

Lexical categories and coding efficiency: A cross-linguistic corpus approach

Yè Jìngtíng

Leipzig University

21.11.2019



UNIVERSITÄT
LEIPZIG



European Research Council
Established by the European Commission

Outline

- 1 Introduction
- 2 Universal Dependencies Database
- 3 Results
- 4 Conclusions

Outline

1 Introduction

2 Universal Dependencies Database

3 Results

4 Conclusions

Background

At the heart of the entire discipline of linguistics is our understanding of part-of-speech systems and a huge number of attempts have been made to puzzle this issue out. e.g. The verb-noun continuum hypothesis:

- Ross (1972): **The category squish**: verb > present participle > perfect participle > passive participle > adjective > preposition (?) > adjectival noun (e.g. fun, snap) > noun
- Comrie (1975): As part of universal grammar, we require a **continuum** from VERB to NOUN, with individual languages requiring different intermediate positions on this hierarchy...Adjectives occupy the intermediate position between nouns and verbs.
- Pustet (1989); Wetzler (1996): verb > verby adj > nouny adj > noun

Background

Croft's theory: the cross-linguistic pattern of syntactic categories is shaped by the **prototypicality** of the correlation between pragmatic functions (Reference, Modification, Predication) and lexical categories (Objects, Properties, Actions).

Table 1: Prototypical correlations between lexical classes and pragmatic functions (Croft 1991)

	Reference	Modification	Predication
Object	unmarked nouns		
Property		unmarked adjectives	
Action			unmarked verbs

Background

Table 2: English examples of unusual and usual associations between semantic classes and pragmatic functions (Croft 1991)

	Reference	Modification	Predication
Object	vehicle	vehicular	<i>be a/the vehicle</i>
Property	whiteness	white	<i>be white</i>
Action	destruction	destroying/destroyed	destroy

The unmarked forms are never longer than the marked forms.

The claim

- The form-frequency correspondence hypothesis: Form-Frequency Correspondence Hypothesis (Zipf 1935; Haspelmath 2008; Haspelmath et al. 2014):
More frequent expressions tend to be coded with shorter forms.
- Dryer (2018): The noun-verb distinction reflects the different frequency with which different sorts of words are used as arguments or as syntactic predicates rather than any semantic or conceptual distinction.
- This talk claims: The noun-verb-**adjective** distinction reflects the different frequency with which different sorts of words are used as arguments, as **modifiers** or as syntactic predicates.
- In addition, the correlation between frequency and inflection potential of adjectives will also be addressed.

Outline

1 Introduction

2 Universal Dependencies Database

3 Results

4 Conclusions

What is Universal Dependencies?

- Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 200 contributors producing more than 100 treebanks in over 70 languages.
- The annotation scheme is based on an evolution of (universal) Stanford dependencies (De Marneffe et al., 2006; De Marneffe and Manning, 2008; De Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2011), and the Intersect interlingua for morphosyntactic tagsets (Zeman, 2008).
- <http://universaldependencies.org/>

The UD is an ongoing project. In the current stage, 15 language families are included.

Universal Dependencies













Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 60 languages.

- [Short introduction to UD](#)
- [UD annotation guidelines](#)
- More information on UD:
 - [How to contribute to UD](#)
 - [Tools for working with UD](#)
 - [Discussion on UD](#)
 - [UD-related events](#)
- Query UD treebanks online:
 - [SETS treebank search](#) maintained by the University of Turku
 - [PML Tree Query](#) maintained by the Charles University in Prague
 - [Kontext](#) maintained by the Charles University in Prague











































If you want to receive news about Universal Dependencies, you can subscribe to the [UD mailing list](#).

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from [WALS Online](#) (IE = Indo-European).

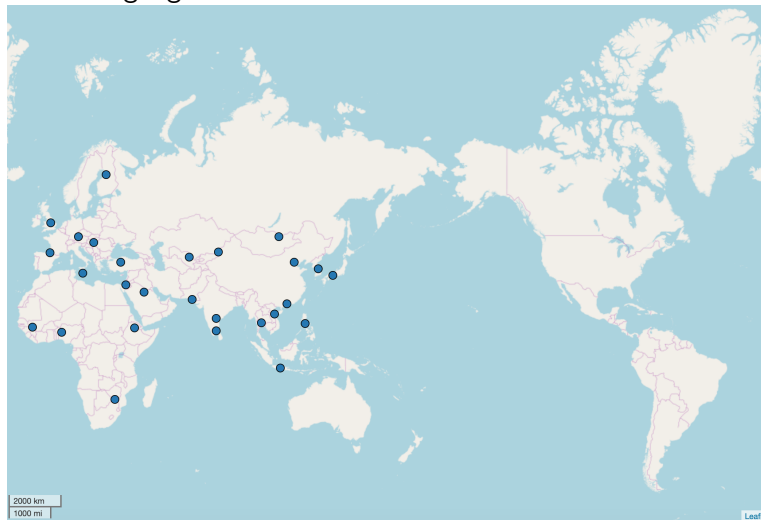
▶		Afrikaans	1	49K		IE, Germanic
▶		Amharic	1	11K		Afro-Asiatic, Semitic
▶		Ancient Greek	2	417K		IE, Greek
▶		Arabic	3	1,042K		Afro-Asiatic, Semitic
▶		Armenian	1	12K		IE, Armenian
▶		Malagasy	1	1K		Malayo-Polynesian

