Corpus-based Syntactic Typological Methods for Dependency Parsing Improvement

Diego Alves
Božo Bekavac
Daniel Zeman
Marko Tadić

06/05/2023

Outline

- Introduction
- Experiment Design
  - Data
  - Corpus-based Typological approaches
  - Dependency Parsing
  - Correlations
- Results
- Conclusions and Perspectives

Introduction

- **Motivation:**
  - Typological approaches used for NLP improvement usually concern phylogenetic characteristics or features provided by typological databases such as WALS
  - Corpus-based typological studies are usually focused on the research of universals, language complexity, or specific syntactic phenomena

- **Objective :**
  - To propose an examination of several corpus-based typological methods in terms of correlation between language distances and dependency parsing scores when languages are combined in the training step

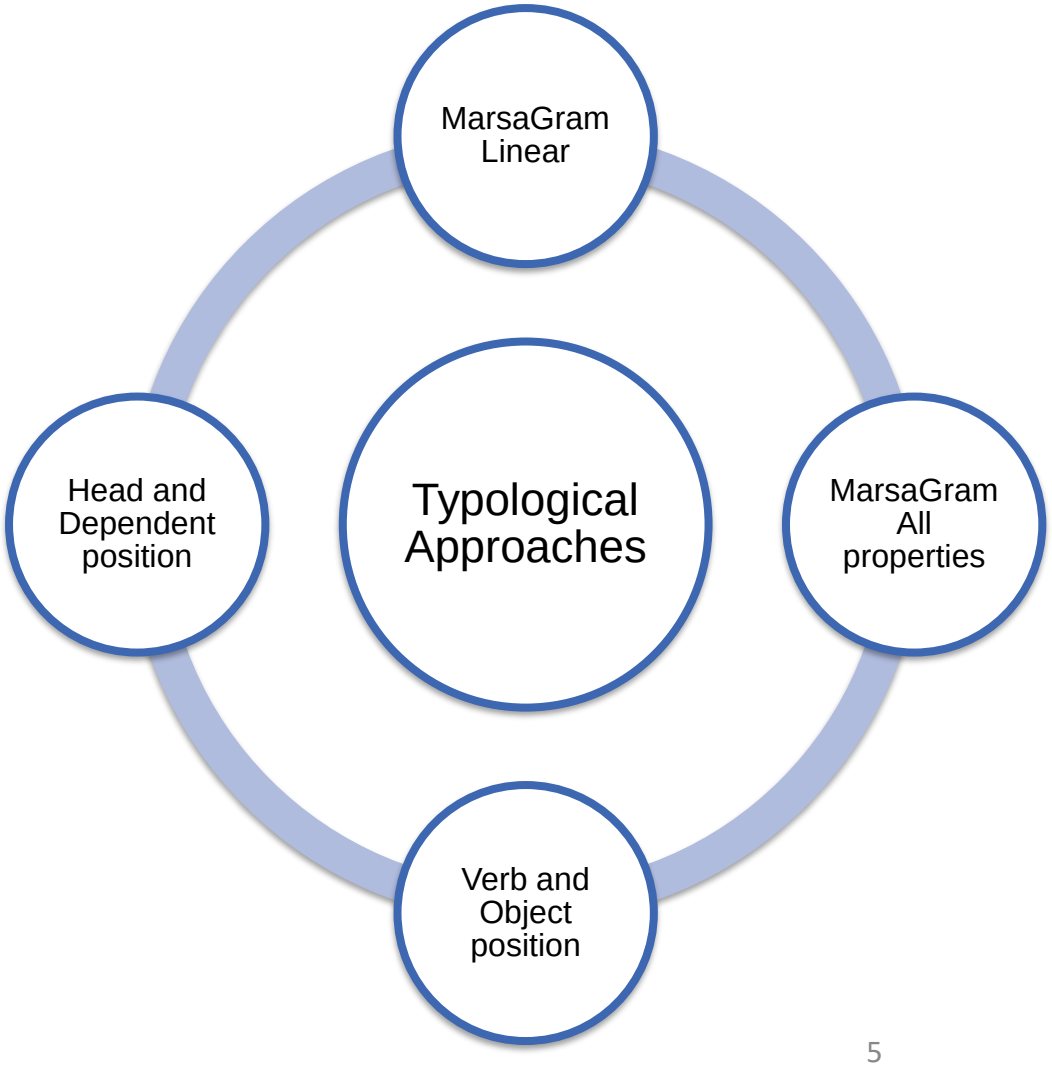
Data

- Parallel Universal Dependencies (PUD):
  - 20 languages: 9 linguistic families, 12 genera
  - 1,000 sentences per language (CoNLL-U)
- Advantages:
  - Homogeneity in terms of size
  - Same semantic content
- Disadvantage:
  - “Translationese”

