

Validating and Merging a Growing Multilayer Corpus - the Case of GUM

Siyao Peng & Amir Zeldes
 Georgetown University
sp1184@georgetown.edu
amir.zeldes@georgetown.edu



AACL2018
 Atlanta, GA
 2018-09-21

Plan

1

1. Introduction: what is GUM?
2. Multilayer Annotation
3. Validation: avoiding errors
4. Merging: catching errors
5. Outlook

Georgetown University Multilayer corpus (GUM)

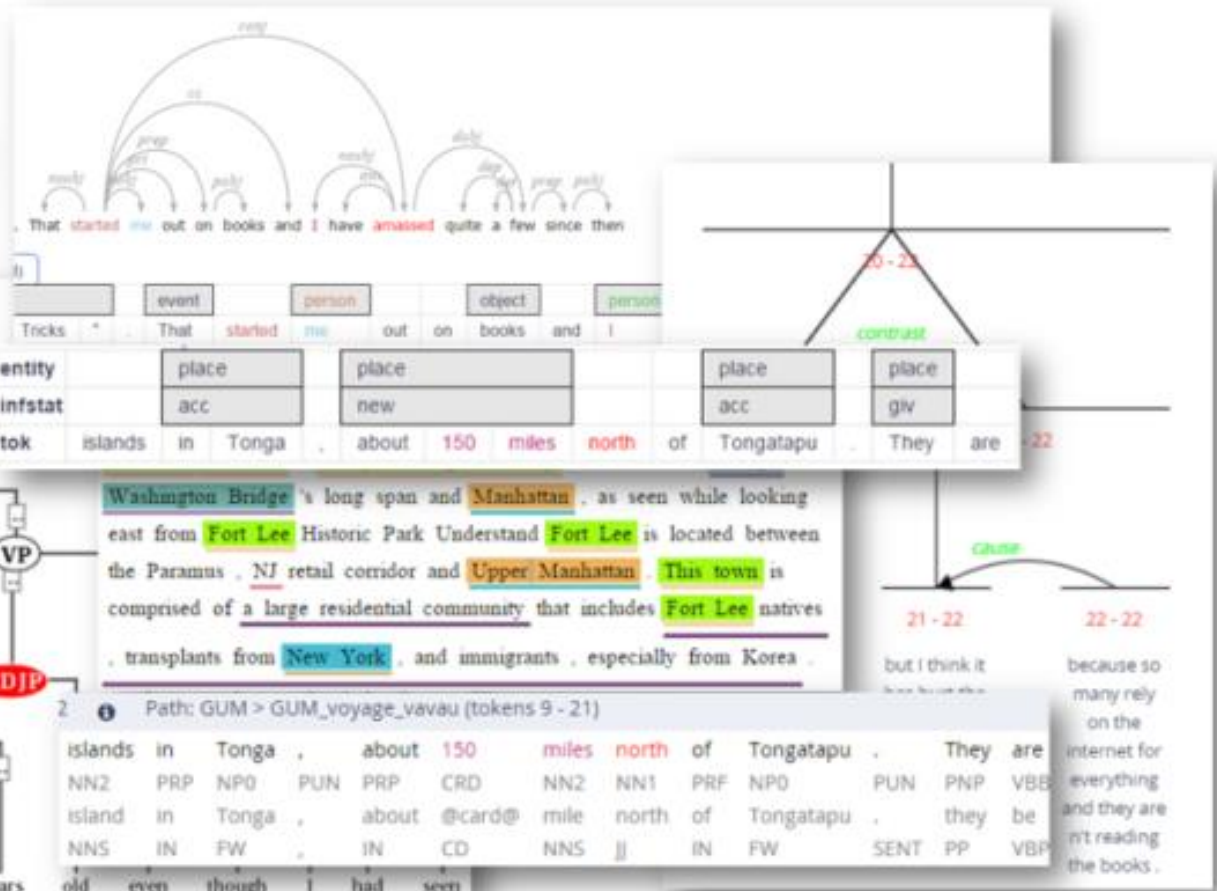
- Open-source multilayer corpus (Zeldes 2017)
- Created by students at Georgetown
- Creative Commons licenses
- Versioned on Github:
 - <https://github.com/amir-zeldes/gum>
- Guidelines online:
 - <https://corpling.uis.georgetown.edu/gum/>

Georgetown University Multilayer corpus (GUM)

- The most recent GUM-v4 consists of:
 - 8 genres: (~100 documents, 85K tokens)
 - 4 existing: *news, interviews, how-to & travel guides*
 - 4 expanding: *academic, bios, fiction & forums*
 - Growth: ~25 docs/20K tokens per year

