Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks

Xinying Chen
School of International Studies
Xi'an Jiaotong University, China
Department of Czech Language
University of Ostrava, Czech Republic
xy@yuyanxue.net

Kim Gerdes
LPP (CNRS)
Sorbonne Nouvelle
France
kim@gerdes.fr

Abstract

This paper shows how the current Universal Dependency treebanks can be used for clustering structural global linguistic features of the treebanks to reveal a purely structural syntactic typology of languages. Different uni- and multi-dimensional data extraction methods are explored and tested in order to assess both the coherence of the underlying syntactic data and the quality of the clustering methods themselves.

1 Introduction

Language universality and language differences are a pair of questions, if not two sides of one question, that relate to most of modern linguistic research, both theoretically and empirically. This is even more true for research in language typology.

Modern language typology research (Croft 2002; Song 2001), mostly based on Greenberg (1963), focuses less on lexical similarity and relies rather on various linguistics indices for language classification, and generally puts much emphasis on the syntactic order (word order), in particular of the principal components in relation to their governing verb (Haspelmath et al. 2005).

However, just as individual constructions can display varying degrees of syntheticity and analyticality (Ledgeway 2011), different syntactic orders can also be found in the very same language. Reality seems to be messier than we would like it to be. Therefore, probabilities or quantitative approaches, which allow gradual transitions and blurred borderlines, could make some unique contributions on this matter (Liu & Xu, 2012). Moreover, empirical studies based on authentic language data can bring richer details, and then corroborate or improve our knowledge of language classification. By relying on quantitative empirical measures we do no longer expect a categorical answer of grouping languages into fixed language groups, but rather tendencies of structural proximity between languages.

Although such efforts have already been made in a few studies (Liu 2010; Liu & Xu 2012), it is not until now, with the appearance of Universal Dependencies, that we can conduct an empirical language classification study based on treebanks of different languages that share the same dependency annotation framework.

1.1 Universal Dependencies

Universal Dependencies (UD) is a project of developing a cross-linguistically consistent treebank annotation scheme for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of (universal) Stanford dependencies (de Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary.

There are two notable advantages of using this data set for language classification studies. Firstly, it is the sheer size of the data set: It includes 70 treebanks of 50 languages, 63 of which have more than 10,000 tokens. And secondly, and most importantly, all UD treebanks use the same annotation scheme. The few previous studies of empirical language classification based on

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 54-63,
Pisa, Italy, September 18-20 2017

treebank data (Liu 2010; Liu & Xu 2011, 2012) still had to rely on much fewer treebanks with heterogeneous annotation schemes. Although already relatively satisfying results were obtained, the question of identifying the source of the observed language variations remains unsolved: They could be actual structural differences between languages or simply annotation schema related differences (or even genre related differences, of course – and thus being due to the underlying text). UD can, to a certain extent, reduce this problem by providing a unique framework for all languages.

However, the drawbacks of the UD 2.0 scheme are also rather obvious. The Universal Dependencies (UD) project is still at an early stage of development and many problems of UD have not been solved appropriately, the most important points being:

1. Many treebanks are a result of multiple transformations of previous phrase-structure and dependency treebanks, therefore often multiplying already existing annotation or even parse errors where no manual correction is available.
2. The UD textual data stems from very different sources and was not conceived as a parallel corpus.[1] Thus, we can never exclude that any observed difference is actually due to genre differences between the texts.
3. The current UD annotation guides are still highly underspecified resulting in low inter-annotator, and more importantly inter-corpus agreement (the authors, submitted). This is particularly true for a series of constructions (cleft, dislocations, disfluencies, …). Also, the attempt to annotate semantic non-compositionality of multi-word expressions in the (syntactic) annotation scheme without actually providing the semantic criteria, necessarily leads to incomparable annotations (Gerdes & Kahane 2016).
4. Most importantly, with the goal of possibly simplifying parsing and other NLP tasks, the basic idea underlying the UD annotation scheme is to make languages look as "similar as possible" based on semantic features, the most prominent of which being to put "content words" higher in the tree. However, the status of *content word* is a semantic distinction. This results in the infamous "Turkish" analysis of English prepositions (Chris Manning, 2016, personal communication).[2] The forced similarity of structurally different languages, like for example Turkish and English, makes the data less valuable for our study of empirical structural language classification: We cannot measure what has been suppressed.