Language	ISO-639-3	Family	Genus
Arabic	arb	Afro-Asiatic	Semitic
Chinese	cmn	Sino-Tibetan	Chinese
Czech	ces	Indo-European	Slavic
English	eng	Indo-European	Germanic
Finnish	fin	Uralic	Finnic
French	fra	Indo-European	Romance
German	deu	Indo-European	Germanic
Hindi	hin	Indo-European	Indic
Icelandic	isl	Indo-European	Germanic
Indonesian	ind	Austronesian	Malayo-Sumbawn
Italian	ita	Indo-European	Romance
Japanese	jpn	Japanese	Japanese
Korean	kor	Korean	Korean
Polish	pol	Indo-European	Slavic
Portuguese	por	Indo-European	Romance
Russian	rus	Indo-European	Slavic
Spanish	spa	Indo-European	Romance
Swedish	swe	Indo-European	Germanic
Thai	tha	Tai-Kadai	Kam-Tai
Turkish	tur	Altaic	Turkic

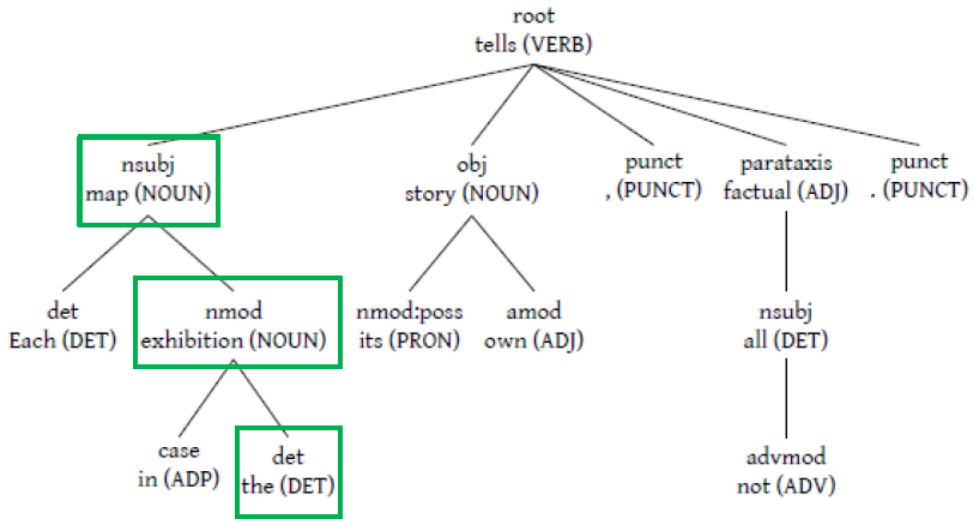
Corpus-based Typological Approaches

- For each method:
  - Definition of language vectors
    - Features: syntactic patterns
    - Values: frequency in the corpus
  - Language comparison:
    - Euclidean and cosine distances → dissimilarity matrix
    - Clustering analysis Ward's method
- Corpus-based approaches are compared with the classification obtained using syntactic features provided by lang2vec tool:
    - 41 features with valid values for PUD languages



Corpus-based Typological Approaches

- MarsaGram Linear:
  - Tool that identifies patterns from context-free grammars extracted from annotated data sets that allow statistical comparison between languages
  - Linear property: Element A precedes element B in a sub-tree with element C as head
    - C_precede_A_B
  - Example : NOUN_precede_DET det_NOUN_nmod
  - 21,242 linear patterns extracted from the PUD corpora



Sentence: “Each map in the exhibition tells its own story, not all factual.”

Corpus-based Typological Approaches

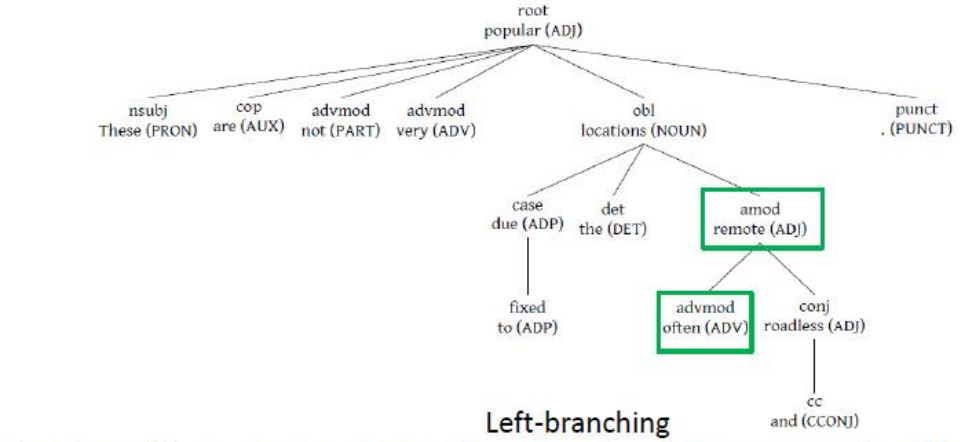
- Marsagram all properties:

  - **Linear**
  - **Require** → The presence of an element A requires the existence of an element B inside the sub-tree
  - **Unicity** → An element A has this property if inside the sub-tree it occurs only once (i.e.: no other element with the same part-of-speech and dependency label is attested)
  - **Exclude** → The presence of element A excludes the occurrence of element B inside the sub-tree.

Property	Number of Patterns	%
Linear	21,242	13.38
Require	6,189	3.90
Unicity	2,144	1.35
Exclude	129,18	81.37

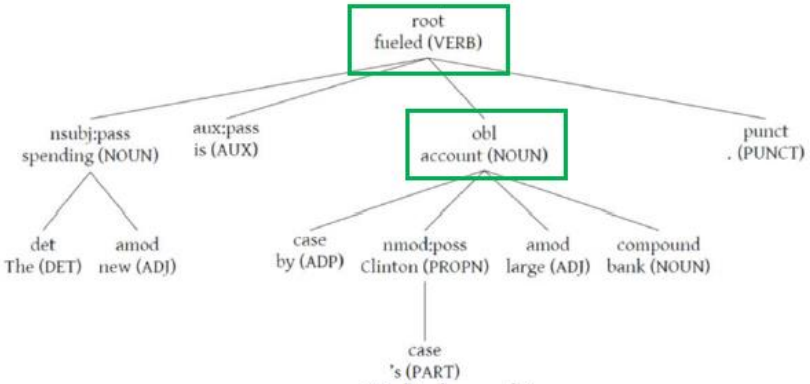
Corpus-based Typological Approaches

- Head and Dependent position
  - Based on the head directionality parameter
    - Left branching
      - Example: ADV_advmod_precedes_ADJ
    - Right branching
      - Example: NOUN_obl_follows_VERB
  - 2,890 patterns extracted from the PUD corpora



Left-branching

Sentence: "These are not very popular due to the often remote and roadless locations."



Right-branching

Sentence: "The new spending is fueled by Clinton's large bank account."

Corpus-based Typological Approaches

- Verb and Object position
  - Based on the importance of these elements in typological studies
  - Extracted patterns are a sub group of the head and dependent approach

  - Head: Verb
  - Object: any POS with obj as the dependency relation

  - 13 OV and 12 VO features were attested in the PUD collection

Dependency Parsing

- UDify tool (Kondratyuk and Straka, 2019)
  - From raw text to dependency parsing using fine-tuning of Multilingual BERT (104 languages)
  - Parameters:
    - Number of epochs: 80
    - Warmup: 500
    - 6 different random seeds
- Baseline: monolingual models
  - Train – 600 sentences
  - Dev – 200 sentences
  - Test – 200 sentences
- Corpora association experiments:
  - Languages combined in pairs
  - Train – 1,200 sentences (600 from L1 and 600 from L2)
  - Dev and test sets → monolingual

Correlations

- Pearson's and Spearman's correlations calculated between:
  - Language distances provided by each corpus-based method and lang2vec
  - LAS deltas:
    - $Delta\ LAS = LAS_{lang_1+lang_2} - LAS_{lang_1}$
  - Strong negative correlation: between -1 and -0,7
  - Moderate negative correlation: between -0,7 and -0,5

Results – Dependency Parsing Experiments

Language	LAS	Std. Dev.
tha	74.68	0.13
cmn	74.84	0.56
tur	76.68	0.21
hin	77.46	0.35
isl	78.90	0.16
fin	82.46	0.28
arb	83.34	0.24
swe	84.69	0.26
ind	85.72	0.19
kor	85.99	0.20
eng	86.63	0.15
ces	86.80	0.40
pol	86.88	0.21
rus	88.42	0.15
ita	89.48	0.14
deu	89.55	0.17
por	89.65	0.16
fra	91.20	0.21
spa	91.24	0.09
jpn	91.57	0.20

