Using modern languages to parse ancient ones

a test on Old English

Luca Brigada Villa & Martina Giarda

{luca.brigadavilla, martina.giarda}@unibg.it

University of Bergamo/Pavia

6 May, 2023
SIGTYP - Dubrovnik

Outline of this talk

1. Introduction
2. Methods
3. Results
4. Parsing errors
5. Conclusion and future aims

Introduction

In our paper we tested the parsing performances of a multilingual parser on Old English data using different sets of languages to train the models:

- support langages alone
- support languages combined with target language

Then, we analyzed more in deep the annotation of some peculiar syntactic constructions and we provided **plausible linguistic explanations** of the errors made by the best performing models.

Why Old English

Reasons for choosing Old English as target language for our study:

- linguistic research
- little attention to the creation of resources to study Old English → scarcity of annotated data for this language
- only constituency treebanks are available for OE
  - York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE - Old English prose)
  - York-Helsinki Parsed Corpus of Old English Poetry (YCOEP - Old English poetry)

Methods

To carry out our research, we followed these steps:

1. we annotated 292 sentences of two prose OE texts:
    - *Adrian and Ritheus*
    - first homily of Ælfric's *Supplemental Homilies*
2. we chose 3 support modern languages:
    - German
    - Icelandic
    - Swedish
3. we used UUParser v2.4 (de Lhoneux et al., 2017) to train the models and parse the test set

Annotation

The texts were annotated following the Universal Dependencies annotation scheme:

1. we converted the texts from the YCOE format to CoNLL-U
2. we converted the annotation of the parts-of-speech in the YCOE to `upos` and `feats` and disambiguate them
3. we annotated the syntax

Old English

- Old English is a **West-Germanic language**, classified with Old Frisian and Old Saxon among the so-called Ingvaeonic languages
- Language spoken in England after Angles, Saxons, Jutes and Frisians came to Britain and settled in the island in the 5th century. → Attestations: from the 7th century (except for some older brief runic inscriptions) to conventionally 1066 (Norman Conquest of England)
- Some OE features:
  - Nominative-accusative alignment
  - 4 cases: nominative, accusative, dative, genitive, (residual instrumental) → case syncretism
  - 3 nominal classes and 2 verbal main conjugational systems (weak and strong)
  - relatively free word order
  - **both pre- and postpositions**
  - discontinuous constituents (often in **relative clauses**)

Choice of the support languages

As support languages we chose three modern languages, namely German, Icelandic and Swedish. The choice was supported by the fact that these languages share with OE some features and belong to the same language family.

To train the multilingual models, we selected 60k tokens from three treebanks (from UD v2.11):

- UD Swedish-Talbanken
- UD Icelandic-Modern
- UD German-GSD

Choice of the support languages

Icelandic

- North-Germanic language
- most ‘archaic’ language
- prenominal definite determiners
- pre- and post-nominal attributive genitive
- the so-called ‘oblique objects’ (i.e. impersonal constructions)
- the presence of verb-auxiliary constructions
- V2

Swedish

- North-Germanic language
- has undergone a process of morphological simplification
- prenominal possessive determiners
- V2

German

- West-Germanic language
- prenominal definite determiners
- pre- and post-nominal attributive genitive
- verb-final order (in subordinate clauses)
- both prepositions and postpositions
- V2

Why not Modern English? Loss of nominal and verbal morphology, strict SVO order, French influence in the lexicon...

Trainig and parsing

For each one of the combinations of the four languages (the target language and the three support languages) we trained a model (following the methodology described in Meechan-Maddon and Nivre 2019):

1. we used UUParser to train the model (30 epochs)
2. we select the best epoch according to the LAS on the OE dev set
3. we parsed the OE test set using the best model

Data and scripts can be found here →

unipv-larl/wundorsmitha-geweorc

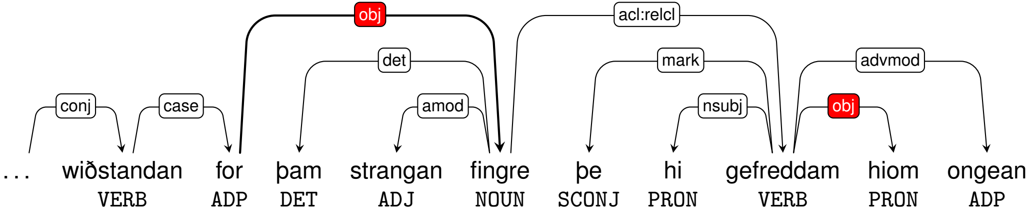
Model performances

	-Target			**+Target**		
	UAS	**LA**	**LAS**	**UAS**	**LA**	**LAS**
Old English				60.79	64.39	47.23
sv	27.06	24.44	9.45	65.07	73.61	57.20
de	32.91	25.34	10.12	65.82	72.19	56.45
is	20.31	22.64	4.57	**68.44**	73.76	**58.70**
sv+de	32.16	25.56	10.42	65.82	72.19	57.42
sv+is	26.39	23.76	9.45	64.62	70.09	54.42
de+is	30.73	27.74	11.17	66.34	**74.29**	57.42
sv+de+is	32.46	24.96	11.02	65.97	71.66	57.57

Deprel `obl`

No significant patterns of error can be identified for the deprel `obl`

- The models fail to recognize *ongean* ‘towards’ as the postposition governing the pronoun *hiom* ‘them’