1.2 What to measure?

In typological studies on word order, Greenberg (1963) proposed 45 linguistic universals, 28 of which are related to the order or position of grammatical units, for instance, the order of subject, object, and verb. According to Dryer's (1992) study of detailed word order correlations based on a sample of 625 languages, there are 17 correlation pairs and 5 non-correlation pairs between a verb and its object.[3] Although the importance of linear order of grammatical units has been addressed for quite a while, more recently statistical investigations of word order also play an increasingly central role in empirical studies, some of which are based on treebanks. Liu (2010) looked through the directional distributions of three pairs of grammatical units, namely, S-V/V-S, V-O/O-V, and Adj-N/N-Adj, in treebanks of 20 languages. He quantified the dependency directions by computing the percentages of positive (head-final) and negative (head-initial) dependencies, thus transforming the sentence internal dependency link into global features of the treebank. He found that these features are relatively efficient for the language classification task, thus being able to dig out human language universals from authentic data.

[1] With the exception of the ParTUT treebanks (Sanguinetti & Bosco 2011).

[2] Contrary to all previous analyses of prepositions in Indo-European languages that we are aware of which see the prepositions as governors of the following noun (giving a *PP* its name), UD annotates prepositions as case markers of the noun, independently of whether it is sub-categorized by the verb (*talk to*) or semantically full (*sleep under*). This leads to a greater structural similarity between English and Turkish than typologically expected and also for example to competing annotations of complex prepositions (*on top of*) in the current treebanks (with *top* as the head of the PP or as a dependent of the embedded noun).

[3] Examples of this type of correlations include the tendency of *O-V* languages to be postpositional, placing adpositions after their objects – while inversely *V-O* languages tend to be prepositional, placing adpositions before their objects. So the *V-O* vs *O-V* feature is correlated with the *preposition* vs. *postposition* feature.

Subsequent empirical studies of language classification have confirmed that combined measures on all dependency links, not only on the verbal and nominal arguments, provides better typological indicators than one or several specific word order measures, which may lead to conflicting conclusions (Liu & Xu 2012). In addition, macroscopic indexes, such as network parameters of dependency treebanks based on language networks, have been shown to perform even better than global measures of word order (Liu & Li 2010; Abramov & Mehler 2011; Liu & Xu 2011, 2012; Liu & Cong 2013). One way of extracting global structural language features is to fuse all equal lexical nodes, resulting in one big syntactic network where every lexical node appears only once (Ferrer-i-Cancho 2001, Chen et al. 2015). In the present work, we completely strip the treebank of the lexical nodes, taking into account only the categories as well as the frequency and directions of the dependency relations.

Although word order is clearly an important index for capturing the typological features of languages, we suspect that it is possible to refine the index by combining it with additional information or to conceive indeces that are better adapted to the classification task, such as network parameters. In the present work, we propose two means of modifying the word order (dependency direction) index for language classification task. To the quantitative measure of the dependency directions, we add the length of the syntactic relations (Liu 2008; Liu et al. 2009), i.e. we compute the *Directional Dependency Distances* (DDD) for each syntactic function with positive/negative values corresponding to the dependency direction. This DDD measure appears to be a straightforward choice of quantitative values that map directly to the dependency direction index.

Although our method follows the same 'quantitative' principle as Liu (2010) and Lu & Xu (2012), it contains different information. Instead of using the distribution percentages of the dependency directions to quantify them, we add the distance information into it and thus create a more integrated value rather than a pure direction index. The second novelty of this work is a more fine-grained dependency direction measurement: Instead of computing an overall value (the average distance or the percentage of positive relations for a whole treebank), the unified annotation scheme of UD allows us to break down the frequency, direction, and length of the links by dependency relationship. Common clustering techniques will allow analyzing and visualizing language similarities.

1.3 Outline

Following the idea of investigating the typological structural universality and diversity of languages based on authentic treebank data, the present work specifically focuses on whether and how the UD treebank set allows us to recognize language families based on purely empirical structural data. The question can be decomposed into various sections:

The following section will describe the dataset used in this study and the principal measures that we apply. In section 3 we start with a global unidimensional measure that imposes a natural order on the set of treebanks. We compare the measure we propose to existing work. Given the abovementioned series of problems of the underlying treebank data, we then move on to assessing whether the current UD data is actually good enough to measure structural differences, the most evident method being whether different treebanks of the same language are actually structurally more similar to each other than to treebanks of other languages. For this, we apply our ranking to the individual treebanks as well as to the data combined by language.

We then make use of the common annotation scheme of the UD treebanks which allows us to split up the measures per syntactic functions. This multi-dimensional dataset can be used for common clustering techniques, whose results we present and discuss. We will conclude with a discussion of the results, problems, and future plans of dependency-based typology.

