amod	6	–	x	x
book	NOUN	**6**	obj	3	x	x	x
on	ADP	7	case	8	–	–	x
linguistics	NOUN	8	nmod	6	–	x	x
that	PRON	9	obj	13	–	–	x
her	DET	10	det	11	–	–	–
professor	NOUN	11	nsubj	13	–	–	x
had	AUX	12	aux	13	–	–	x
recommended	VERB	13	acl	6	–	x	x

Calculating more NP lengths

		word index		index of head		is part of argument NP cycle 1	cycle 2	cycle 3
Jane	PROPN	1	nsubj	3	–	–	–	–
was	AUX	2	aux	3	–	–	–	–
reading	VERB	3	root	0	–	–	–	–
a	DET	4	det	6	–	x	x	x
new	ADJ	5	amod	6	–	x	x	x
book	NOUN	**6**	obj	3	x	x	x	x
on	ADP	7	case	8	–	–	x	x
linguistics	NOUN	8	nmod	6	–	x	x	x
that	PRON	9	obj	13	–	–	x	x
her	DET	10	det	11	–	–	–	x
professor	NOUN	11	nsubj	13	–	–	x	x
had	AUX	12	aux	13	–	–	x	x
recommended	VERB	13	acl	6	–	x	x	x

Calculating more NP lengths

		word index		index of head	is part of argument NP (cycle 1)	(cycle 2)	(cycle 3)
Jane	PROPN	1	nsubj	3	–	–	–	–
was	AUX	2	aux	3	–	–	–	–
reading	VERB	3	root	0	–	–	–	–
a	DET	4	det	6	–	x	x	x
new	ADJ	5	amod	6	–	x	x	x
book	NOUN	**6**	obj	3	x	x	x	x
on	ADP	7	case	8	–	–	x	x
linguistics	NOUN	8	nmod	6	–	x	x	x
that	PRON	9	obj	13	–	–	x	x
her	DET	10	det	11	–	–	–	x
professor	NOUN	11	nsubj	13	–	–	x	x
had	AUX	12	aux	13	–	–	x	x
recommended	VERB	13	acl	6	–	x	x	x

10 words

Making associations

- *how do we make the associations between word indices and their heads automatically?*

Making associations

- *how do we make the associations between word indices and their heads automatically?*

- **(repeatedly) join the table with itself, matching the head indices with the word indices, then copy over the markers for NP constituency**

Stopping criteria

- *how do we know when to stop running additional cycles,*
  i.e. when do we know that all NP subconstituents have been associated?

Stopping criteria

- *how do we know when to stop running additional cycles,*
  i.e. when do we know that all NP subconstituents have been associated?

- **count the number of associations made each cycle:**
  - **(A)** if the previous cycle made N > 0 associations, *run another cycle*
  - **(B)** if the previous cycle made N = 0 associations, *stop cycling*

Position in the clause

- *how do we tell where an argument NP is located in the clause, relative to the predicate?*

Position in the clause

- *how do we tell where an argument NP is located in the clause, relative to the predicate?*

- **compare the word index of the head of the NP with the word index of its head (i.e. the predicate of the clause):**
  - **(A)** if the former is lower than the latter, the NP *precedes* the predicate
  - **(B)** if the former is higher than the latter, the NP *follows* the predicate

Implementation in R

- **workshop website:** `tinyurl.com/2sja5stc`
- **download R scripts and corpus data**
- *two options:*
  - **(A)** reduced data set (c. 25 MB)
  - **(B)** full data set (c. 380 MB)

- *here:*
  use *RDS files*, a compressed serialization format for R objects;
  raw *CoNLL-U files* can be found on the UD website

Preparations

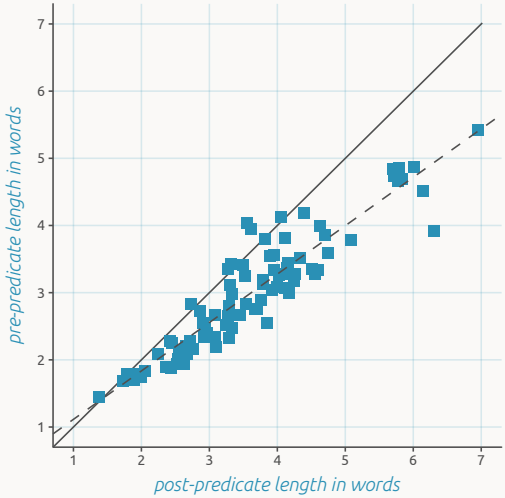
- **step –3**:
  download and unzip the *R scripts*
- **step –2**:
  unzip your *data set* of choice directly into the "data/" folder
  (or alter the file path in the script to your liking)
- **step –1:**
  open the "_preparation.R" script in RStudio and install the *packages* listed there

- **R not built for efficiently handling very large amounts of data**
- use the **data.table** package (Dowle & Srinivasan 2021):
  - chiefly a replacement for base R's `data.frame`
  - hugely more efficient with large data sets
  - bonus: more elegant syntax