Upcoming UD Languages

▶		Assamese	1	-		IE, Indic
▶		Bengali	2	-	 	IE, Indic
▶		Bhojpuri	1	-		IE, Indic
▶		Cusco Quechua	1	-		Quechuan
▶		Dargwa	1	-		Nakho–Dagestanian
▶		Georgian	1	-		Kartvelian
▶		Kannada	1	-		Dravidian, Southern
▶		Komi Permyak	1	-		Uralic, Permic
▶		Kyrgyz	1	-		Turkic, Northwestern
▶		Livvi	1	-		Uralic, Finnic
▶		Macedonian	1	-		IE, Slavic
▶		Maghrebi Arabic				
▶		French	1	-		Code switching
▶		Pnar	1	-	 	Austro–Asiatic, Khasian
▶		Romansh	2	-		IE, Romance
▶		Scottish Gaelic	1	-		IE, Celtic
▶		Shipibo Konibo	1	-		Panoan
▶		Sindhi	1	-		IE, Indic
▶		Somali	1	-		Afro–Asiatic, Cushitic
▶		Sorani	1	-		IE, Iranian
▶		Swiss German	1	-		IE, Germanic

The sample

The sample consists of 26 languages from 15 language families, and 1 creole language.



How to extract the data from UD

- Word classes are annotated, which makes it relatively easy to get frequency information.
- The second step is to count the actual occurrences according to different usage of reference, modification and predication in the texts. This is also tagged in the corpora. For instance, 'amod' represents 'adjective as modifier'; 'nobj' represents 'noun as object', etc.
- In addition, inflectional categories (case, gender, tense, aspect, etc.) can also be extracted from the corpora.

A glimpse into the actual raw data (from Kazakh)

```
# sent_id = Eurovision_ән_конкурсы_2010.tagged.txt:1:0
# text = Еуровидение 2010 ән конкурсы Еуровидениенің 55-ші конкурсы болады.
1   Еуровидение   Еуровидение   PROPН   np           Case=Nom      4           nmod:poss   _           _
2   2010          2010          NUM     num          NumType=Ord  4           amod        _           _
3   ән             ән            NOUN    n            Case=Nom      4           nmod:poss   _           _
4   конкурсы      конкурс       NOUN    n            Case=Nom|Number[psor]=Plur,Sing|Person[psor]=3  7           nsubj
5   Еуровидениенің Еуровидение   PROPН   np           Case=Gen      7           nmod:poss   _           _
6   55-ші         55           NUM     num          NumType=Ord  7           amod        _           _
7   конкурсы      конкурс       NOUN    n            Case=Nom|Number[psor]=Plur,Sing|Person[psor]=3  0           root
8   болады бол     AUX        v         Mood=Ind|Number=Sing|Person=3|Tense=Aor|VerbForm=Fin  7           cop
9   .             .            PUNCT   sent         _           7           punct       _           _
```

Outline

- 1 Introduction
- 2 Universal Dependencies Database
- 3 Results**
- 4 Conclusions

To see the world in a grain of sand: data from Japanese

Table 3: Token counts

	Reference	Modification	Predication	Total
Verb	7575	2941	658	14356
Adjective	701	3055	162	3925
Noun	2171	10738	27614	40527

Table 4: The percentage* of different pragmatic functions for each word class

	Predication	Modification	Reference
Verb	52%	20%	5%
Adjective	18%	78%	4%
Noun	5%	26%	68%

*Note that the percentage does not add to 100%. This is because the number in the column of 'Total' actually also includes peripheral functions. The percentage represented here is the relative frequency of each word class in different syntactic functions, which can be compared across languages.