Our images contain very small fonts but the image resolution allows zooming in. For the PCA images, the color zones, which we describe in the text, are generally sufficient for the understanding of the clusters. Since we compute data on close to 50 languages, 70 treebanks, and 30 dependency relations, we cannot provide all numerical data in the Annex of this paper. All scripts, data, and images are freely available on https://gerdes.fr/papiers/2017/dependencyTypology/ thus allowing reproducing our results, in particular as the underlying UD treebank target is a fast moving target.

2 Methods

The main analysis includes four main steps: 1) data selection and description, 2) determina-

tion and extraction of the parameters to investigate, 3) quantitative description of the parameters, 4) clustering analysis based on measurements of step 3.

In step 1, we remove the relatively sparse languages, namely treebanks with less than 10,000 tokens, from the dataset. We also only kept ***syntagmatic core relations***, removing *fixed*, *flat*, *conj*, and *root* relations from our distance measures as their direction and length are universally fixed in the annotation guide and don't indicate any interesting difference between languages.[4] Different treebanks of the same language are firstly kept separate for consistency measures and secondly combined for the main classification tasks.

In step 2, we extract dependency function distribution, direction, and distance measures from the combined treebanks. More specifically, we compute the relative frequency distributions of dependency functions and the ***Directional Dependency Distance (DDD)***, which we define as the product of the dependency distance and the direction, thus including negative values. We obtain three different central observations, as shown in Table 1, which we will also compare to other frequency measures.

Observations	Distributions (frequency)	Directional dependency distance (DDD)
1	√	×
2	×	√
3	√	√

Table 1: 3 observations based on 2 parameters

For Observation 1, we only look at the distributions of dependency functions of different languages. For the observation 2, we compute the *DDD* per syntactic function by computing the difference of the node index and the governor index for each node, adding those values up and dividing by the number of links[5]. The DDD of a dependency relation R is thus defined as follows:

$$DDD(R) = \frac{\sum_{r \in R} distance(r)}{frequency(R)}$$

In the third measure, we quantify this average *DDD* by means of the relative frequency of the UD functions by multiplying *DDD* with the relative frequency of the corresponding function.

At step 3, we conduct clustering analyses based on the data of the observations. We compare the results of these three observations to each other as well as to previous language classification studies to see whether they can distinguish different known language families in order to assess which observations provide the best result.

3 Unidimensional measures

To start, let us first look at the simple measures, where we get a unique numerical value per treebank or language.

We computed the DDD of all dependency relations combined. The DDD takes head-final relations as negative values and head-initial relations as positive values. Languages that have an equal number of left-spanning and right-spanning links of similar average length, will have a value close to zero.

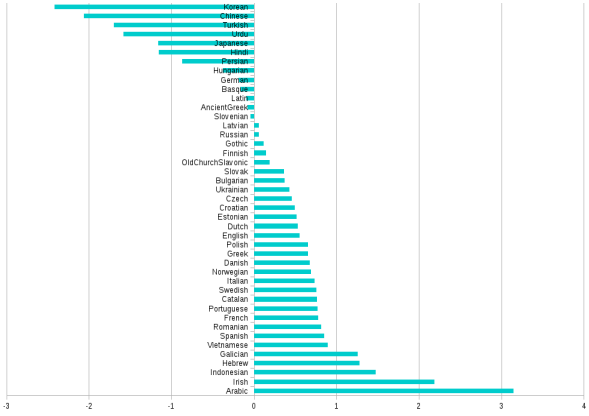

Figure 1: Languages ordered by dependency distance

This graph gives a good idea of what kind of insights we want to gain from dependency mea-

[4] *Fixed* and *flat* are used for multi-word expressions, *conj* for coordinations. These three dependency relations have arbitrarily been assigned to a left-to-right bouquet structure (all subsequent tokens depend on the first token). See Gerdes & Kahane 2016 for a description and for alternatives to this choice. The *root* link is often thought of and drawn as a line straight up from the root node but it is encoded in CoNLL as a link to the zero node. Taking the root "length" into account would artificially add left-right relations to mainly head-final language (and the way around), thus lowering the average distance measures.