More preparations

- **step 0a**:
  open the "example-A_NP-lengths-in-UDs.R" script
- **step 0b**:
  you may have to set your working directory to the location of the scripts

Results



Mean length of lexical NP arguments by position relative to the predicate in Universal Dependency corpora from 88 languages

Observations

1. **post-predicate argument NPs** tend to be **longer** than pre-predicate ones

2. **substantial cross-linguistic variation** obtains in the **overall length of argument NPs** (at least partially due to variability in text types)

3. there is a noticable trend for languages with **longer NPs overall** to skew more strongly towards **heavier post-predicate arguments**

Observations

1. **post-predicate argument NPs** tend to be **longer** than pre-predicate ones
2. **substantial cross-linguistic variation** obtains in the **overall length of argument NPs** (at least partially due to variability in text types)
3. there is a noticable trend for languages with **longer NPs overall** to skew more strongly towards **heavier post-predicate arguments**

- missing piece: *word order preferences in each language*

Taking this further

- **juxtapose NP lengths with preferred word order in each language,**
  either as a categorical variable,
  or as proportion of subjects/objects before/after predicate
  (though the latter may require filtering by type of clause first)

- **compare written and spoken texts:**
  long NPs are largely the domain of written modes of language,
  majority of NPs in spoken language tend to be quite short and flat

Taking this further

- **juxtapose NP lengths with preferred word order in each language,**
  either as a categorical variable,
  or as proportion of subjects/objects before/after predicate
  (though the latter may require filtering by type of clause first)

- **compare written and spoken texts:**
  long NPs are largely the domain of written modes of language,
  majority of NPs in spoken language tend to be quite short and flat

- *but that's all beyond the scope of this workshop!*

The Multi-CAST collection

- **The Multilingual Corpus of Annotated Spoken Texts** (Haig & Schnell 2015)
  in development since 2014
  at the Dept. of General Linguistics, University of Bamberg
- entirely **CC-BY** (some recordings are in the public domain)
- website:
  `multicast.aspra.uni-bamberg.de/`

The Multi-CAST collection

- chiefly based on **language documentation data**
  from small, underrepresented languages
  (plus some long-hanging fruit, e.g. English, Mandarin, Persian)
- **exclusively spoken data**
- as of January 2021 (version 2101):
  corpora from **13 languages**, c. 100 000 words
- (more to be added in the near-ish future)

Multi-CAST languages

Content and annotations

0. audio recordings

1. transcriptions

2. English translations

3. morphological glossing (as per LGR)

4. multiple levels of specialized annotations (more to be added)

- annotations **applied uniformly to all corpora**
  (some corpora still lack certain annotation levels)

Annotations: GRAID

- **annotations of the form and function of major clause constituents**
  (GRAID, Haig & Schnell 2014)
- small set of annotation symbols with a simple combinatory syntax
- align with individual words, but target entire phrases
- capture **broad cross-linguistic categorizations**,
  but can be refined through **symbol extensions**
  (e.g. ⟨pro⟩ 'free definite pronoun' to ⟨dem_pro⟩ 'demonstrative used as pronoun')

Annotations: GRAID

- crucial aspects of the annotations for our purposes:
  - mark the **syntactic function of expressions**,
  - mark the **form of expressions**, including **elliptical arguments** (i.e. zero),
  - mark the left and right **boundaries of clauses**

Annotations: RefIND

- **annotations for co-reference relations**
  (RefIND, Schiborr et al. 2018)
- also: information status of new referents (Schiborr et al. 2018: 15)
- see also other schemas for co-reference tracking,
  e.g. UCREL, etc.

Annotations: RefIND

- assign a **unique numerical index**
  to each occurrence of a **discourse referent** in a text
- allows **tracking of referents** throughout a text
- extend annotations with GRAID:
  align with heads of phrases,
  yielding a tight bundle of information on discourse expressions

Annotations

- designed for investigations into the **interface between discourse and grammar**, **reference in discourse**, and other discourse phenomena
  (Haig & Schnell 2016; Haig et al. to appear; Schnell et al. to appear; Schiborr 2021)

- i.e. *more focused* than the generalist UD treebanks and PoS tags, but still *extremely versatile*

File formats

- annotation values are provided as

  EAF files (used by the linguistic annotation software ELAN),
  XML files, and
  TSV files

- metadata as *TSV files*

- **companion R package:**
  multicastR (Schiborr 2020)

- directly access annotation data and metadata in R,
  plus a number of utility functions

Anaphora

- what are **discourse anaphors**?
  textual references to previously mentioned entities (i.e. discourse referents)
- effected via **referring expressions**

Referential choice

- speakers choose the most appropriate referring expressions
  based on the *salience/activation/accessibility*
  of the underlying discourse referent
  (Chafe 1976; Givón 1983; Ariel 1990; etc.)
- the more *salient/activated/accessible* a referent,
  the less informative and distinctive the expression needs to be
- **referential scales:**
  zero > *pronominal NPs* > *lexical NPs* > *proper names*

Referential choice

- notion of **topicality**

- **some criteria**:

animacy:	*human*	>	*non-human*
recency:	*recent*	>	*distant*
discourse prominence:	*more frequent*	>	*less frequent*
syntactic prominence:	*subject*	>	*object (> other positions)*
etc.			