Annotations include:

- Token annotations (POS, lemmas)
- Text encoding initiatives (TEI tags, cf. Burnard & Bauman 2008)
- Dependency trees
- Discourse referents and coreference
- Rhetorical structure theory (RST, cf. Mann & Thompson 1988)



Philosophy of annotation

- GUM has been richly annotated
 - Different annotation → different research purpose
- Choosing the right tool for each layer of annotation
 - ease of annotation & validation
- Mantra of annotation:
 - A consistent, complete but flawed annotation guideline is better than an idealized but inconsistent one
- Value of multilayer corpora:
 - Challenging to manage multiple layers
 - But easier to find errors and inconsistencies across layers

Classroom annotation

- Course: LING367, students build guidelines in Wiki:
 - <https://corpling.uis.georgetown.edu/wiki/>
- Annotation in multiple Interfaces
- Validation to avoid errors
- TA and instructor do QA
- Use ANNIS search engine to compare previous practices:
 - <http://corpus-tools.org/annis/> (Krause & Zeldes 2014)
 - Search and visualization architecture for multilayer corpora
 - Powerful and flexible search queries

Metadata & structure markup

- Annotated using GitDox (Zhang & Zeldes 2017):
 - <https://corpling.uis.georgetown.edu/gitdox/>
- Learn XML markup:
 - Encode interesting features of data `<head>Wikipedia</head>`
 - Use TEI vocabulary, e.g. `<hi rend="bold">NOTE</hi>`
- Rough speech acts (based on SPAAC, Leech et al. 2003)
 - E.g. *decl, imp, inf, wh, intj*, etc.
- Validation of annotation vocabulary:
 - If students forgot to assign a type to a sentence...
 - If they assign invalid annotation values...


```

1 <text id="GUM_bio_jespersen">
2
3 <head><s type="frag">Otto Jespersen</s></head>
4
5 <p><s type="decl"><hi rend="bold">Jens Otto Harry Jespersen</hi> or <hi rend="bold">Otto Jespersen</hi>
6 (Danish: [ʌtˢoˈjɛsbɐsn̩]; 16 July 1860 – 30 April 1943) was a Danish linguist who specialized in the
7 grammar of the English language.</s></p>
8
9 <head><s type="frag">Early life</s></head>

```

☐ TEI markup (grid)[illegible]

degree in <ref>French</ref>, with English and <ref>Latin</ref> as his secondary languages. </s><s type="decl"> He supported himself during his studies through part-time work as a schoolteacher and as a

Metadata

9

Info for salt:/GUM/GUM_bio_jespersen

Metadata

document: GUM_bio_jespersen

Name	Value
author	Wikipedia, The Free Encyclopedia
dateCollected	2017-09-13
dateCreated	2001-12-19
dateModified	2017-07-10
id	GUM_bio_jespersen
shortTitle	jespersen
sourceURL	https://en.wikipedia.org/wiki/Otto_Jespersen
speakerCount	0
speakerList	none
title	Otto Jespersen
type	bio
annis:doc	GUM_bio_jespersen

GitDox validation for metadata and XML

■ XML validation using XSD

Editor | [back to document list](#)


Document Name:	<input type="text" value="GUM_bio_test"/>		<div>Validate</div> <p>XML schema: 5: element r: Schemas validity error : Element 'r': This element is not expected. Expected add, figure, hi, sp, q, w, quote, ref, date, gap).XML schema fails to validate Metadata for type does not match pattern ... No metadata for speakerList</p>
Corpus Name:	<input type="text" value="GUM"/>		
Git Repo:	<input type="text" value="account/repo_name"/>		
XML Schema:	<input type="text" value="gum_schema"/>	▼	
Assigned to:	<input type="text" value="sp1184"/>	▼	
Status:	<input type="text" value="markup"/>	▼	
Mode:	<input type="text" value="xml"/>	▼	

```

1 <text id="GUM_bio_test">
2 <p>
3 <s type="decl"> This
4 is
5 <r>
6 an
7 </r>
8 example
9 </s>
10 </p>
11 </text>
  
```

Spreadsheet mode validation

- Second step – annotation spans in spreadsheet view:
 - Limit spans: <figure> can have <caption> but not <head>
 - Every word should be in some <s>
 - <head> spans must properly nest <s> spans
 - No nesting of <s> inside of <s>...