[5] This means that we do not take into account the variance of these links, e.g. a language that has symmetric links around each governor will have a zero distance, independently of the length of these dependency links. We also computed the standard deviation of each relation and included this value in the clustering, but this did not significantly improve the result.

sures: It comes as no surprise to find Korean at the top of the list of the most centripetal (Tesnière's term for head-final[6]) languages, and, inversely, Arabic at the bottom of the list, being the most centrifugal of the analyzed languages. The appearance of Chinese, however, between Korean and Turkish at the second place affirms how strictly head-final Chinese actually is – a fact that does not really show when classifying languages in the discrete categories of SVO, SOV etc. We see how the numerical analysis allows for new empirically-based groupings and ordering of languages that are hard to perceive on purely categorical classifications.

The Germanic language group is spread across the spectrum, starting from the negatively distanced German to the highly positively distanced Swedish. The Romance languages, however, are all very well clustered around an average distance of about 0.8.

Compare this with a measure that does not take into account the actual length of the dependencies but only the direction percentages (proposed by Liu (2010)):

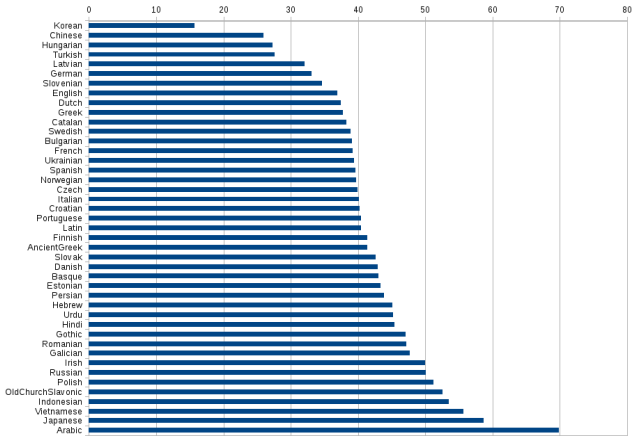

Figure 2: Languages ordered by % of positive links

Although the two extremes (Korean and Arabic) are the same, the results correspond less to well-known language classifications. Observe how Japanese finds its natural position close to Korean, Turkish and Hungarian in the DDD measure, whereas the direction percentage measure places it right next to Arabic, presumably because of the high number of (postpositional) particles.[7]

[6] Tesnière's language classification terminology (1959) precedes Greenberg's by 4 years but was not cited by the latter.

[7] Although functionally analogous, equivalent postpositions are traditionally seen as morphological case-marking in Korean. This leads to quite diverse treebanks for structurally similar languages, calling for a more precise tokenization specification.

3.1 Corpus or Language differences?

A basic coherence measure of our data can be done by comparing not languages as a whole but treebanks which have usually been created by different groups of developers. If we encounter strong differences among treebanks of the same language that genre differences cannot account for, then this points to underspecification of guidelines – or possibly to systematic errors in one treebank.

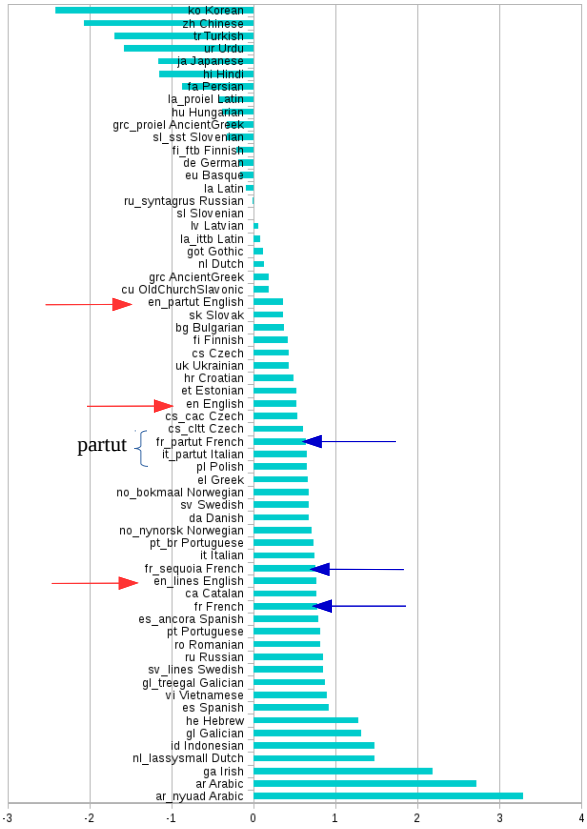

Figure 3: Treebanks ordered by dependency distance with positions for English and French. The language names are preceded by the ISO language code and the complete treebank name if there is more than one treebank per language.