Textual distances

- *how to calculate the textual "distance" between two elements (words, phrases, referring expressions, etc.) in a text?*

Textual distances

- *how to calculate the textual "distance" between two elements (words, phrases, referring expressions, etc.) in a text?*

- **manually:**
  start from one and count backwards/forwards to the other (in whatever unit we're measuring in)

Textual distances

- *how to calculate the textual "distance" between two elements (words, phrases, referring expressions, etc.) in a text?*

- **manually:**
  start from one and count backwards/forwards to the other (in whatever unit we're measuring in)

- **computationally:**
  let's find out!

Textual distances

- start with **word distances**, as easier to understand,
  if less useful than other measures
- *probably best:*
  distance measurements in **clause units**,
  here defined as a predicate plus arguments and any adjuncts
- closest representation of how speech is parsed cognitively (cf. Chafe 1976)

Calculating word distances

Then Jane looked out the window. She could see the blue sky.

Calculating word distances

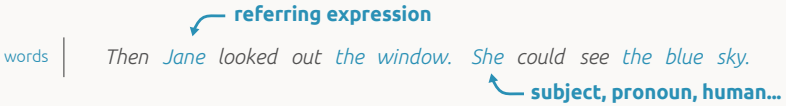
words	Then Jane looked out the window. She could see the blue sky.

Calculating word distances

referring expression

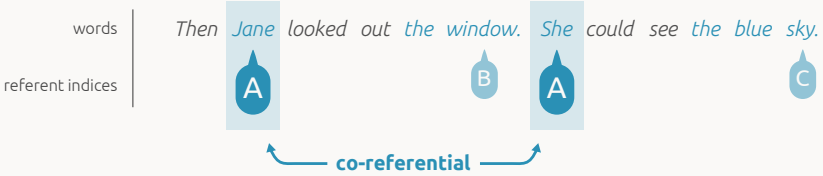
words	Then	Jane	looked	out	the	window.	She	could	see	the	blue	sky.