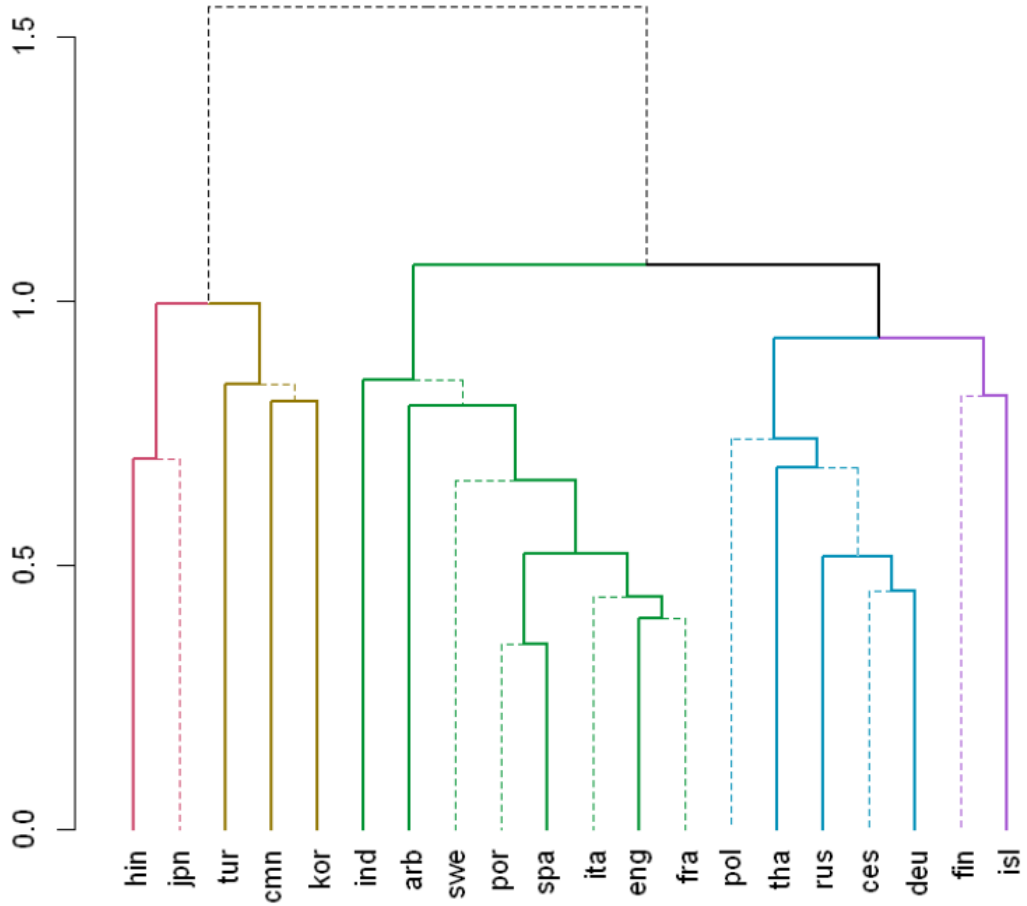
	Positive LAS Deltas ($p<0.01$)	Negative LAS Deltas ($p<0.01$)
hin	0	0
jpn	0	6
kor	0	14
ind	1	1
tha	1	6
arb	2	0
fra	3	0
cmn	4	0
tur	4	1
deu	6	0
pol	9	0
ita	10	0
por	11	0
spa	11	0
ces	12	0
eng	14	0
isl	14	0
swe	14	0
rus	15	0
fin	16	0

Results - Correlations

		MarsaGram All		MarsaGram Linear		Head and Dependent		VO_OV		Lang2vec	
		Euc.	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**	**Euc.**	**cos**
Pearson	Strong	0	0	0	0	0	1	1	2	1	1
	Moderate	3	8	3	10	7	7	5	2	6	5
	Total	3	8	3	10	7	8	6	4	7	6
Spearman	Strong	0	1	0	0	1	2	2	0	1	1
	Moderate	3	2	3	7	6	5	5	5	5	5
	Total	3	3	3	7	7	7	7	5	6	6

- Strongest correlation → MarsaGram Linear typological strategy
  - Moderate correlation for 10 out of 20 PUD languages
  - Better results than SOTA lang2vec tool

MarsaGram Linear – Cluster analysis



Conclusions and Perspectives

- From the selected corpus-based strategy, the MarsaGram Linear one presented better results in terms of correlation (Pearson's) than the other methods
- However, this strategy only showed a moderate correlation for half of the PUD languages
- Thus, in the future, our aim is to:
  - Increase the language-set → test with non-parallel corpora
  - Optimize this strategy by identifying the patterns which may play a major role in dependency parsing experiments when different languages are combined

Questions?

dfvalio@ffzg.hr