The separation of our data by treebank generally puts languages at similar positions independently of the treebank. Nevertheless, this also reveals some of the aforementioned incoherences of the current state of the annotations – and thus also the limits of our approach. The following figure indicates the different places taken by the English (left side, red arrows) and the French (right side, blue arrows) treebanks of UD 2.0. Although the absolute values are not as extremely different as the position suggests (en: 0.4, 0.5, 0.8; fr: 0.6, 0.8, 0.8), any derived typological

classification seems to remain quite treebank dependent at the current state of UD. Note also that the treebanks from the ParTUT team coherently have a lower dependency direction than their counterparts for English, French, and Italian. It is tempting to attribute this difference to differences in the guidelines used by different teams in the annotation process, but for Italian, the other Italian treebank has also been created by the ParTUT team. So maybe the difference is rather due to the syntactic structure of "Translationese", that has shorter dependency links for the mostly head-initial languages included in ParTUT.

More generally, this shows how these methods also allow for detecting common ground and outliers in the process of treebank development. They can be used for error-mining the treebank.

4 Multi-dimensional clustering

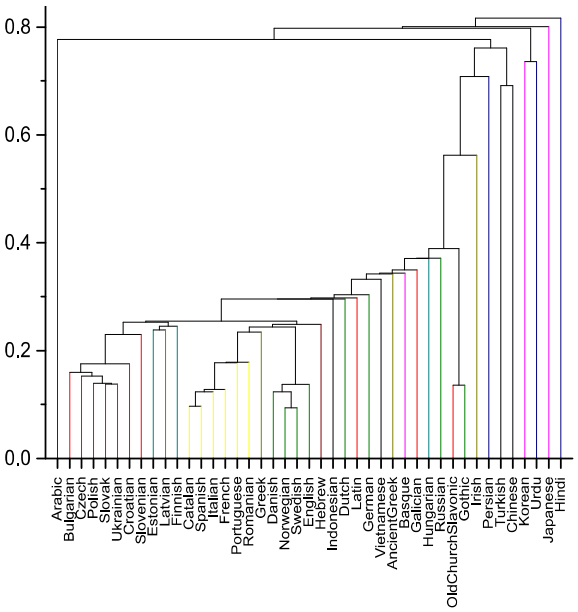

Figure 4: Dendrogram of DDD vectors per function

Measures on our set of treebanks that distinguish dependency relations give rise to multi-dimensional vectors. The clustering analysis can be done by the usual Principal Component (PCA[8]) and the Hierarchical Cluster Analysis (HCA[9]).

UD allows the introduction of idiosyncratic sub-classes of syntactic functions. English, for example has the *nmod:poss* function, the possessive subclass of nominal modifiers used for the annotation of genitives. To make the values comparable, we are measuring the direction and distribution of simple functions, i.e. function names stripped of what follows the colon.

4.1 Directional Dependency Distance (DDD) by syntactic function

Instead of comparing the single DDD value, we can use the whole vector of DDDs, one for each of the 33 syntactic functions. Contrarily to what we have seen for the global DDD, the multi-dimensional HCA clustering of Figure 4 groups relatively correctly: the Slavic language family (red, except Russian), Romance (yellow, without Galician) and Germanic (green, without German and Dutch).