Calculating word distances

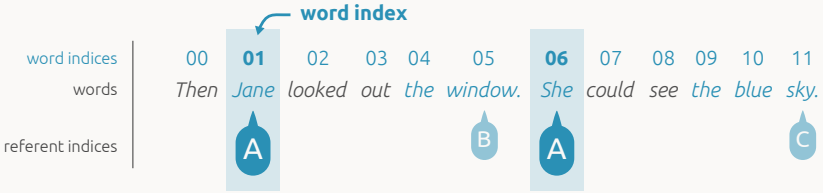


Calculating word distances

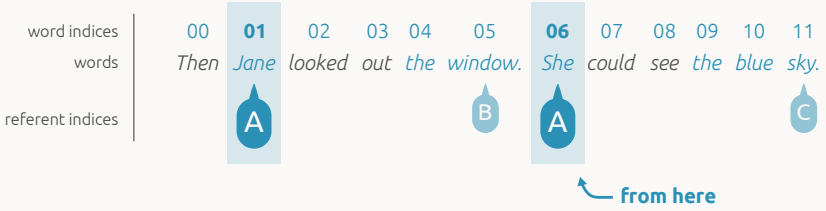
Calculating word distances



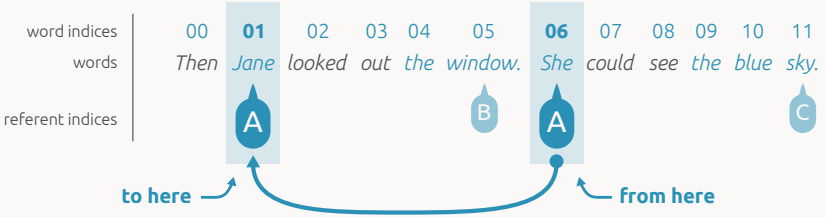
Calculating word distances



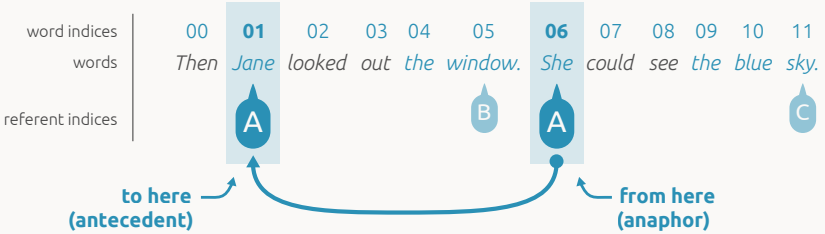
Calculating word distances



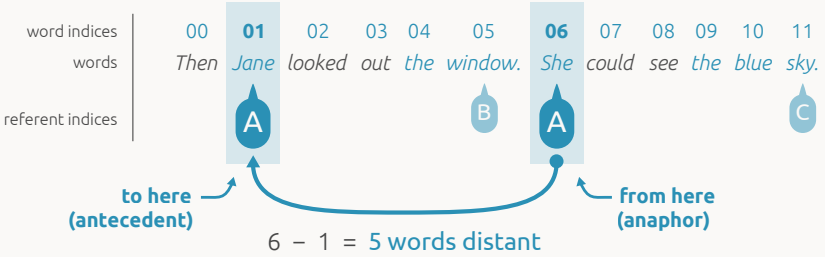
Calculating word distances



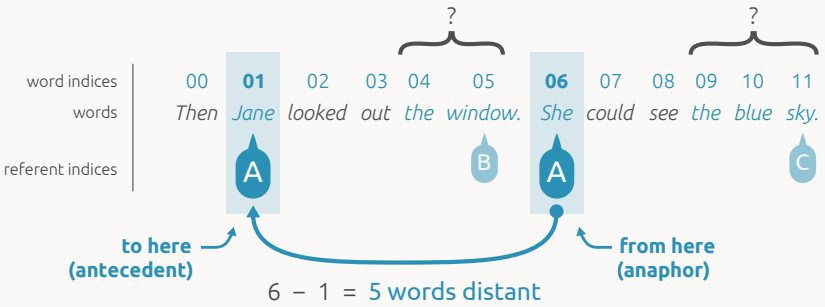
Calculating word distances



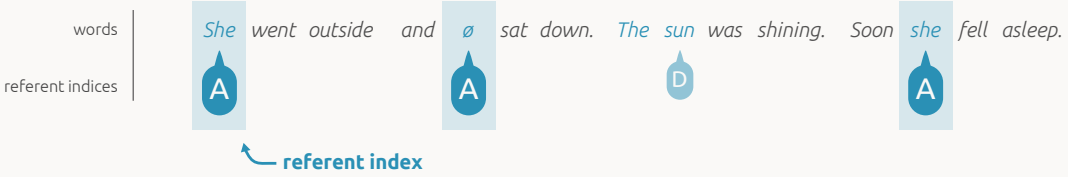
Calculating word distances



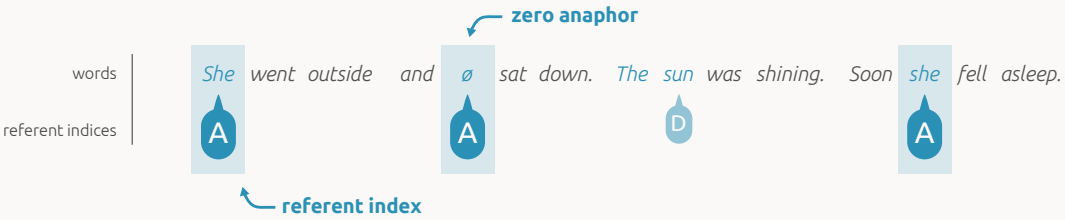
Calculating word distances



Calculating clause distances

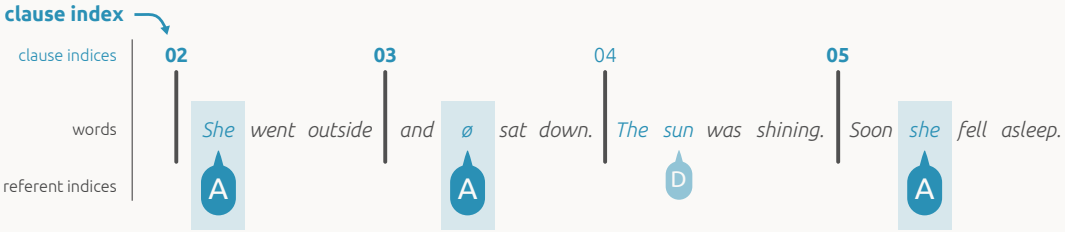


Calculating clause distances

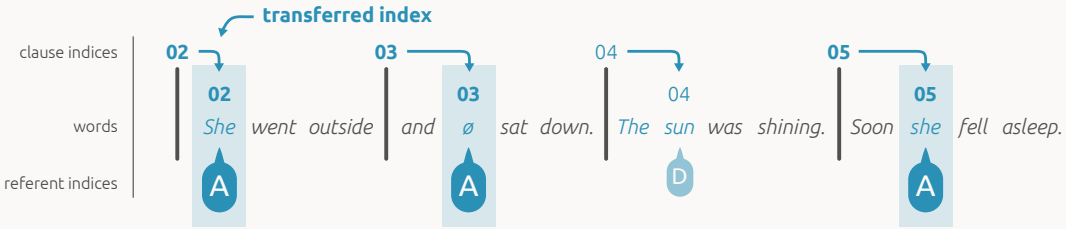


Calculating clause distances

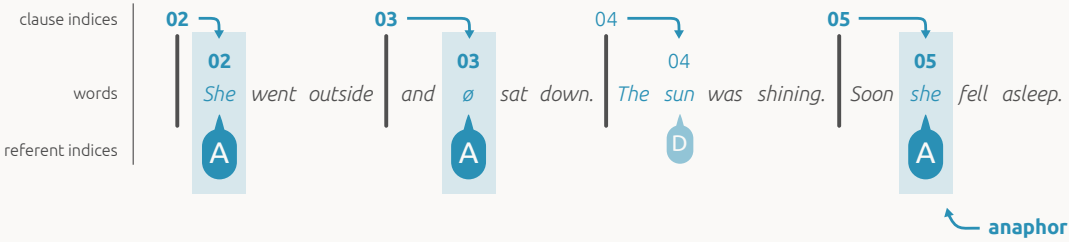
Calculating clause distances



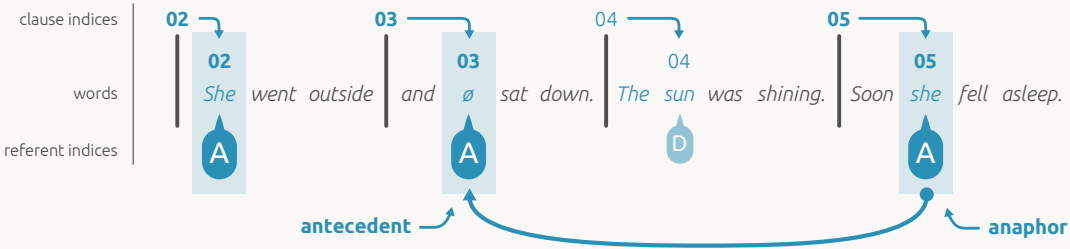
Calculating clause distances



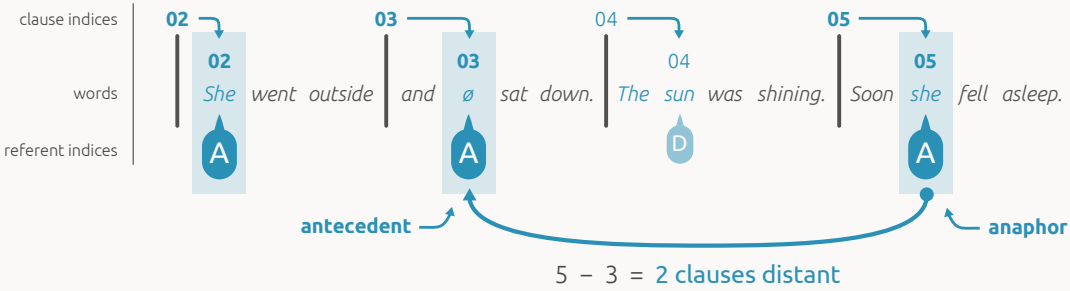
Calculating clause distances



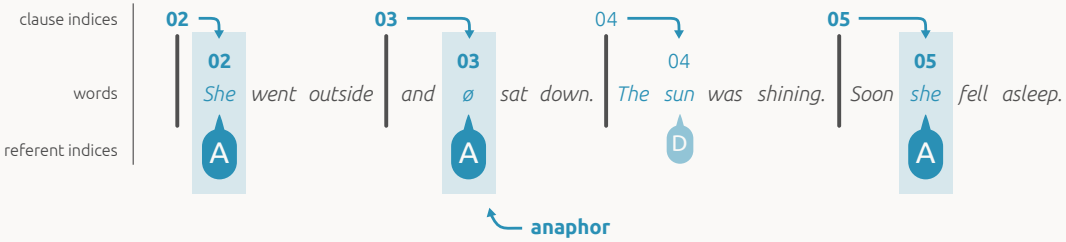
Calculating clause distances



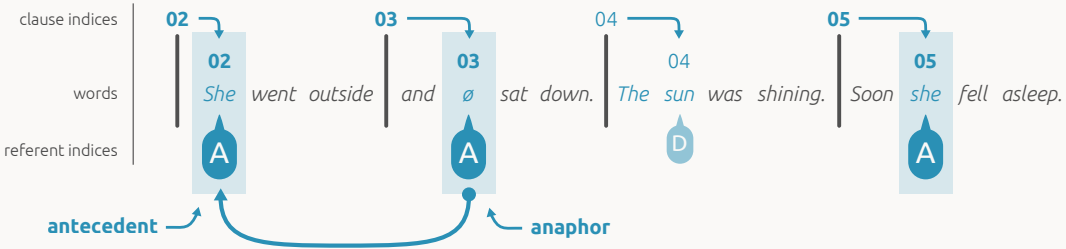
Calculating clause distances



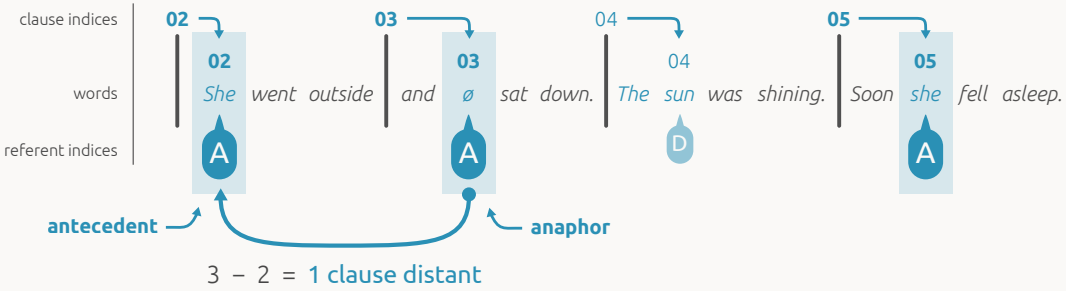
