Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting

Café TAL

Hee-Soo Choi, Bruno Guillaume, Karën Fort, Guy Perrier

August 31th, 2021

Goal

Determine a dominant word order of **Subject (S), Object (O), Verb (V)** on **74 languages (141 corpora)** of **Universal Dependencies** (UD) using **GREW**, a Graph Rewriting tool and compare our results with other references.

grew

Motivations

Typology and NLP

- Typological feature of dominant word order of Subject, Object and Verb

Typology and NLP

- Typological feature of dominant word order of Subject, Object and Verb
- Utility of linguistic typology in NLP [Bender, 2016, O'Horan et al., 2016]:
  - In language transfer [Naseem et al., 2012, Ahmad et al., 2019]
  - In developping language-independent systems [Bender, 2009]

Typology and NLP

- Typological feature of dominant word order of Subject, Object and Verb
- Utility of linguistic typology in NLP [Bender, 2016, O'Horan et al., 2016]:
  - In language transfer [Naseem et al., 2012, Ahmad et al., 2019]
  - In developping language-independent systems [Bender, 2009]
- Universal annotations of UD allow typological experiments on several languages

Methodology

Taking Corpora as Basis

- Corpus’ degree of representativeness of the language

Taking Corpora as Basis

- Corpus' degree of representativeness of the language

- Corpora with fewer than 1,000 sentences eliminated:
  → UD 2.7: 104 languages, 183 corpora
  → **UD 2.7_{1K}: 74 languages, 141 corpora**

Taking Corpora as Basis

- Corpus' degree of representativeness of the language

- Corpora with fewer than 1,000 sentences eliminated:
  → UD 2.7: 104 languages, 183 corpora
  → **UD 2.7_{1K}: 74 languages, 141 corpora**

- Corpus level to observe variations between corpora of a given language:
  → 29 languages with more than one corpus
  → 45 languages with only one corpus

Defining a Dominant Word Order

- Count frequencies of the six possibles orders: SVO, SOV, VSO, VOS, OVS, OSV

Defining a Dominant Word Order

- Count frequencies of the six possibles orders: SVO, SOV, VSO, VOS, OVS, OSV

- Two meanings according to WALS [Dryer and Haspelmath, 2013]:
  → the order is **the only possible one** for the language
  → the language exhibits several different orders and **one is more frequently used**

Defining a Dominant Word Order

- Count frequencies of the six possibles orders: SVO, SOV, VSO, VOS, OVS, OSV

- Two meanings according to WALS [Dryer and Haspelmath, 2013]:
  → the order is **the only possible one** for the language
  → the language exhibits several different orders and **one is more frequently used**

- The most frequent order considered as the dominant order if it is **at least twice as frequent** as the next most frequent:
  → ratio ≥ 2: the most frequent order is the dominant order
  → ratio < 2: No Dominant Order (NDO)

Using Graph Rewriting

GREW: Graph Rewriting Tool

Graph rewriting tool dedicated to NLP applications

- Query corpora using graph patterns
- Count the occurrences of each pattern in each corpus

GREW pattern for SVO order:

```
pattern {
    V [upos=VERB];
    V -[1=nsubj]-> S;
    V -[1=obj]-> O;
    S << V; V << O
}
```

Limits of UD annotations

- A subject can not be linked to several verbs

Limits of UD annotations

- A subject can not be linked to several verbs
- The subject and the object may not be related to the same verb

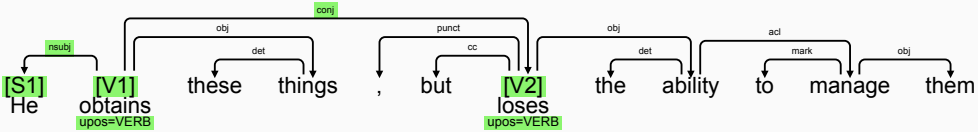
Enriching UD annotations

Two cases recovered by adding implicit subjects (isubj):

Enriching UD annotations

Two cases recovered by adding implicit subjects (isubj):