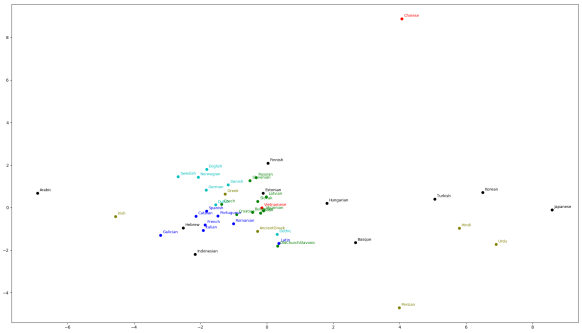

Figure 5: PCA of DDD vectors per function

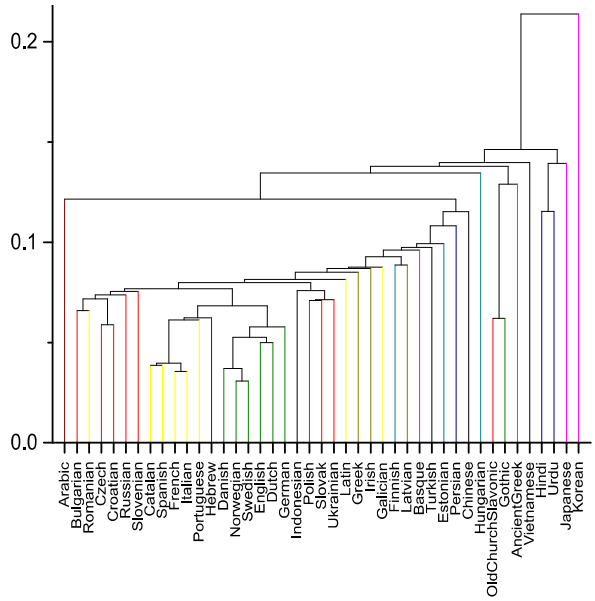

Figure 6: Dendrogram of relative frequencies of dependency relations

The PCA of the same data provides clustering of comparable quality, cf. Figure 5: Romance in blue, Germanic in turquoise, and, less clearly clustered, Slavic in green. Note also the rectangle containing Altaic languages in the following order but quite fare from one another: Hungarian, Turkish, Korean, and Japanese.

[8] The PCAs are performed with the decomposition package of the scikit-learn project. See github.com/scikit-learn/scikit-learn

[9] The HCA in this paper are conducted by OriginPro 9 (Cluster Method: Nearest neighbor, Distance Type: Euclidean).

4.2 Clustering relative frequency distributions

Do we actually need to take into account the length and direction of the dependency relation to obtain correct language families? Or will the simple frequency of dependency relation labels do? Figure 6, another clustering analysis, only on the relative frequency of each dependency label, shows that the analysis successfully distinguishes Indo-European languages and also obtains rather good results for three big sub-groups, namely the Germanic branch (in green color), Italic branch (in yellow color), and the Slavic branch (in red color). Although some intermingling of these three branches still exists, the result is slightly better than the result that we obtain based on simple dependency directions.

It is noteworthy that we cannot further simplify the underlying data and dispense with the tree structure altogether. If for example we only use relative POS frequencies, we obtain an PCA analysis where language groups are not coherent clusters (Figure 7).

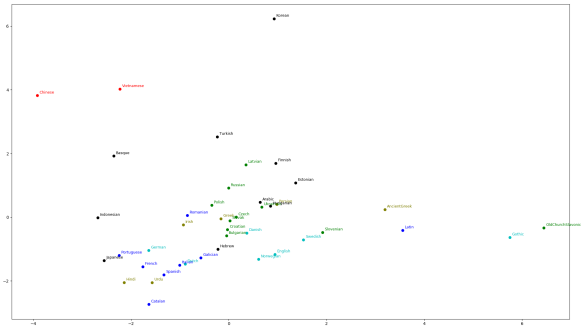

Figure 7: PCA of POS frequencies

Inversely, complexifying the features gives sparse data and unrecognizable results. If, for example, we combine function and category and measure the frequencies of function-category couples, one couple being for example (*nsubj→NOUN*), we obtain the following uninterpretable graph (Figure 8). Although many of UD's syntactic functions are actually redundant (*nsubj* contains the information that the dependent is a NOUN), the higher-dimensional space projects less clearly into two-dimensional space (~500 dimensions), presumably because of data sparsity. This experiment could be redone when some UD treebanks will have attained a significantly greater size.

Note that in both the pure POS and the function-POS analysis, the two ancient languages Gothic and Old Church Slavonic are strong outliers (on the right of the graph), not far from Ancient Greek and Latin. This suggests that the

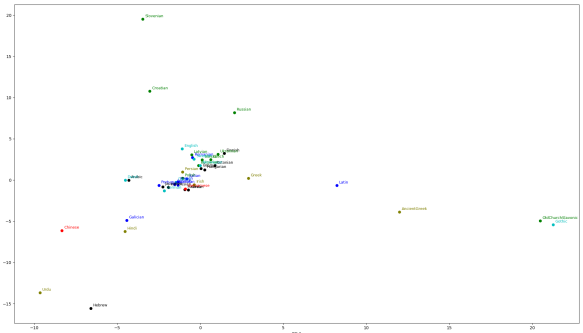

Figure 8: PCA of function-POS frequencies

POS annotations of these languages has been done by the same team or at least has been under mutual influence.

This shows that all measures are not created equal. The actual structural information of the treebank is crucial to obtain satisfying language groups.

