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 Slavic languages, CorpAfroAs on Afroasiatic languages.

- Others aim at more general typological representativity, e.g. Universal Dependencies, DoReCo, etc.
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 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)
  - *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
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
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 (?)
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
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)
Jane was reading a new book on linguistics.
Jane was reading a new book on linguistics.
<table>
<thead>
<tr>
<th>words</th>
<th>Jane</th>
<th>was</th>
<th>reading</th>
<th>a</th>
<th>new</th>
<th>book</th>
<th>on</th>
<th>linguistics.</th>
</tr>
</thead>
</table>

Jane was reading a new book on linguistics.
### Dependencies

<table>
<thead>
<tr>
<th>words</th>
<th>part-of-speech tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
</tr>
</tbody>
</table>
Jane was reading a new book on linguistics.

Words:
- Jane
- was
- reading
- a
- new
- book
- on
- linguistics

Part-of-speech tags:
- PROPN
- AUX
- VERB
- DET
- ADJ
- NOUN
- ADP
- NOUN

Syntactic relations:
- nsubj

Relation type:
- nsubj
Jane was reading a new book on linguistics.

- **Words**: Jane, was, reading, a, new, book, on, linguistics.
- **Part-of-speech tags**: PROPN, AUX, VERB, DET, ADJ, NOUN, ADP, NOUN
- **Syntactic relations**: nsubj, obj
Jane was reading a new book on linguistics.

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

Part-of-speech tags:
- Jane: PROPN
- was: AUX
- reading: VERB
- a: DET
- new: ADJ
- book: NOUN
- on: ADP
- linguistics: NOUN

Syntactic relations:
- nsubj: Jane
- obj: book
- det: a
- amod: new
Jane was reading a new book on linguistics.

- **Jane**: PROPN
- **was**: AUX
- **reading**: VERB
- **a**: DET
- **new**: ADJ
- **book**: NOUN
- **on**: ADP
- **linguistics.**: NOUN

**Part-of-speech tags:**
- PROPN
- AUX
- VERB
- DET
- ADJ
- NOUN
- ADP
- NOUN

**Syntactic relations:**
- nsubj
- obj
- det
- amod
- case
Jane was reading a new book on linguistics.
Jane was reading a new book on linguistics.

- **Jane**: PROPN
- **was**: AUX
- **reading**: VERB
- **a**: DET
- **new**: ADJ
- **book**: NOUN
- **on**: ADP
- **linguistics.**: NOUN

**Syntactic relations:***
- nsubj: Jane
- aux: was
- obj: reading
- det: a
- amod: new
- nmod: book
- case: on
- nmod: linguistics.
Dependencies

Jane was reading a new book on linguistics.

words

Jane
was
reading

PROPN
AUX
VERB

part-of-speech tags

a
new
book
on
linguistics.

DET
ADJ
NOUN
ADP
NOUN

syntactic relations

nsubj
aux
obj
det
amod
nmod
case
Dependencies

Jane was reading a new book on linguistics.

<table>
<thead>
<tr>
<th>Word</th>
<th>Part-of-Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
</tr>
</tbody>
</table>

Word index:
1. Jane
2. was
3. reading
4. a
5. new
6. book
7. on
8. linguistics

Syntactic relations:
- nsubj: Jane
- det: a
- amod: new
- nmod: on
- case: linguistics

Part-of-speech tags:
- PROPN
- AUX
- VERB
- DET
- ADJ
- NOUN
- ADP
- NOUN

Numbered sequentially
Jane was reading a new book on linguistics.
Dependences

Jane was reading a new book on linguistics.

PROPN AUX VERB DET ADJ NOUN ADP NOUN
nsubj aux det amod obj case nmod

word index
1 2 3 4 5 6 7 8
3 3 6 6 6 8 3

index of head
root

index of head
0

root

syntactic relations

word index

part-of-speech tags

index of head

a new book on linguistics.
Jane was reading a new book on linguistics.

syntactic relations

words

part-of-speech tags

word index

index of head

nsubj

aux

obj

det

amod

nmod

case

Jane was reading

PROPN AUX VERB

1 2 3

3 3 6

a new book

DET ADJ NOUN

4 5 6

6 6 3

on linguistics.