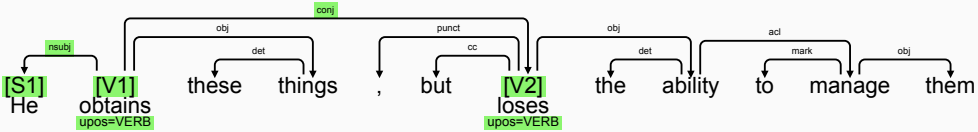
- Coordination:



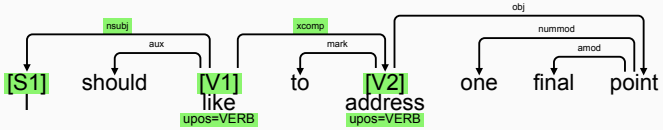
Enriching UD annotations

Two cases recovered by adding implicit subjects (isubj):

- Coordination:



- Control or raising:



Results

Dominant Word Order in Multi-Corpora Languages

Intra-language Consistency

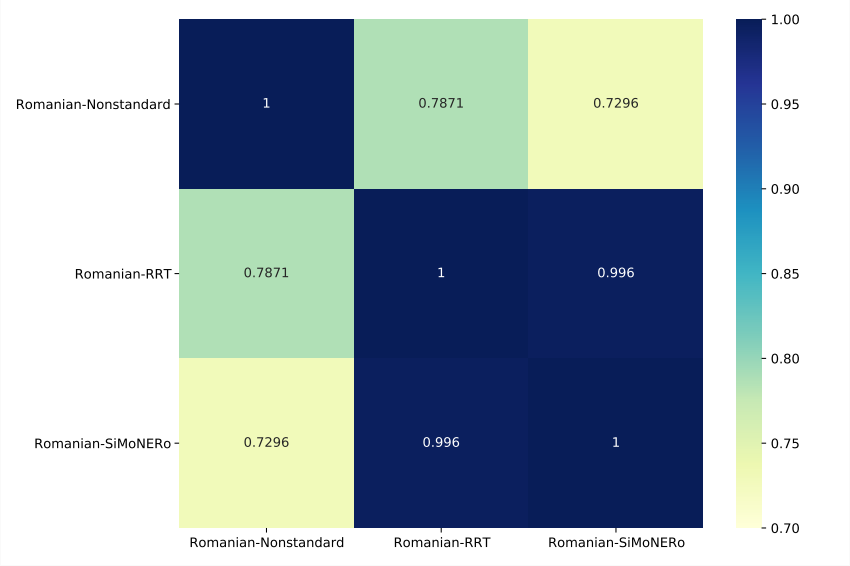
- Corpora as vectors with the frequencies of the six orders

	SVO	SOV	VSO	VOS	OSV	OVS
Romanian_Nonstandard	**38.07%**	**31.87%**	9.66%	3.97%	1.71%	14.72%
Romanian_RRT	**85.32%**	7.76%	1.12%	0.70%	1.18%	3.91%
Romanian_SiMoNERo	**97.61%**	0.97%	0.09%	0.09%	0.13%	1.10%

Distribution vectors for the Romanian corpora.

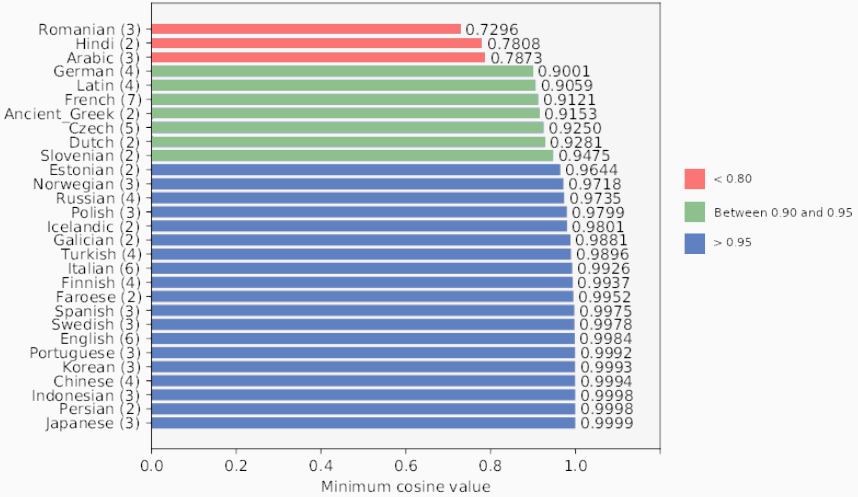
- Computing the cosine between the vectors for each corpus
- Cosine value close to 1 when two corpora display similar distributions