Calculating clause distances



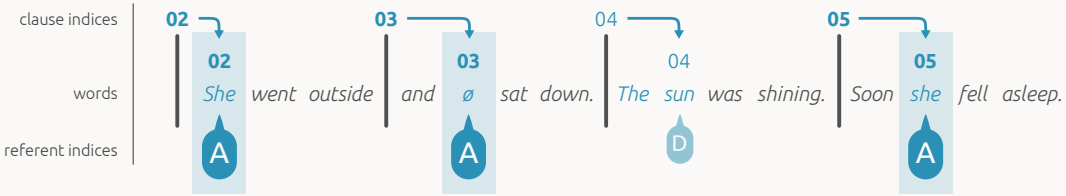
Calculating clause distances



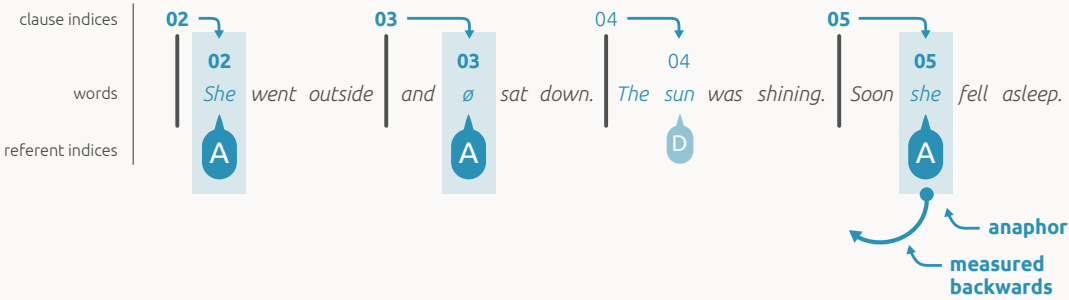
Calculating clause distances



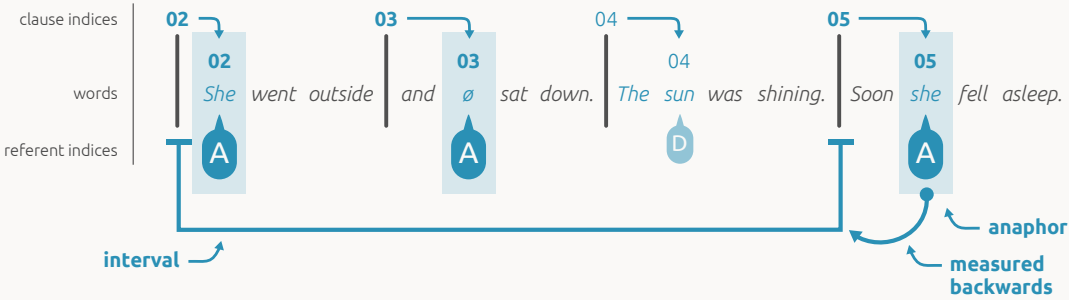
Distance intervals



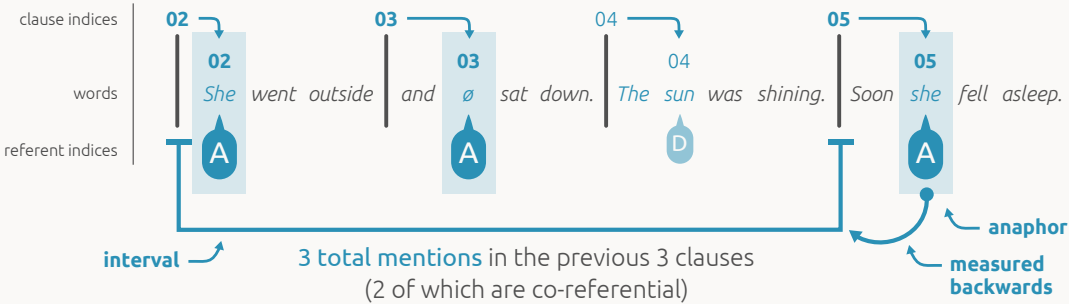
Distance intervals



Distance intervals



Distance intervals



Selection criteria

- only **subjects** (for simplicity's sake)
  (with GRAID functions ⟨:s⟩ 'subject of an intransitive clause',
  ⟨:a⟩ 'subject of a transitive clause', or
  ⟨:ncs⟩ 'non-canonical subject')
- only **pronominal NPs, lexical NPs,** or **zero anaphors**
  (with GRAID forms ⟨pro⟩, ⟨np⟩, ⟨0⟩ or subspecifications thereof)

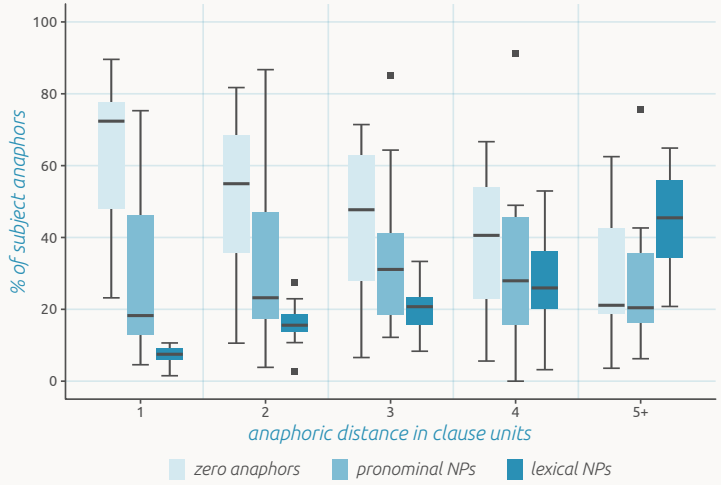
Selection criteria

- only **referring expressions**
  (i.e. no non-referential expressions;
  i.e. only those with a referent index)
- only **anaphoric mentions**
  (i.e. no newly introduced referents, deixis, etc.;
  i.e. only the second and later occurrences of a referent index)

Implementation

- **step –1**:
  download and unzip the *R scripts*
- **step 0**:
  open the "example-B_anaphoric-distance-in-MC.R" script

Results