4.3 DDD multiplied by relative frequency

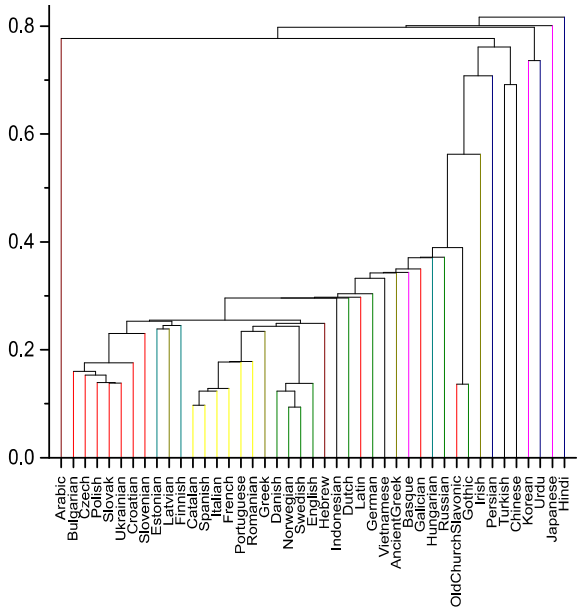

Figure 9: Dendrogram of distance × frequency clustering per language

Both the pure frequency measures and the directional dependency measures (DDD) measures give interesting results. When combining these two measures by multiplying the DDD by the relative frequencies, we obtain even more satisfying results: Figure 9 shows a first red subtree corresponding to Slavic langugaes, only Latvian, Russian, and Old Slavonic being outliers. The next yellow subtree hosts Romance language with Latin and Galician later following alone. The green sub-tree shows the proximity of the Germanic languages Danish, Norwegian, Swedish, and English – with Dutch and German following separately. As in the PCA analysis, Old Slavonic and Gothic form again a close sub-

group – presumably due to a common annotation process.

Even when grouping by treebanks and not by languages, the subtrees cut neatly into the set of languages. In Figure 9, the red subtree on the left groups together nearly all Slavic languages, the yellow subtree contains nearly all Romance languages, and the green subtree most Germanic languages (see the Annex for the names of the language codes). Then there is another separate green subtree for German and Dutch and two more Germanic outliers: Gothic and another Dutch corpus. If this is not a genre difference, we can suppose that this Dutch Lassymal UD treebank follows different annotation guidelines. Note also how close are Finnish and Estonian (small light brown subtree). This subtree then groups together with Latvian, a language considered coming from a different group of languages. This structural similarity mimicking geographic proximity is an interesting result suggesting cross-language-group influences not only on the lexicon but also on the syntactic structure itself.

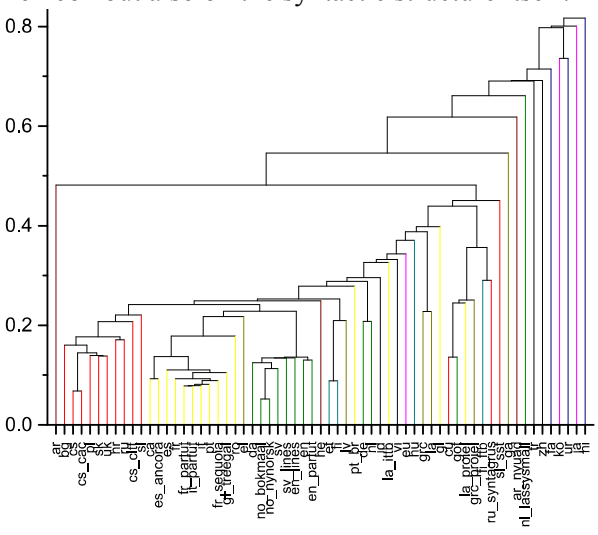

Figure 10: Dendrogram of distance × frequency clustering per corpus

Similarly, note that the distance × frequency measures consistently cluster Romanian in the Romance language group, but simple relative frequency measures show Romanian close to Bulgarian and other Slavic languages. In a sense, the simple frequency captured some features of language groups better than DDD and the multiplied values. We have to leave it to further research to determine which kind of proximity is better captured by which measure.

We can see that a well-chosen measure, here the combined frequency and distance measure, can abstract away from the many annotation errors and incoherences of the current UD.

Even using PCA on the language treebank data (Figure 11), we see that the right hand side of the PCA diagram contains the same languages as the most independent languages of the dendrogram: Japanese (black dot to the right) Chinese (red on top), Hindi, Korean, and Urdu stand out the furthest from the crowd in both projections, showing the relative robustness of the data concerning the actual choice of the clustering technique.