ADP NOUN

7 8

8 3 6
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)
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)?
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?)
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!
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

<table>
<thead>
<tr>
<th>word index</th>
<th>Jane</th>
<th>PROPN</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>was</td>
<td>AUX</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>reading</td>
<td>VERB</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>DET</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>new</td>
<td>ADJ</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>book</td>
<td>NOUN</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>on</td>
<td>ADP</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>linguistics</td>
<td>NOUN</td>
<td>8</td>
</tr>
</tbody>
</table>
### Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1$nsubj$ 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2$aux$ 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3$root$ 0</td>
</tr>
<tr>
<td>a</td>
<td>DET 4$det$ 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5$amod$ 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6$obj$ 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7$case$ 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8$nmod$ 6</td>
</tr>
</tbody>
</table>
Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>1 — nsubj — 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>2 — aux — 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>3 — root — 0</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>4 — det — 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>5 — amod — 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>6 — obj — 3 x</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>7 — case — 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>8 — nmod — 6</td>
</tr>
</tbody>
</table>
Calculating NP lengths

<table>
<thead>
<tr>
<th>Index</th>
<th>Word</th>
<th>Part of Speech</th>
<th>Index of Head</th>
<th>Is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jane</td>
<td>PROPN</td>
<td>nsubj 3</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>was</td>
<td>AUX</td>
<td>aux 3</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>reading</td>
<td>VERB</td>
<td>root 0</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>DET</td>
<td>det 6</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>new</td>
<td>ADJ</td>
<td>amod 6</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>book</td>
<td>NOUN</td>
<td>obj 3</td>
<td>x</td>
</tr>
<tr>
<td>7</td>
<td>on</td>
<td>ADP</td>
<td>case 8</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>linguistics</td>
<td>NOUN</td>
<td>nmod 6</td>
<td>-</td>
</tr>
</tbody>
</table>
## Calculating NP lengths

<table>
<thead>
<tr>
<th>Word</th>
<th>Part of Speech</th>
<th>Index of Head</th>
<th>Is Part of Argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>1</td>
<td>nsubj 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>2</td>
<td>aux 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>3</td>
<td>root 0</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>4</td>
<td>det 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>5</td>
<td>amod 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>6</td>
<td>obj 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>7</td>
<td>case 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>8</td>
<td>nmod 6</td>
</tr>
</tbody>
</table>

*Note: The cycle 1 indicates the start of the argument NP cycle.*
Calculating NP lengths

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

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>1 nsubj 3</td>
<td>-</td>
</tr>
<tr>
<td>was</td>
<td>2 aux 3</td>
<td>-</td>
</tr>
<tr>
<td>reading</td>
<td>3 root 0</td>
<td>-</td>
</tr>
<tr>
<td>a</td>
<td>4 det 6</td>
<td>-</td>
</tr>
<tr>
<td>new</td>
<td>5 amod 6</td>
<td>-</td>
</tr>
<tr>
<td>book</td>
<td>6 obj 3</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>7 case 8</td>
<td>-</td>
</tr>
<tr>
<td>linguistics</td>
<td>8 nmod 6</td>
<td>-</td>
</tr>
</tbody>
</table>
Calculating NP lengths

Jane was reading a new book on linguistics

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1</td>
<td>nsubj 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2</td>
<td>aux 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3</td>
<td>root 0</td>
</tr>
<tr>
<td>a</td>
<td>DET 4</td>
<td>det 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5</td>
<td>amod 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6</td>
<td>obj 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7</td>
<td>case 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8</td>
<td>nmod 6</td>
</tr>
</tbody>
</table>
Calculating NP lengths

Jane was reading a new book on linguistics

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td></td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td></td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td></td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td></td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td></td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td></td>
</tr>
</tbody>
</table>
Calculating NP lengths

<table>
<thead>
<tr>
<th>Word</th>
<th>Part of Speech</th>
<th>Index</th>
<th>Head Index</th>
<th>Argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>1</td>
<td>3</td>
<td>x</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>6</td>
<td>3</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>8</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nsubj</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>aux</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>root</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>det</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>amod</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>obj</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>case</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>nmod</td>
<td>6</td>
</tr>
</tbody>
</table>