Proportion of expressions used for subject anaphors in 10 Multi-CAST corpora

Discussion

- **rate of more informative expressions increases linearly with distance**

- substantial **cross-linguistic variation between zero and pronoun rates** (both across distance categories and overall)
- comparatively much **more stable rates of lexical expression** (both across distance categories and overall)

Discussion

- referential choice is primarily about
  the **selection of lexical vs. non-lexical expressions**
  (Kibrik 2011, Schiborr 2021)
- **special significance of the d = 1** (i.e. antecedent in previous clause) **context**,
  especially for subjects: **same-subject chains** (Givón 2017)
  here find lowest rate of lexical NPs and highest rate of zero across languages

Summary

- **multilingual corpora**
- *Universal Dependencies* and *Multi-CAST*
- *NP lengths by position relative to the predicate*
- *anaphoric distance and form of subjects*
- **deriving complex measures from simple annotations**
- **using multiple layers of annotations in conjunction**

Thanks!

References

Ariel, Mira. 1990. *Accessing noun-phrase antecedents.* London: Routledge.

Arnold, Jennifer & Losongco, Anthony & Wasow, Thomas & Ginstrom, Ryan. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1): 28–55.

Barth, Danielle & Evans, Nicholas. 2017. The Social Cognition Paralax Corpus (SCOPIC). *Language Documentation and Conservation* special publication 12.

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11(1): 239–251.

Bickel, Balthasar. 2011. Grammatical relations typology. In Song, Hae Jung (ed.), *The Oxford handbook of linguistic typology*, 399–444. Oxford: Oxford University Press.

Bickel, Balthasar & Zakharko, Taras & Nichols, Johanna. 2016. *Better data with late aggregation: AUTOTYP and beyond.* Paper presented at the 36rd Poznań Linguistic Meeting, Posnań, Poland, 15 September 2016.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, Charles N. (ed.), *Subject and topic*, 25–55. New York: Academic Press.

References

Cysouw, Michael & Wälchli, Bernhard. 2007. Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung (STUF)* 60(2): 95–99.

Dahl, Östen. 2015. *How WEIRD are WALS languages?* Paper presented at the Closing Conference of the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 1–3 May 2015.