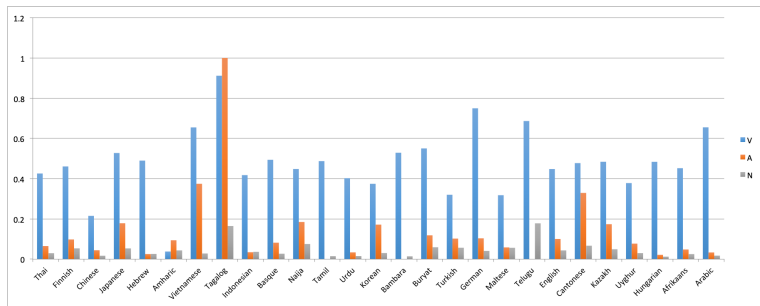
Results

→ Compare the relative frequency for each pragmatic function.

Results

The ranking the frequency of the predicative use is “V > A > N” in most languages, which is in line with the V-N continuum. However, there are two exceptions: Amharic and Tagalog, which may due to the small size of the data for the two languages.

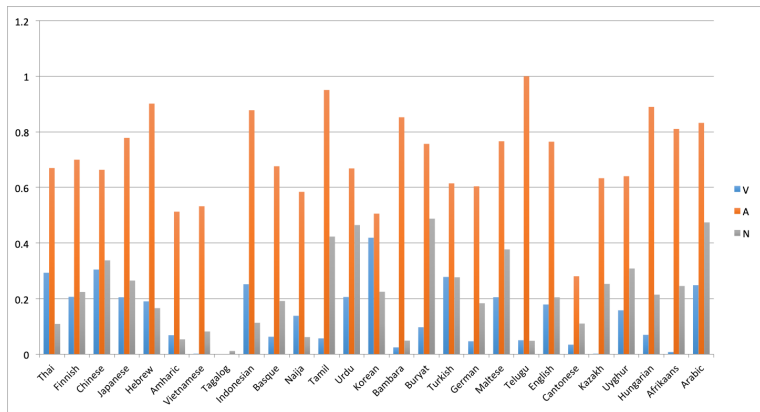
Figure 1: The relative frequency of Predicative function for different word classes



Results

The most frequent word class used as modifiers is adjective.

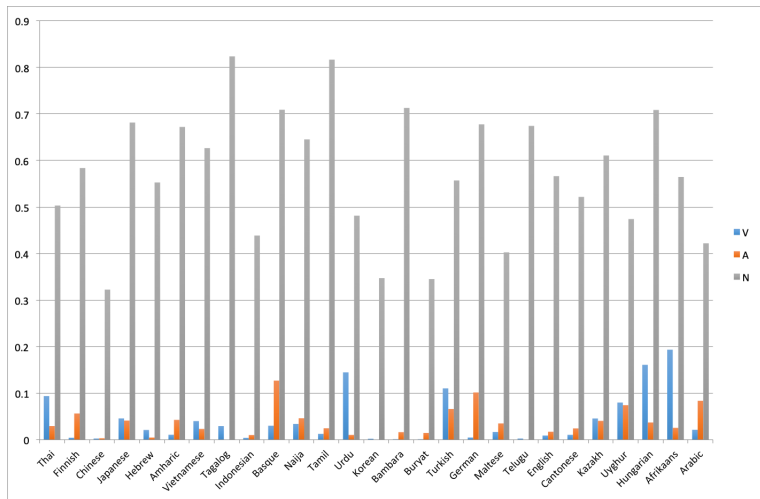
Figure 2: The relative frequency of Modifier function for different word classes



Results

The most frequent word class used as reference is noun.

Figure 3: The relative frequency of Reference function for different word classes



Results

→ Compare the relative frequency for each word class.

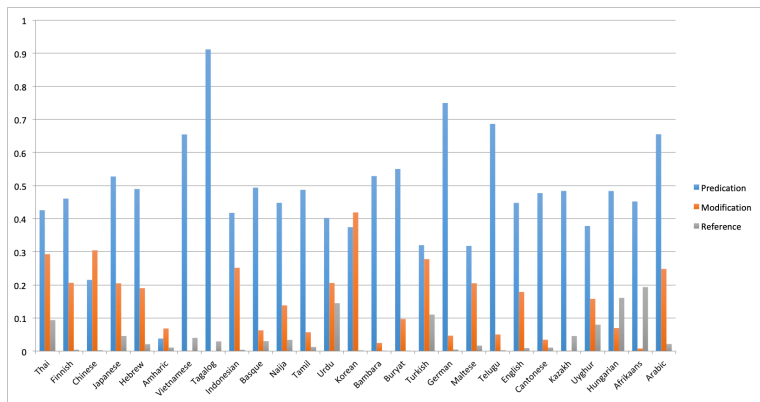
Results

Verbs are most frequently used for predication.