	A	B	C	D	E	F
1	tok	pos	lemma	text_id	head	s_type
38	English	JJ	English	Span break on line 40 in column head but not s_type		
39	language	NN	language			
40	.	SENT	.			
41	Early	JJ	early		head	frag
42	life	NN	life			

Tokenization & POS tagging

- Tokenization:
 - Human corrections (PTB guidelines) after auto-tokenization
 - Annotated using GitDox:
<https://corpling.uis.georgetown.edu/gitdox/>
- POS tagging:
 - Annotated from scratch
 - Use extended TreeTagger version (Schmid 1994) of PTB tagset (Santorini 1990)

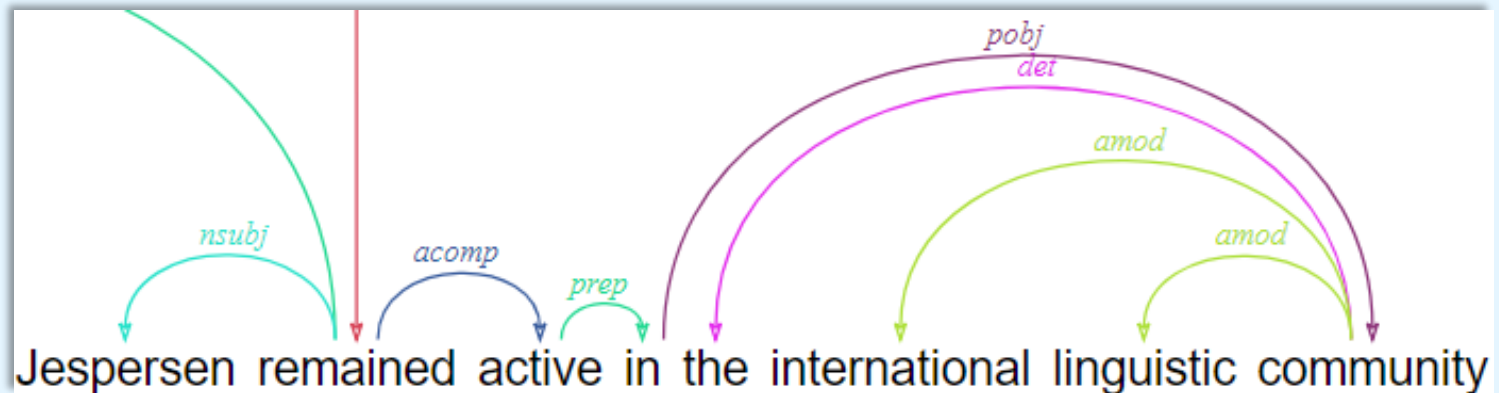
Tokenization & POS tagging

After his retirement in 1925, Jespersen remained active in the international linguistic community.

TEI markup (grid)														
						date								
						1925								
p														
note														
cl														
decl														
.	"	After	his	retirement	in	1925	,	Jespersen	remained	active	in	the	international	linguistic community .
path: GUM > GUM_bio_jespersen (tokens 841 - 851) left context: 15 ▼ right context: 5														
After	his	retirement	in	1925	,	Jespersen	remained	active	in	the	international	linguistic	community	.
IN	PRP\$	NN	IN	CD	,	NNP	VBD	JJ	IN	DT	JJ	JJ	NN	.
PRP	DPS	NN1	PRP	CRD	PUN	NPO	VVD	AJ0	PRP	ATO	AJ0	AJ0	NN1	PUN
prep	poss	pobj	prep	pobj	punct	nsubj	root	acomp	prep	det	amod	amod	pobj	punct
after	his	retirement	in	@card@	,	Jespersen	remain	active	in	the	international	linguistic	community	.
IN	PP\$	NN	IN	CD	,	NP	VVD	JJ	IN	DT	JJ	JJ	NN	SENT