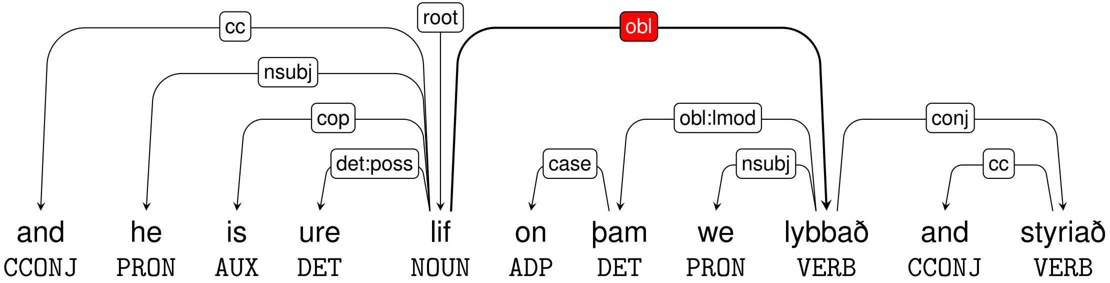


Translation: ‘[and they could no longer] withstand [Moyses] for that strong finger that they felt against them’

Deprel `acl:relcl` - variability

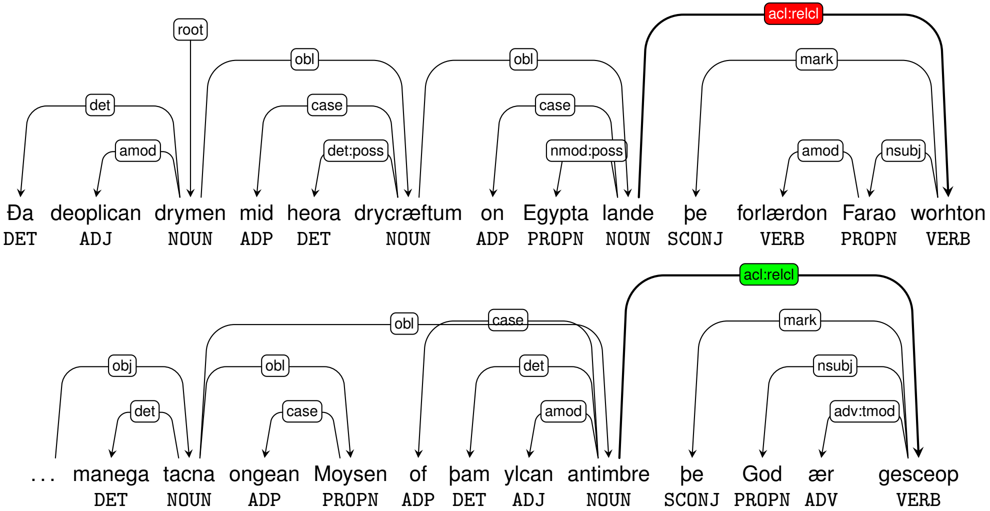
Most problems in the annotation of relative clauses are:

- the great variability in the relative pronouns marking them
- non-projectivity.



Translation: ‘and he is our life, in which we live and move, in which we are, so as Paul said to us’

Deprel `acl:relcl` - non-projectivity



Đa	deoplican	drymen	mid	heora	drycræftum	on	Egypta	lande	þe	forlærdon	Farao	worhton
DET	ADJ	NOUN	ADP	DET	NOUN	ADP	PROPN	NOUN	SCONJ	VERB	PROPN	VERB

. . .	manega	tacna	ongean	Moysen	of	þam	ylcan	antimbre	þe	God	ær	gesceop
	DET	NOUN	ADP	PROPN	ADP	DET	ADJ	NOUN	SCONJ	PROPN	ADV	VERB

Translation: ‘The deep joys, **which corrupted the Pharaoh** with their magical arts in the lands of Egypt, made towards Moyses many signs of the same substance, **which God had created before**...’

Correction of some deprels

We noticed some recurrent errors that the models could have avoided. These errors are due to the fact that the generated tree and the annotation of dependency relations do not take into account the POS of the tokens.

We corrected the output following these rules:

form	**upos**	**xpos**	**deprel**
ne	CCONJ	any	cc
ne	PART	any	advmod:neg
any	any	starts with MD	aux
any	any	ADV^L	advmod:lmod
any	any	ADV^T	advmod:tmod

The correction improved the LAS of the parsed sentences by 1%

Conclusion

In our paper we showed that:

- the model trained just using data of the target language achieved far better results than the models (both monolingual and multiliguals) trained without target language data
- **Icelandic and German combined better with OE** than Swedish according to the scores reached parsing OE test data
- some poor results might be due to the peculiarity of such constructions in OE
- using support languages in combination with the target language to train the models improve the results of the parsing, in particular when:
    - support languages **are related with the target language** or
    - when they **share a significant number of features with the target language**

Future aims

In the future, we would like to:

- have an alternative to a rule-based conversion of the YCOE(P) treebanks
- develop a tool to annotate other OE texts, which are not included in the above-mentioned treebanks

Thank you for your attention!

✉ luca.brigadavilla@unibg.it

✉ martina.giarda@unibg.it

unipv-larl