- Jane: PROPN
- was: AUX
- reading: VERB
- a: DET
- new: ADJ
- book: NOUN
- on: ADP
- linguistics: NOUN
Calculating NP lengths

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  —  

is part of argument NP  cycle 1

index of head

word index

x

is part of argument NP

cycle 1
Calculating NP lengths

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  —
book  NOUN  6  —  obj  —  3  x
on  ADP  7  —  case  —  8  —
linguistics  NOUN  8  —  nmod  —  6  —

index of head

word index

is part of argument NP
cycle 1

x is part of argument NP

cycle 1
### Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1 — nsubj 3</td>
<td>-</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2 — aux 3</td>
<td>-</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3 — root 0</td>
<td>-</td>
</tr>
<tr>
<td>a</td>
<td>DET 4 — det 6</td>
<td>-</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5 — amod 6</td>
<td>-</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6 — obj 3</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7 — case 8</td>
<td>-</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8 — nmod 6</td>
<td>-</td>
</tr>
</tbody>
</table>
Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>book</strong></td>
<td><strong>NOUN</strong></td>
<td><strong>cycle 1</strong></td>
</tr>
<tr>
<td><strong>Jane</strong></td>
<td><strong>PROPN</strong></td>
<td></td>
</tr>
<tr>
<td><strong>was</strong></td>
<td><strong>AUX</strong></td>
<td></td>
</tr>
<tr>
<td><strong>reading</strong></td>
<td><strong>VERB</strong></td>
<td></td>
</tr>
<tr>
<td><strong>a</strong></td>
<td><strong>DET</strong></td>
<td></td>
</tr>
<tr>
<td><strong>new</strong></td>
<td><strong>ADJ</strong></td>
<td></td>
</tr>
<tr>
<td><strong>on</strong></td>
<td><strong>ADP</strong></td>
<td></td>
</tr>
<tr>
<td><strong>linguistics</strong></td>
<td><strong>NOUN</strong></td>
<td></td>
</tr>
</tbody>
</table>

- 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
- on (ADP): 7 — case — 8
- linguistics (NOUN): 8 — nmod — 6

*is part of argument NP*
### Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>1 — nsubj — 3</td>
<td>— —</td>
</tr>
<tr>
<td>was</td>
<td>2 — aux — 3</td>
<td>— —</td>
</tr>
<tr>
<td>reading</td>
<td>3 — root — 0</td>
<td>—</td>
</tr>
<tr>
<td>a</td>
<td>4 — det — 6</td>
<td>— x</td>
</tr>
<tr>
<td>new</td>
<td>5 — amod — 6</td>
<td>—</td>
</tr>
<tr>
<td>book</td>
<td>6 — obj — 3</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>7 — case — 8</td>
<td>— —</td>
</tr>
<tr>
<td>linguistics</td>
<td>8 — nmod — 6</td>
<td>—</td>
</tr>
</tbody>
</table>

*Note: The cycle 1 column indicates whether the word is part of the argument NP.*
### Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1 — nsubj 3</td>
<td>x</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2 — aux 3</td>
<td>x</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3 — root 0</td>
<td>x</td>
</tr>
<tr>
<td>a</td>
<td>DET 4 — det 6</td>
<td>x</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5 — amod 6</td>
<td>x</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6 — obj 3</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7 — case 8</td>
<td>x</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8 — nmod 6</td>
<td>x</td>
</tr>
</tbody>
</table>
Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1</td>
<td>nsubj 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2</td>
<td>aux 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3</td>
<td>root 0</td>
</tr>
<tr>
<td>a</td>
<td>DET 4</td>
<td>det 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5</td>
<td>amod 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6</td>
<td>obj 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7</td>
<td>case 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8</td>
<td>nmod 6</td>
</tr>
</tbody>
</table>
### Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1 → nsubj 3</td>
<td>- -</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2 → aux 3</td>
<td>- -</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3 → root 0</td>
<td>- -</td>
</tr>
<tr>
<td>a</td>
<td>DET 4 → det 6</td>
<td>- x</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5 → amod 6</td>
<td>- x</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6 → obj 3</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7 → case 8</td>
<td>- -</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8 → nmod 6</td>
<td>x</td>
</tr>
</tbody>
</table>
Calculating NP lengths