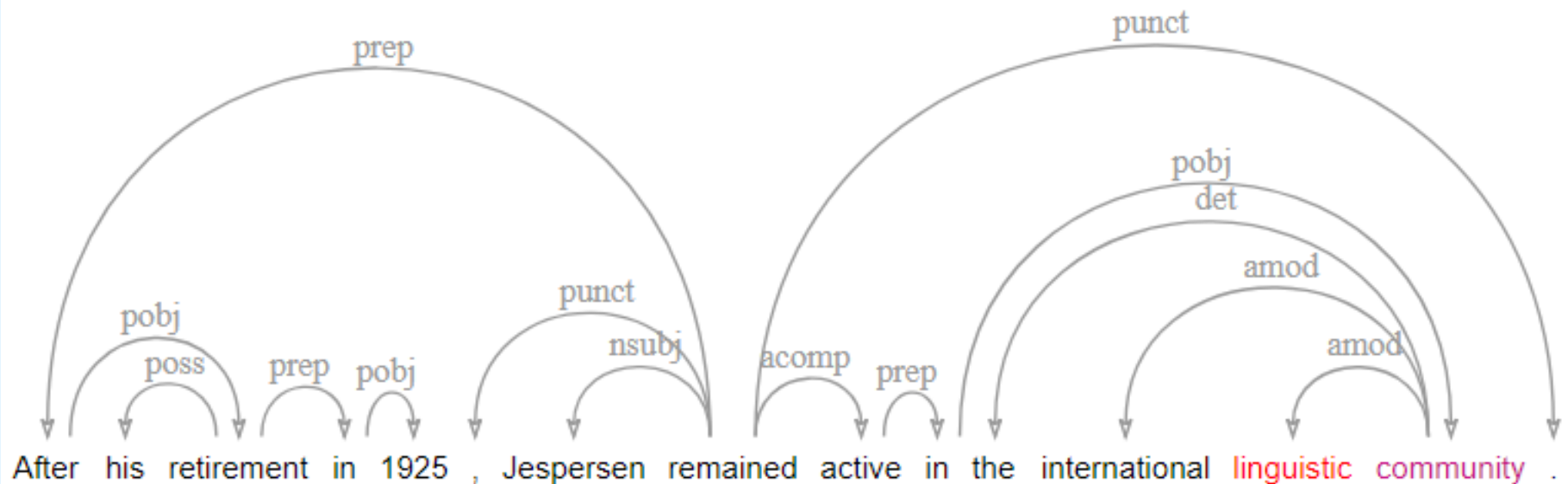
Syntactic annotation - dependencies

- Annotated using Arborator (Gerdes 2013):
 - <http://corpling.uis.georgetown.edu/arborator/>
- Correct auto-parsed Stanford Dependencies
 - Stable scheme for English based on PTB tags
 - Lexical head dependencies
 - Automatic conversion to Universal Dependencies (UD)



Syntactic annotation - SD

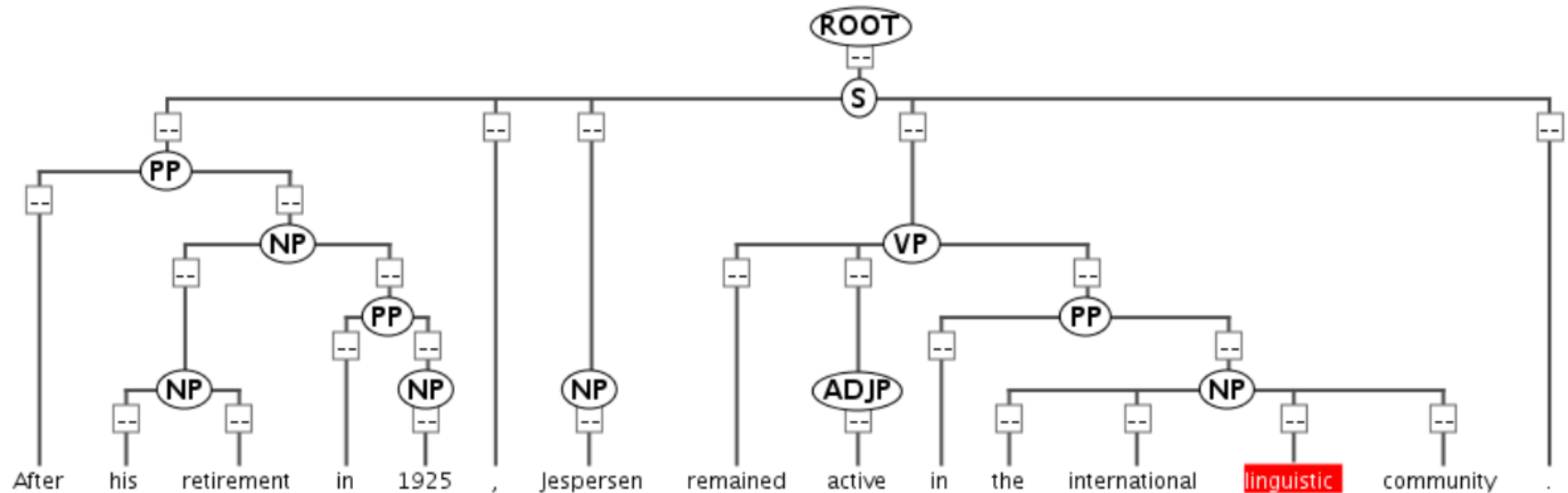
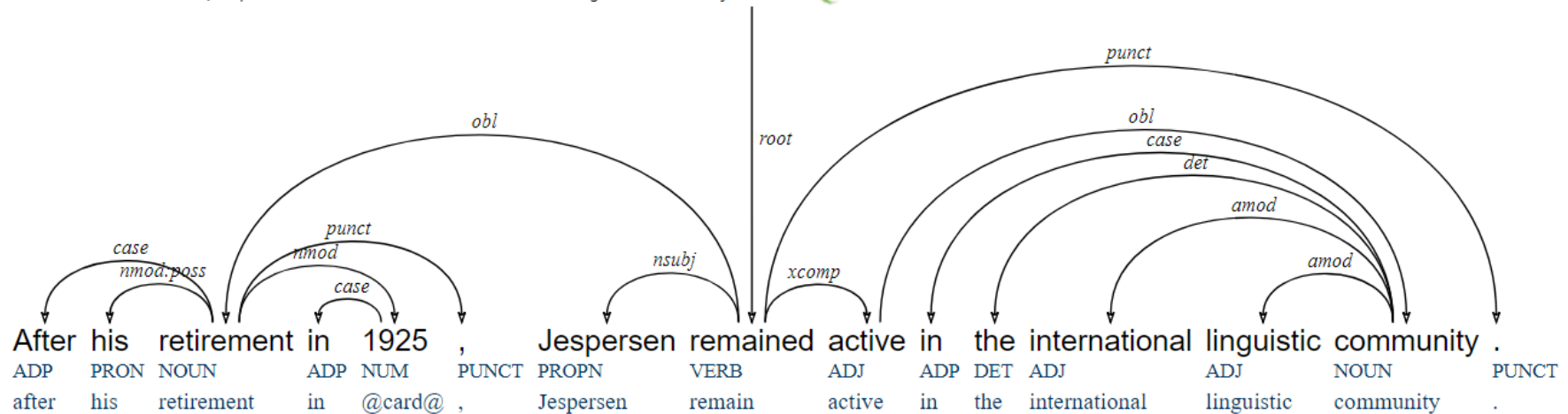
- Searchable in ANNIS – students can look up similar cases