Intra-language Consistency



Cosine values between the three Romanian corpora in UD 2.7_{1K}.

Intra-language Consistency



Multi-corpora (nb in parenthesis) languages ordered by minimum cosine value.

Basic Analysis

Possible explanations for intra-language inconsistency:

Basic Analysis

Possible explanations for intra-language inconsistency:

- Different text genres: `Romanian-NonStandard`, `French-FQB`

Basic Analysis

Possible explanations for intra-language inconsistency:

- Different text genres: `Romanian-NonStandard`, `French-FQB`
- Different text periods: `Latin`, `Ancient Greek`, `German-LIT`

Basic Analysis

Possible explanations for intra-language inconsistency:

- Different text genres: `Romanian-NonStandard`, `French-FQB`
- Different text periods: `Latin`, `Ancient Greek`, `German-LIT`
- Non-standard annotations: `Hindi-HDTB` where the object is a verb

Basic Analysis

Possible explanations for intra-language inconsistency:

- Different text genres: `Romanian-NonStandard`, `French-FQB`
- Different text periods: `Latin`, `Ancient Greek`, `German-LIT`
- Non-standard annotations: `Hindi-HDTB` where the object is a verb
- Language specifics: `Arabic-PADT` with topicalization

Comparison with other sources

Comparison with WALS

59 languages in common, same dominant word order for 48

Language	UD 2.7_{1K}	WALS
Amharic	**1 NDO**	**SOV**
Arabic	1 VSO, 2 NDO	VSO
Belarusian	**1 SVO**	**NDO**
Estonian	1 SVO, 1 NDO	SVO
German	2 SOV, 2 NDO	NDO
Greek	**1 SVO**	**NDO**
Hindi	1 SOV, 1 NDO	SOV
Mbya Guarani	**1 NDO**	**SVO**
Romanian	2 SVO, 1 NDO	SVO
Slovenian	1 SVO, 1 NDO	SVO
Urdu	**1 NDO**	**SOV**

Differences with WALS.

Comparison with Östling [Östling, 2015]

- Word order typology based upon the translated and aligned New Testament
- 52 languages in common, same dominant order for 38

Language	UD 2.7_{1K}	Östling
Amharic	**1 NDO**	**SOV**
Ancient Greek	2 NDO	SVO
Armenian	1 NDO	SVO
Basque	**1 SOV**	**SVO**
Dutch	2 NDO	SOV
Estonian	1 SVO, 1 NDO	SVO
German	2 SOV, 2 NDO	SOV
Hindi	1 SOV, 1 NDO	SOV
Hungarian	1 NDO	SVO
Latin	1 SOV, 3 NDO	SVO
Mbya Guarani	1 NDO	SVO
Romanian	2 SVO, 1 NDO	SVO
Slovenian	1 SVO, 1 NDO	SVO
Welsh	**1 VSO**	**SVO**

Differences with Östling.