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

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1</td>
<td>nsubj 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2</td>
<td>aux 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3</td>
<td>root 0</td>
</tr>
<tr>
<td>a</td>
<td>DET 4</td>
<td>det 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5</td>
<td>amod 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6</td>
<td>obj 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7</td>
<td>case 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8</td>
<td>nmod 6</td>
</tr>
</tbody>
</table>

- cycle 1
- cycle 2

A checkmark (x) indicates that the word is part of an argument NP.
Calculating NP lengths

<table>
<thead>
<tr>
<th>word</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>1 — nsubj — 3</td>
<td>— —</td>
</tr>
<tr>
<td>was</td>
<td>2 — aux — 3</td>
<td>— —</td>
</tr>
<tr>
<td>reading</td>
<td>3 — root — 0</td>
<td>— —</td>
</tr>
<tr>
<td>a</td>
<td>4 — det — 6</td>
<td>— x</td>
</tr>
<tr>
<td>new</td>
<td>5 — amod — 6</td>
<td>— x</td>
</tr>
<tr>
<td>book</td>
<td>6 — obj — 3</td>
<td>x x</td>
</tr>
<tr>
<td>on</td>
<td>7 — case — 8</td>
<td>— —</td>
</tr>
<tr>
<td>linguistics</td>
<td>8 — nmod — 6</td>
<td>— x</td>
</tr>
</tbody>
</table>
### Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN 1 — nsubj 3</td>
<td>– –</td>
</tr>
<tr>
<td>was</td>
<td>AUX 2 — aux 3</td>
<td>– –</td>
</tr>
<tr>
<td>reading</td>
<td>VERB 3 — root 0</td>
<td>– –</td>
</tr>
<tr>
<td>a</td>
<td>DET 4 — det 6</td>
<td>– x</td>
</tr>
<tr>
<td>new</td>
<td>ADJ 5 — amod 6</td>
<td>– x</td>
</tr>
<tr>
<td>book</td>
<td>NOUN 6 — obj 3</td>
<td>x x</td>
</tr>
<tr>
<td>on</td>
<td>ADP 7 — case 8</td>
<td>– –</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN 8 — nmod 6</td>
<td>– x</td>
</tr>
</tbody>
</table>
Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>1 — nsubj — 3</td>
<td>— —</td>
</tr>
<tr>
<td></td>
<td>2 — aux — 3</td>
<td>— —</td>
</tr>
<tr>
<td>was</td>
<td>3 — root — 0</td>
<td>— —</td>
</tr>
<tr>
<td></td>
<td>4 — det — 6</td>
<td>— x</td>
</tr>
<tr>
<td>reading</td>
<td>5 — amod — 6</td>
<td>— x</td>
</tr>
<tr>
<td>book</td>
<td>6 — obj — 3</td>
<td>x x</td>
</tr>
<tr>
<td></td>
<td>7 — case — 8</td>
<td>— —</td>
</tr>
<tr>
<td>new</td>
<td>8 — nmod — 6</td>
<td>— x</td>
</tr>
</tbody>
</table>

For the word "book", the index of head is 6, and it is part of argument NP cycle 1 and cycle 2.
Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>1 — nsubj — 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>2 — aux — 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>3 — root — 0</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>4 — det — 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>5 — amod — 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>6 — obj — 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>7 — case — 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>8 — nmod — 6</td>
</tr>
</tbody>
</table>
## Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>-</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>-</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>-</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>-</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>-</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>x</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>-</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>cycle 1</th>
<th>cycle 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>book</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>linguistics</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

*Note:* A cross (x) indicates that the position is part of the argument NP cycle.
Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>nsubj 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>aux 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>root 0</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>det 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>amod 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>obj 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>case 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>nmod 6</td>
</tr>
</tbody>
</table>
### Calculating NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cycle 1</td>
</tr>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>1 — nsubj — 3</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>2 — aux — 3</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>3 — root — 0</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>4 — det — 6</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>5 — amod — 6</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>6 — obj — 3</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>7 — case — 8</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>8 — nmod — 6</td>
</tr>
</tbody>
</table>