Syntactic conversions - UD & Const

16

0: After his retirement in 1925, Jespersen remained active in the international linguistic community.



IS - entities and coreference

- Annotated using WebAnno (Yimam et al. 2013):
 - <https://webanno.github.io/>
- Manual corrections of *xrenner* output:
 - <https://corpling.uis.georgetown.edu/xrenner/>
- Entity types, e.g. ***place, organization, person***, etc.
 - Entity markables can be nested
 - Each entity could be *new, given, or accessible*
- Coreference types, e.g. ***ana, cata, bridge***, etc.
 - Coreference produces chains: ***William Evans*** <- ***President*** <- ***he***
 - Bridging anaphora: ***the house ...*** <- ***the door (=of the house)***

Correcting entities and coreference

18



IS - entities and coreference

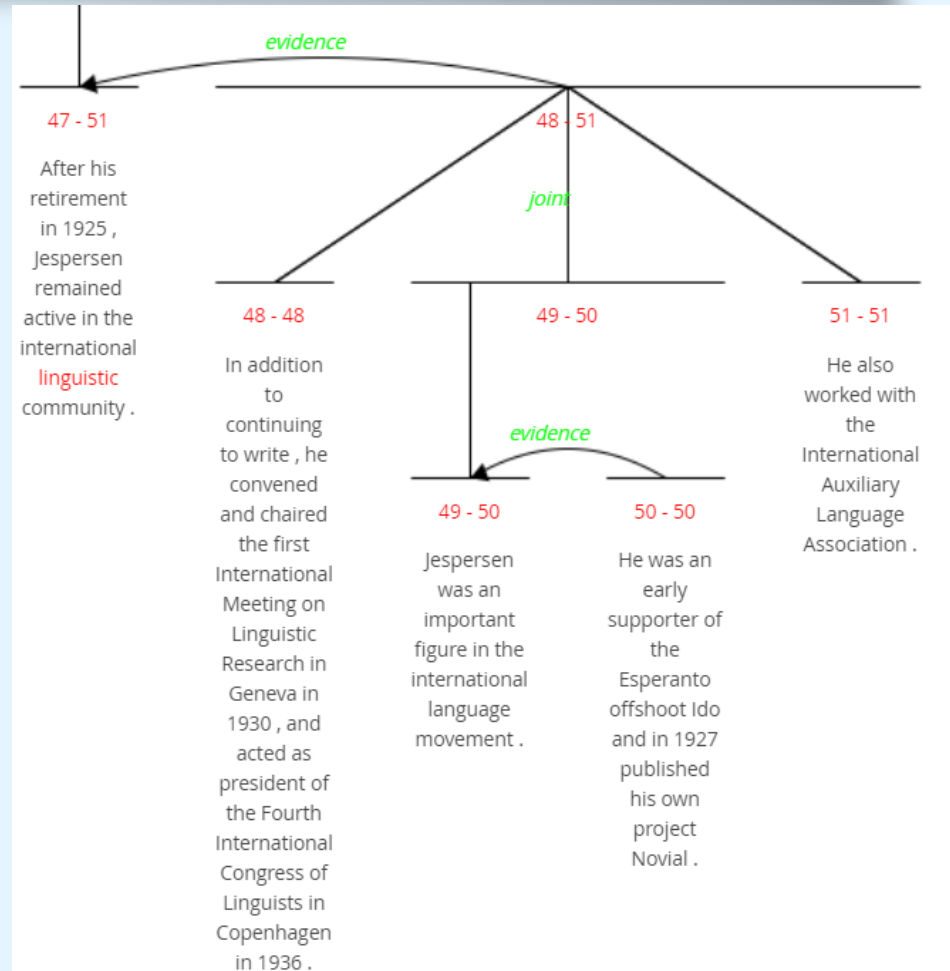
☐ referents (grid)