Exceptions: Korean, Amharic, Chinese.

–These languages seem to have a higher percentage of modification use.

Figure 4: The relative frequency of different syntactic functions for verbs

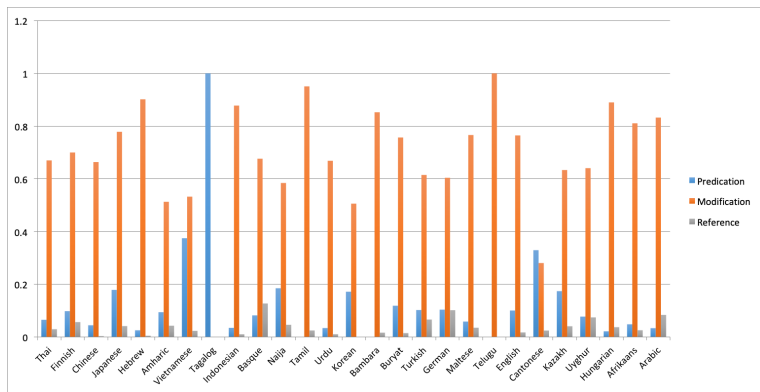


Results

Adjectives are most frequently used for modification.

Exceptions: Tagalog, Cantonese

Figure 5: The relative frequency of different syntactic functions for adjectives



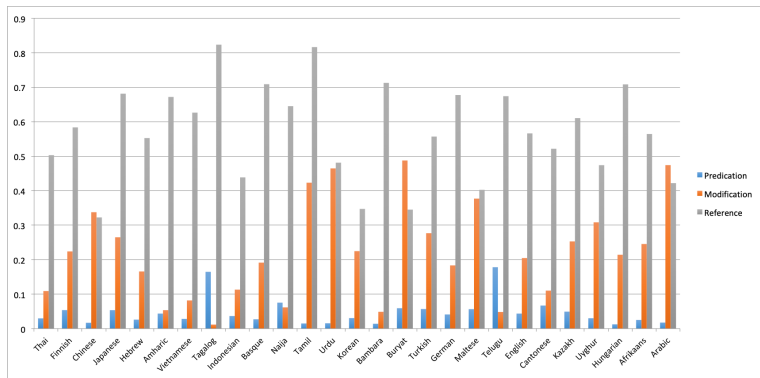
Results

Nouns are most frequently used for reference.

Exceptions: Chinese, Buriat

These languages seem to have a higher proportion for nouns to be used as modifiers than as references.

Figure 6: The relative frequency of different syntactic functions for nouns



More about the adjective

- The adjective is a heterogeneous syntactic category, which shows discrepancies in its functions. The primary function of adjectives is modification, but they are also very often used as predicates. (the **white** snow vs. The snow is **white**.)
- The morpho-syntactic features differ across languages. It would be interesting to see whether there is a correlation between these formal features and frequency of use.

Illustrations:

- (1) Japanese inflection of the verbal domain
naga-i *ressya*
long-NONPAST train
'long train' (Backhouse 2004, 53)
- (2) German inflection of the nominal domain
ein schön-es *Kleid*
one beautiful-NEUT dress
'a beautiful dress'

The claim

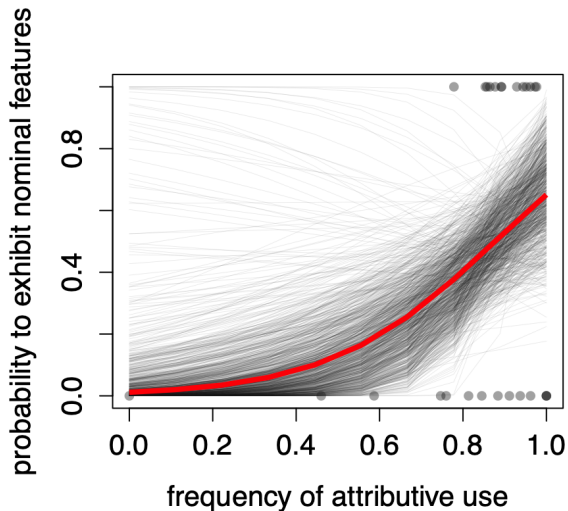
- Adjectives occur less frequently in attributive positions tend to be less likely to exhibit morpho-syntactic features of the nominal domain (case, gender, number, definiteness).
- Adjectives occur less frequently in predicative positions tend to be less likely to exhibit morpho-syntactic features of the verbal domain (tense, aspect, mood, person indexes).
- The less frequent is one element in certain domain, the less entrenched it is for the domain, and hence the less likely for it to exhibit typical morpho-syntactic features of the domain in question (if the canonical members of the domain also exhibit these features).
☺ One takes the colour of one's company, so do words!