5 words
### Calculating more NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nsbj</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>aux</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>root</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>det</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>amod</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>obj</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>case</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>nmod</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>obj</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>det</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>nsbj</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>aux</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>acl</td>
<td>6</td>
</tr>
</tbody>
</table>
Calculating more NP lengths

| word index | index of head | is part of argument NP
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>---</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>---</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>---</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>x</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>book</th>
<th>NOUN</th>
<th>x</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>on</td>
<td>ADP</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>that</td>
<td>PRON</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>her</td>
<td>DET</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>professor</td>
<td>NOUN</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>had</td>
<td>AUX</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>recommended</td>
<td>VERB</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
Calculating more NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>that</td>
<td>PRON</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>her</td>
<td>DET</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>professor</td>
<td>NOUN</td>
<td>cycle 1 cycle 2</td>
</tr>
<tr>
<td>had</td>
<td>AUX</td>
<td>cycle 1 cycle 2</td>
</tr>
</tbody>
</table>

- 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
- 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

Table showing the index of head for each word along with its syntactic role and whether it is part of an argument NP.
Calculating more NP lengths

<table>
<thead>
<tr>
<th>word index</th>
<th>index of head</th>
<th>is part of argument NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>PROPN</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>1 — nsubj 3</td>
<td>- - -</td>
</tr>
<tr>
<td>was</td>
<td>AUX</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>2 — aux 3</td>
<td>- - -</td>
</tr>
<tr>
<td>reading</td>
<td>VERB</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>3 — root 0</td>
<td>- - -</td>
</tr>
<tr>
<td>a</td>
<td>DET</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>4 — det 6</td>
<td>x x x</td>
</tr>
<tr>
<td>new</td>
<td>ADJ</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>5 — amod 6</td>
<td>x x x</td>
</tr>
<tr>
<td>book</td>
<td>NOUN</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>6 — obj 3</td>
<td>x x x</td>
</tr>
<tr>
<td>on</td>
<td>ADP</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>7 — case 8</td>
<td>- - x</td>
</tr>
<tr>
<td>linguistics</td>
<td>NOUN</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>8 — nmod 6</td>
<td>- x x</td>
</tr>
<tr>
<td>that</td>
<td>PRON</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>9 — obj 13</td>
<td>- - x</td>
</tr>
<tr>
<td>her</td>
<td>DET</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>10 — det 11</td>
<td>- - -</td>
</tr>
<tr>
<td>professor</td>
<td>NOUN</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>11 — nsubj 13</td>
<td>- - x</td>
</tr>
<tr>
<td>had</td>
<td>AUX</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>12 — aux 13</td>
<td>- - x</td>
</tr>
<tr>
<td>recommended</td>
<td>VERB</td>
<td>cycle 1 cycle 2 cycle 3</td>
</tr>
<tr>
<td></td>
<td>13 — acl 6</td>
<td>x x x</td>
</tr>
</tbody>
</table>
### Calculating more NP lengths

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

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
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
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 noticeable 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 noticeable 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. \langle pro \rangle ‘free definite pronoun’ to \langle dem_pro \rangle ‘demonstrative used as pronoun’)

An essential aspect of the annotations for our purposes:

- Mark the **syntactic function of expressions**, where
- 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

words

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

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

referring expression

subject, pronoun, human...
Calculating word distances

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

words

referent indices

A

B

A

C

referent index
Calculating word distances

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

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

<table>
<thead>
<tr>
<th>word indices</th>
<th>00 01 02 03 04 05 06 07 08 09 10 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>Then Jane looked out the window. She could see the blue sky.</td>
</tr>
<tr>
<td>referent indices</td>
<td>A       B       A       C</td>
</tr>
</tbody>
</table>
Calculating word distances

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

| word indices | 00 Then | 01 Jane | 02 looked | 03 out | 04 the | 05 window. | 06 She | 07 could | 08 see | 09 the | 10 blue | 11 sky. |
| words         |        |         |          |       |       |           |       |          |       |       |        |        |
| referent indices | A      |          |          |       |       |           |       |          |       |       |        |        |
Calculating word distances

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

Word indices:
- 00: Then
- 01: Jane
- 02: looked
- 03: out
- 04: the
- 05: window.
- 06: She
- 07: could
- 08: see
- 09: the
- 10: blue
- 11: sky.