entity					event								organization				
entity					person			time		person							
infstat					new								new				
infstat					giv			giv		giv							
tok	for	.	"	After	his	retirement	in	1925	,	Jespersen	remained	active	in	the	international	linguistic community	.

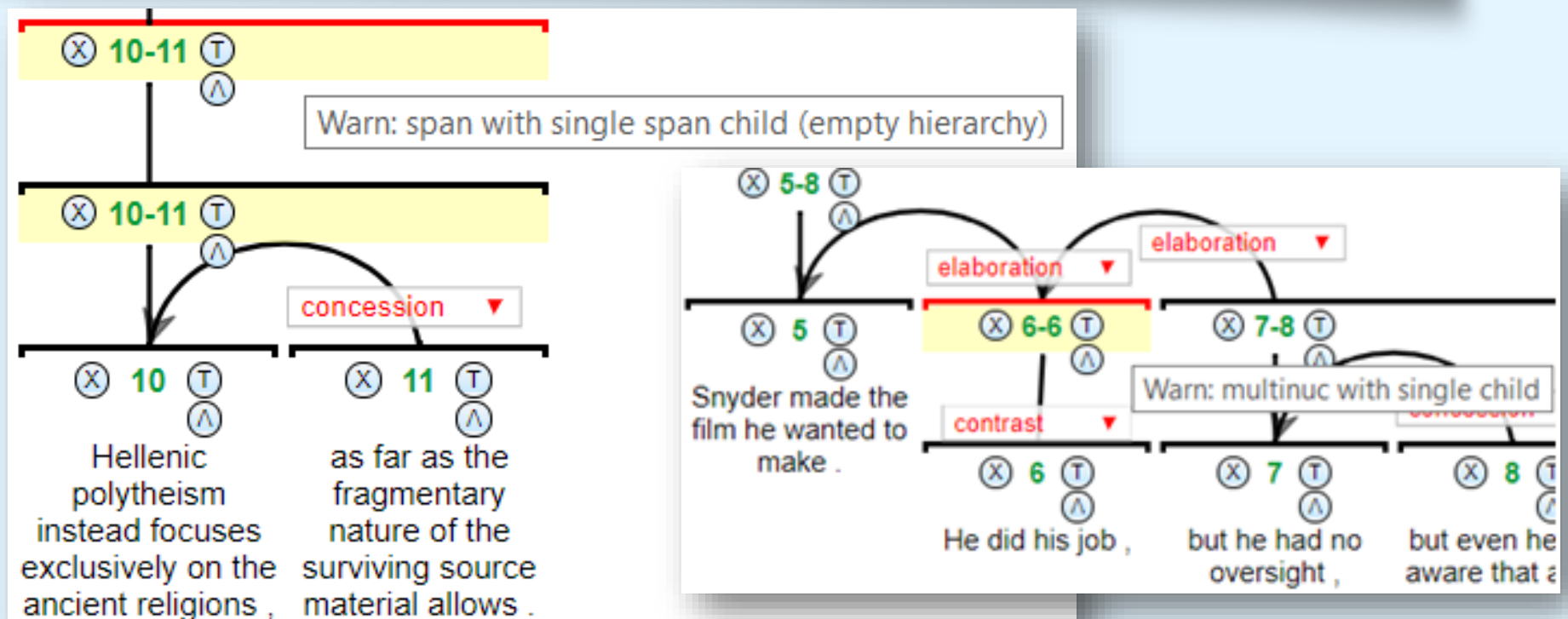
ideas they stand for . " After his retirement in 1925 , Jespersen remained active in the international linguistic community . In addition to continuing to write , he convened and chaired the first International Meeting on Linguistic Research in Geneva in 1930 , and acted as president of the Fourth International Congress of Linguists in Copenhagen in 1936 . Jespersen was