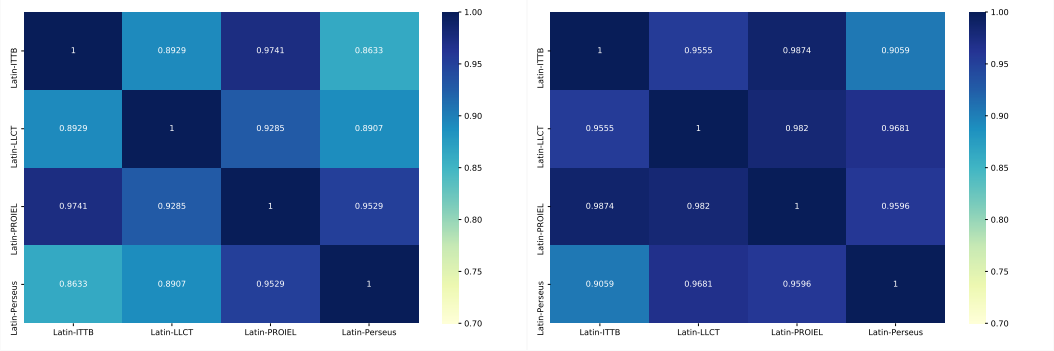
Influence of Implicit Subjects

Comparison with/without isubj

		Without isubj		With isubj	
Language	Corpora	Order	Ratio	Order	Ratio
Czech	CAC	SVO	4.27	SVO	5.28
	CLTT	SVO	6.85	SVO	8.18
	FicTree	**NDO (SVO/SOV)**	**1.97**	**SVO**	**2.20**
	PDT	SVO	3.36	SVO	3.96
	PUD	SVO	6.58	SVO	6.17
Estonian	EDT	SVO	3.80	SVO	3.19
	EWT	**SVO**	**2.05**	**NDO (SVO/SOV)**	**1.70**
German	GSD	NDO (SOV/SVO)	1.03	NDO (SOV/SVO)	1.03
	HDT	**NDO (SOV/SVO)**	**1.87**	**SOV**	**2.01**
	LIT	SOV	2.30	SOV	2.53
	PUD	NDO (SOV/SVO)	1.47	NDO (SOV/SVO)	1.62
Latin	ITTB	NDO (SVO/SOV)	1.22	NDO (SVO/SOV)	1.12
	LLCT	**NDO (OSV/SVO)**	**1.07**	**NDO (SOV/SVO)**	**1.40**
	PROIEL	NDO (SOV/SVO)	1.21	NDO (SOV/SVO)	1.16
	Perseus	SOV	2.42	SOV	2.17

Corpora for which the word order changes with/without isubj and associated ratio.

Comparison with/without isubj



Cosine values between the Latin corpora, without isubj on the left, with isubj on the right.

Conclusion

Conclusion

- Results obtained on UD corpora consistent with typology databases

Conclusion

- Results obtained on UD corpora consistent with typology databases

- They can be used to NLP applications or to complete databases:
  → WALS does not cover dead-languages
  → WALS does not provide feature 81A for six languages: Faroese, Galician, Kazakh, Maltese, Naija and Slovak

Conclusion

- Results obtained on UD corpora consistent with typology databases
- They can be used to NLP applications or to complete databases:
  → WALS does not cover dead-languages
  → WALS does not provide feature 81A for six languages: Faroese, Galician, Kazakh, Maltese, Naija and Slovak
- GREW useful to query UD corpora and to overcome limits of UD annotations

Thank you for your attention

References

References i

Ahmad, W., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., and Peng, N. (2019).
On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Bender, E. M. (2009).
Linguistically naïve != language independent: Why NLP needs linguistic typology.
In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

References ii


Bender, E. M. (2016).
Linguistic typology in natural language processing.
Linguistic Typology, 20:645–660.


Dryer, M. S. and Haspelmath, M., editors (2013).
***WALS Online*.**
Max Planck Institute for Evolutionary Anthropology, Leipzig.


Naseem, T., Barzilay, R., and Globerson, A. (2012).
Selective sharing for multilingual dependency parsing.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

References iii


O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., and Korhonen, A. (2016).
Survey on the use of typological information in natural language processing.
In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.


Östling, R. (2015).
Word order typology through multilingual word alignment.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

GREW rule for isubj

```
rule conj {
    pattern {
        V1 [upos=VERB]; V2 [upos=VERB];
        V1 -[1=conj]-> V2;
        V1 -[1=nsubj]-> S1;
    }
    without { V2 -[1=nsubj]-> S2; }
    commands { add_edge V2 -[isubj]-> S1; }
}
```

GREW rule adding the isubj relation.