Referent indices:
- A: from here to here
- B: from here
- C: to here

6 - 1 = 5 words distant
Calculating word distances

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

Then

00 01 02 03 04 05 06 07 08 09 10 11

words

word indices

referent indices

from here
to here

(anaphor) (antecedent)

6 − 1 = 5 words distant

01 Jane

02 looked

03 out

04 the

05 window.

06 She

07 could

08 see

09 the

10 blue

11 sky.
Calculating word distances

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

- **Word indices**: 00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11
- **Words**: Then, Jane, looked, out, the, window, She, could, see, the, blue, sky
- **Referent indices**: A, B, C

**To here (antecedent)**: 6 - 1 = 5 words distant

**From here (anaphor)**: 6 - 1 = 5 words distant
Calculating word distances

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

6 - 1 = 5 words distant

(word indices)

(antecedent)

(to here)

(from here)

(anaphor)

(referent indices)
Calculating clause distances

She went outside and sat down. The sun was shining. Soon she fell asleep.
Calculating clause distances

She went outside and sat down. The sun was shining. Soon she fell asleep.

zero anaphor

referent index

words

referent indices

She A went outside and Ø sat down. The sun was shining. Soon she A fell asleep.
Calculating clause distances

She went outside and sat down. The sun was shining. Soon she fell asleep.

words

referent indices

clause boundary
Calculating clause distances

She went outside and sat down. The sun was shining. Soon fell asleep.

clause index

clause indices

words

referent indices

02

She

A

A

03

went outside

and

ø

sat down.

04

The

sun

was

shining.

05

Soon

she

fell asleep.

A

D

A
Calculating clause distances

She went outside and sat down. The sun was shining. Soon fell asleep.

words
She went outside and sat down. The sun was shining. Soon fell asleep.

transferred index

clause indices

referent indices

A
Calculating clause distances

She went outside and sat down. The sun was shining. Soon fell asleep.

words

referent indices

clause indices

02 02
She
A

03 03
ø
A

04 04
The
D

05 05
she
A

anaphor

5 - 3 = 2 clauses distant
Calculating clause distances

She went outside and sat down. The sun was shining. Soon she fell asleep.
Calculating clause distances

She went outside and sat down. The sun was shining. Soon fell asleep.

5 − 3 = 2 clauses distant
Calculating clause distances

She went outside and sat down. The sun was shining. Soon fell asleep.

<table>
<thead>
<tr>
<th>clause indices</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>She</td>
<td>ø</td>
<td>The</td>
<td>she</td>
</tr>
<tr>
<td>referent indices</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>A</td>
</tr>
</tbody>
</table>

anaphor

3 − 2 = 1 clause distant
Calculating clause distances

```
She went outside and sat down. The sun was shining. Soon fell asleep.
```

<table>
<thead>
<tr>
<th>clause indices</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>She</td>
<td>⌀</td>
<td>The</td>
<td>she</td>
</tr>
<tr>
<td>referent indices</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>A</td>
</tr>
</tbody>
</table>

antecedent

anaphor

clause distances:

3 - 2 = 1 clause distant
Calculating clause distances

She went outside and sat down. The sun was shining. Soon fell asleep.  

3 - 2 = 1 clause distant
Distance intervals

She went outside and sat down. The sun was shining. Soon she fell asleep.

3 total mentions in the previous 3 clauses (2 of which are co-referential)
Distance intervals

She went outside and sat down. The sun was shining. Soon fell asleep.

3 total mentions in the previous 3 clauses (2 of which are co-referential)

Anaphor measured backwards
She went outside and sat down. The sun was shining. Soon fell asleep.

Distance intervals

02: She went outside
03: and sat down.
04: The sun was shining.
05: she fell asleep.

Words: She, and, the, soon, she
Referent indices: A
Clause indices: 02, 03, 04, 05

3 total mentions in the previous 3 clauses (2 of which are co-referential)

Interval measured backwards
Anaphor
She went outside and sat down. The sun was shining. Soon fell asleep.

3 total mentions in the previous 3 clauses (2 of which are co-referential)
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
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


References


References


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


References