Rhetorical Structure Theory (RST)

- Annotated using rstWeb:
 - <https://corpling.uis.georgetown.edu/rstweb>
- Elementary Discourse Units (EDUs) are related via rhetorical relations, e.g. *evidence*, *explanation*, etc.
- “Tree of clauses”



Validation for RST



- No non-multinuclear with multiple non-span children
- No multinuclear span with single child
- No empty hierarchy

GUM Build bot

```
=====
Validating files...
=====

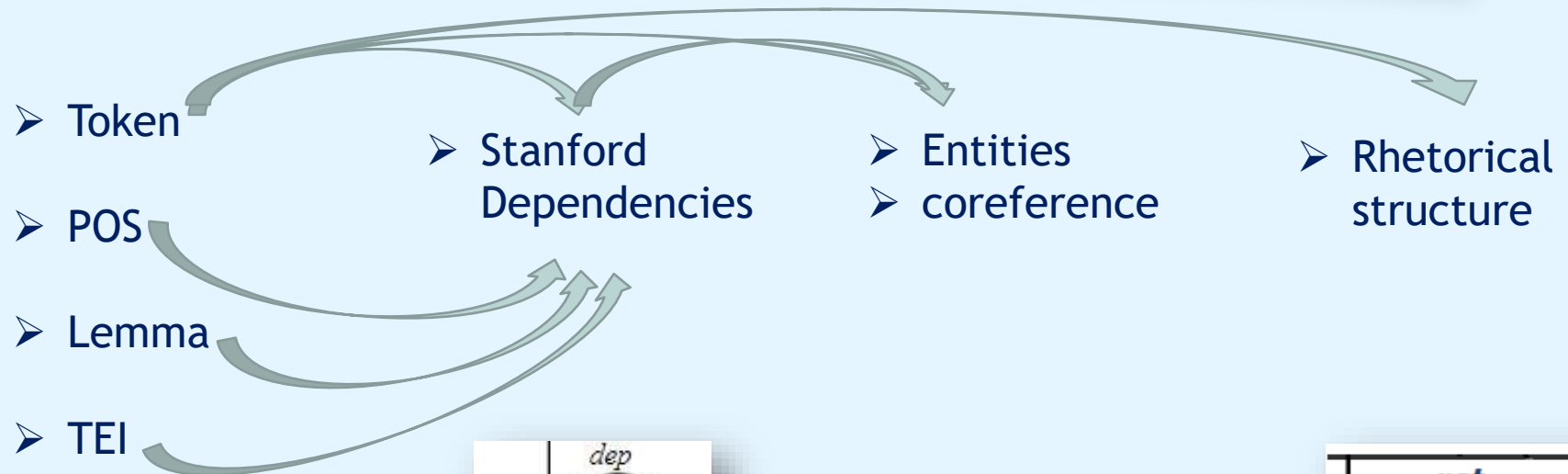
o Found 101 documents
o File names match
o Token counts match across directories
o 101 documents pass XSD validation

WARN: back-pointing mwe in *discrimination.xml @ token 675 (more <- than)
WARN: new markable has antecedent in *discrimination.xml:50-7=abstract (sample)
WARN: coref clash in *thrones.xml:18-6=object -> 16-4=abstract (books->book)
WARN: unlisted mwe in *nida.xml @ token 510 (in -> opposition)
WARN: frag root may not have nsubj in *nida.xml @ token 757 (ROOT -> aim)
```

GUM Build bot

- Github is great for version control and collaboration 😊
- GUM (<https://github.com/amir-zeldes/gum>) separates edit files for different annotations into four directories:
 - `_build/src/xml`, `_build/src/dep`, `_build/src/tsv`, `_build/src/rst`
- The GUM Build Bot merges sources using SaltNPepper (Zipser & Romary 2010)
- We will discuss three functionalities:
 - ***Propagation***
 - ***Conversion***
 - ***Validation/Error-catching***

Propagation & Conversion



`<w>xml</w>`

\downarrow *dep*
 GUM dep
 N N

tsv object[0] giv[0]

rst
 Web

✓ Vanilla PTB
 ✓ Constituents

✓ Universal
 Dependencies

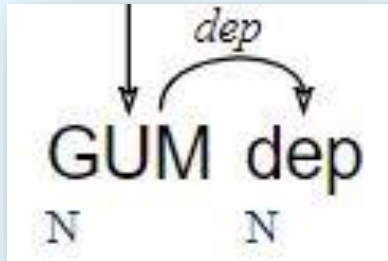
Within-directory validation - /xml

`<w>xml</w>`

- Token
- POS
- Lemma
- TEI

Validation Error	#
Pos=/POS/ must have lemma=/s/	70
Non-ASCII characters for punctuations	36
Pos=/VB*/ must have lemma=/be/	8
Pos=/VH*/ must have lemma=/have/	5
Pos=/VV*/ must not have lemma=/be/	1
(TEI) Sent_type=/intj/ must have pos=/UH/	NA
(TEI) Sent_type=/imp/ cannot have pos=/VVP/	NA