Dowle, Matt & Srinivasan, Arun. 2021. *data.table: Extension of 'data.frame'.* R package version 1.14.0. (http://CRAN.R-project.org/package=data.table)

de Marneffe, Marie-Catherine & Manning, Christopher D. 2008. The Stanford typed dependencies representation. *COLING 2008: Proceedings of the workshop on Cross-framework and Cross-domain Parser Evaluation.*

de Marneffe, Marie-Catherine & Dozat, Timothy & Silveira, Natalia & Haverinen, Katri & Ginter, Filip & Nivre, Joakim & Manning, Christopher D. 2014. Universal Stanford Dependencies: A cross-linguistic typology. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014.*

References

Futrell, Richard & Levy, Roger P. & Gibson, Edward. 2020. Dependency locality as an explanation principle for word order. *Language* 96(2): 371–412.

Gerdes, Kim & Kahane, Sylvain & Chen, Xinying 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa* 6(1): 1–31.

Givón, Talmy. 1983. Topic continuity in discourse: An introduction. In Givón, Talmy (ed.), *Topic continuity in discourse*, 1–42. Amsterdam: John Benjamins.

Givón, Talmy. 2017. Zero, pronouns and clause-chaining: Toward a diachronic understanding. *Lingua* 185(1): 96–120.

Haig, Geoffrey & Schnell, Stefan & Wegener, Claudia. 2011. Comparing corpora from endangered languages: Explorations in language typology based on original texts. In Haig, Geoffrey & Nau, Nicole & Schnell, Stefan & Wegener, Claudia (eds.), *Documenting endangered languages: Achievements and perspectives*, 55–86. Berlin: Mouton de Gruyter.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators.* Version 7.0. (https://multicast.aspra.uni-bamberg.de/data/pubs/graid/Haig+Schnell2014_GRAID-manual_v7.0.pdf)

References

Haig, Geoffrey & Schnell, Stefan. 2015. *Multi-CAST: The Multilingual Corpus of Annotated Spoken Texts.* (https://multicast.aspra.uni-bamberg.de/)

Haig, Geoffrey & Schnell, Stefan. 2016. The discourse basis of ergativity revisited. *Language* 92(3): 591–618.

Haig, Geoffrey & Schnell, Stefan. 2021. *Multi-CAST: The Multilingual Corpus of Annotated Spoken Texts.* Version 2101. (https://multicast.aspra.uni-bamberg.de/)

Haig, Geoffrey & Schnell, Stefan & Schiborr, Nils N. To appear. Universals of reference in discourse and grammar: Evidence from the Multi-CAST collection of spoken corpora. To appear in *Language Conservation and Documentation*.

Kibrik, Andrej A. 2011. *Reference in discourse*. Oxford: Oxford University Press.

Koehn, Philipp. 2005. EuroParl: A parallel corpus for statistical machine translation. *MT Summit* 6: 79–86.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Languages in Contrast* 23(3): 533–572.

References

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21–27 May 2012.*

MacWhinney, Brian. 1991. *The CHILDES project: Tools for analyzing talk.* Hillsdale, NJ: Erlbaum.

Mayer, Thomas & Cysouw, Michael. 2014. Creating a massively parallel Bible corpus. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014.*

Mettouchi, Amina & Martine, Vanhove & Caubet, Dominique. 2015. *Corpus-based studies of lesser-described languages: The CorpAfroAs corpus of spoken AfroAsiatic languages.* Amsterdam: John Benjamins.

Paschen, Ludger & Delafontaine, François & Draxler, Christoph & Fuchs, Susanne & Stave, Matthew & Seifart, Frank. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'20), Marseille, France, 13–16 May 2020.*

References

Schiborr, Nils N. 2021. Lexical anaphora: A corpus-based typological study of referential choice. PhD dissertation, University of Bamberg.

Schiborr, Nils N. 2021. *multicastR: A companion to the Multi-CAST collection.* R package version 2.0.0. (https://cran.r-project.org/package=multicastR)

Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2018. RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines. Version 1.1. (https://multicast.aspra.uni-bamberg.de/data/pubs/refind/Schiborr+etal2018_RefIND-guidelines_v1.1.pdf)

Schnell, Stefan & Barth, Danielle. To appear. *Corpus linguistics.*

Schnell, Stefan & Schiborr, Nils N. Submitted. Cross-linguistic corpus studies in linguistic typology. To appear in *Annual Review of Linguistics.*

Schnell, Stefan & Schiborr, Nils N. & Haig, Geoffrey. To appear. Efficiency in discourse processing: Does morphosyntax adapt to accommodate new referents?. To appear in *Linguistic Vanguard.*

Tošović, Branko. 2008. Das Gralis-Korpus. In Tošović, Branko (ed.), *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, 653–692. Graz: LIT.

References

von Waldenfels, Ruprecht & Meyer, Roland. 2006. *ParaSol, a corpus of Slavic and other languages.* (http://parasolcorpus.org/)

Wälchli, Bernard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13: 77–94.

Wasow, T. 1997. Remarks on grammatical weight. *Language Variation and Change* 9(1): 81–105.

Zeman, Daniel & Nivre, Joakim & Abrams, Mitchell & et alii. 2020. *Universal Dependencies.* Version 2.7. Prague: Universal Dependencies Consortium.