The logistic Regression Model

- Logistic Regression Model (calculate with the function glm in R)
- In statistics, the logistic model (or logit model) is a statistical model with input (independent variable) a continuous variable and output (dependent variable) a binary variable, where a unit change in the input multiplies the odds of the two possible outputs by a constant factor.
- What we want to test is whether the lower frequency causes the lower probability of exhibiting certain feature. Frequency is a continuous variable, and whether the feature exists or not in certain language is a binary variable. Because of the property of the two variables, the logistic regression model would be a suitable model to test our claim.

Testing the correlation between frequency of attributive use and the features of the nominal-domain.

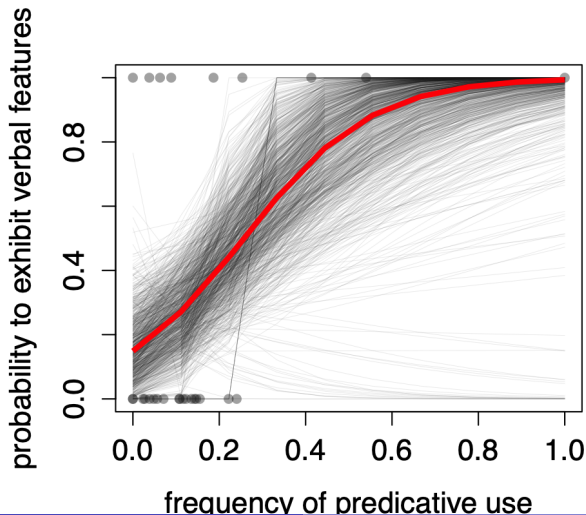
Coefficient: 5.074 , P-value = 0.182 → not significant ☹



Testing the correlation between frequency and the features of the verbal domain

Coefficients: 6.784, P value = 0.0143

→ significant 😊



From the statistical analysis, we could see that there is a significant correlation between the frequency of predicative use and the probability for adjectives to exhibit verbal features.

Outline

- 1 Introduction
- 2 Universal Dependencies Database
- 3 Results
- 4 Conclusions

Conclusions (1)

- The noun-verb-adjective distinction reflects the different frequency with which different sorts of words are used as arguments, as modifiers or as syntactic predicates.
→ This claim is to a large extent supported by the frequency data.
- Lexemes used most frequently as arguments are most likely to be coded with nouns; Lexemes used most frequently as predicates are most likely to be coded with verbs; Lexemes used most frequently as modifiers are most likely to be coded with adjectives.

Conclusions (2)

- Based on the data I have extracted from UD database, it is clear that the frequency of occurrence in the predicative position is ranked as $V > A > N$. This is a quite robust trend observed in UD, and also fit in nicely with the verb-noun continuum hypothesis in the literature.
- The few exceptions may due to the size of the corpus, or peculiarities of certain languages. (Such as, in Cantonese, adjectives are more frequently used as predicates, instead of modifiers. This may due to the fact that adjectives in Cantonese are verb-like.)
- It is also shown that there is a significant correlation between frequency of predicative use and the verbal features (or verbal-domain-codings) of adjectives.

References

- Backhouse, A. E. (2004). Inflected and Uninflected Adjectives in Japanese. In Dixon, R. and Aikhenvald, A., editors, *Adjective classes: A cross-linguistic typology*, pages 50–73. Oxford University Press.
- Comrie, B. (1975). Polite plurals and predicate agreement. *Language*, pages 406–418.
- Croft, W. (1991). *Syntactic categories and grammatical relations: The cognitive organization of information*. University of Chicago Press.
- De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.
- De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.
- De Marneffe, M.-C. and Manning, C. D. (2008). The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics.
- Dryer, M. S. (2018). On the order of demonstrative, numeral, adjective, and noun. *Language*, 94(4):798–833.
- Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, 19(1):1–33.
- Haspelmath, M., Calude, A., Spagnol, M., Narrog, H., and Bamyaci, E. (2014). Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics*, 50(3):587–625.
- Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.
- Pustet, R. (1989). *Die Morphosyntax des "Adjektivs" im Sprachvergleich*. P. Lang.
- Ross, J. R. (1972). The category squish: endstation hauptwort. In *Chicago Linguistic Society*, volume 8, pages 316–328.
- Wetzer, H. (1996). *The typology of adjectival predication*. Walter de Gruyter.
- Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *LREC*, volume 2008, pages 28–30.
- Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. MIT Press.

Thank you!