Within-directory validation - */dep* & */tsv*



➤ Stanford Dependencies

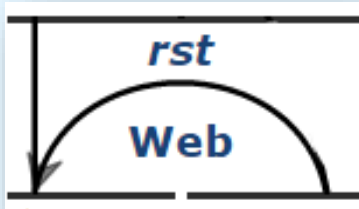
<i>Validation Error</i>	<i>#</i>
Unlisted <i>mwe</i>	66
Back-pointing func <i>mwe</i>	60
Back-pointing func <i>conj</i>	8
No cyclic dependency chains	NA
No invalid dependency relation label	NA

`tsv object[0] giv[0]`

➤ Entities ➤ coreference

<i>Validation Error</i>	<i>#</i>
No coreference clash	35
New markable should not have antecedent	12

Within-directory validation - /rst



➤ Rhetorical structure

<i>Validation Error</i>	<i>#</i>
Non-multinuclear with multiple non-span children	161
Span with single span child	12

Cross-directory validation - */xml+/dep*

`<w>xml</w>`

- Token
- POS
- Lemma
- TEI



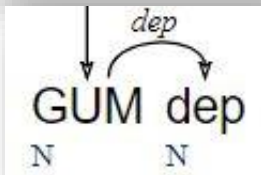
- SD

Validation Error

Validation Error	#
Root of s_type= <i>/frag/</i> may not have func= <i>/nsubj/</i>	14
Root of s_type= <i>/imp/</i> may not have func= <i>/nsubj/</i>	9
Pos= <i>/POS/</i> token must have func= <i>/possessive/</i>	7
Func= <i>/aux/</i> must be lemma= <i>/(be) (have) (do)/</i>	6
Func= <i>/possessive/</i> token must have pos= <i>/POS/</i>	3
Func= <i>/auxpass/</i> must be lemma= <i>/(be) (get)/</i>	2

Cross-directory validation - *all dirs*

<w>xml</w>



tsv object[0] giv[0]



Validation Error

	#
Same doc should have same sentence lengths	3
Dirs should have same number and names of files	NA
Same doc should have same token & sent counts	NA

- UD conversion of GUM4 is available (Peng & Zeldes 2018):
 - https://github.com/UniversalDependencies/UD_English-GUM
- GUM continues to grow – look for version 5 in winter!

Thank you!

References (1/2)

- Burnard, L., & Bauman, S. (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Technical report. Available at: <http://www.tei-c.org/Guidelines/P5/>.
- Gerdes, K. (2013). Collaborative Dependency Annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*. Prague, 88–97.
- Krause, T., & Zeldes, A. (2014). ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities*. Available at: <http://dsh.oxfordjournals.org/content/digitalsh/early/2014/12/02/llc.fqu057.full.pdf>.
- Leech, G., McEnery, T. & Weisser, M. (2003). SPAAC Speech-Act Annotation Scheme. Lancaster University, Technical Report, Lancaster University. Available at: <http://ucrel.lancs.ac.uk/SPAAC/>
- Mann, W. C. & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3), 243–281.
- Peng, S. and Zeldes, A. (2018). All Roads Lead to UD: Converting Stanford and Penn Parses to English Universal Dependencies with Multilayer Annotations. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018) at COLING2018*. Santa Fe, NM, 167-177.

References (2/2)

- Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). University of Pennsylvania, Technical Report, University of Pennsylvania.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK, 44–49.
- Yimam, S. M., Gurevych, I., Castilho, R. Eckart de, & Biemann, C. (2013). WebAnno: A Flexible, Webbased and Visually Supported System for Distributed Annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, 1–6.
- Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation* 51(3), 581–612.
- Zhang, S. and Zeldes, A. (2017). GitDOX: A Linked Version Controlled Online XML Editor for Manuscript Transcription. In: *Proceedings of FLAIRS 2017, Special Track on Natural Language Processing of Ancient and other Low-resource Languages*. Marco Island, FL, 619-623.
- Zipser, F. & Romary, L. (2010). A Model Oriented Approach to the Mapping of Annotation Formats using Standards. In: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. Valletta, Malta, 7–18.