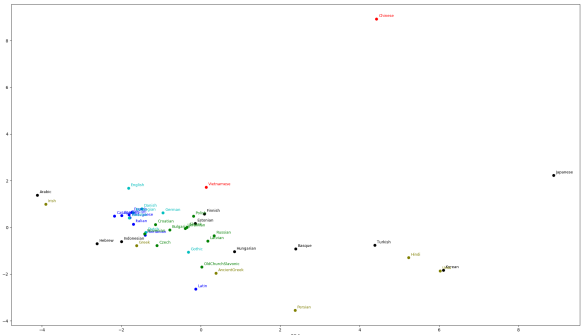

Figure 11: PCA of distance × frequency

5 Conclusion

The various data extraction and clustering techniques that we have carried out, only the most emblematic of which we could present in this paper, show that the UD treebanks succeed rather well for language classification even if we solely base our study on the delexicalized tree structures. The coherent cross-language annotation scheme makes it possible to split up the measures by dependency functions. Although modern language typology studies are mainly focused on word order, the different measures and methods we proposed show that the classical word order classification alone is no longer sufficient to classify languages based on authentic clustering data, which is a similar result to Liu (2012). Usually we get better results if we consider the actual dependency relations, no matter under which format: relative distribution, network, and network variations. For single parameters alone, the dependency relationship distribution is performing better than the dependency direction. However, combining the criteria provides us with the best language clustering results attainable on the sole basis of syntactic treebanks.

Meanwhile, it is necessary to further assess in future research the robustness of our clustering approach to typology across different annotation schemes, for instance by comparing the UD treebanks with data that can be obtained from crosslingual parsers (Ammar et al. 2016; Guo et al. 2016).

Since the distribution of dependency relationships is very uneven and the majority of links consists of a small subset of all types, it seems possible that the most frequent relations are sufficient for classifying languages. If they are, then some functions may have different effects on the clustering process. The decisive functions in the clustering represent language diversity, the others have a more universal character. This process transforms the categorical opposition between principles and parameters into a gradual scale where syntactic features and constructions can be positioned based on empirical data from treebanks.

A basic epistemological question arises from two types of results that we can obtain in our approach: We have measures that group languages according to well-known classes, and measures that show new groupings and relationships. Both results are interesting, the latter requiring further explorations and explanations – and, as in any truly empirical approach, it requires returning to the data to ascertain the actual causes of the observed distances between treebanks.

Here we encounter the difficulty of assessing the nature of the results: Are they possibly due to annotation errors and incoherences? Are they due to genre differences of the underlying texts? The methodology we propose will grow and improve with the coherence of the UD treebanks. – Or possibly with the emergence of other more syntactically oriented treebank collections, in particular if they are conceived as parallel treebanks, with identical genres. This would dispel any doubts on clustering results, as each cluster would solely and directly express an empirical typological relation.

References

Abramov, Olga, and Alexander Mehler. "Automatic language classification by means of syntactic dependency networks." *Journal of Quantitative Linguistics*, 18.4 (2011): 291-336.

Chen, Xinying, Haitao Liu, and Kim Gerdes. "Classifying Syntactic Categories in the Chinese Dependency Network." *Depling 2015* (2015): 74.

Croft, William. *Typology and universals*. Cambridge University Press, 2002.

De Marneffe, Marie-Catherine, et al. "Universal Stanford dependencies: A cross-linguistic typology." *LREC*. Vol. 14. 2014.

Dryer, Matthew S. "The Greenbergian word order correlations." *Language*, (1992): 81-138.

Ferrer-i-Cancho, Ramon, and Richard V. Solé. "The small world of human language." *Proceedings of the Royal Society of London B: Biological Sciences*, 268.1482 (2001): 2261-2265.

Gerdes, Kim, and Sylvain Kahane. "Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies." *LAW X (2016)*

Greenberg, Joseph H. "Some universals of grammar with particular reference to the order of meaningful elements." *Universals of language*, 2 (1963): 73-113.

Haspelmath, Martin. *The world atlas of language structures*. Vol. 1. Oxford University Press, 2005.

Ledgeway, Adam. "Syntactic and morphosyntactic typology and change." *The Cambridge history of the Romance languages*, 1 (2011): 382-471.

Liu, Haitao. "Dependency distance as a metric of language comprehension difficulty." *Journal of Cognitive Science*, 9. 2 (2008): 159-191.

Liu, Haitao. "Dependency direction as a means of word-order typology: A method based on dependency treebanks." *Lingua*, 120.6 (2010): 1567-1578.

Liu, Haitao, and Chunshan Xu. "Can syntactic networks indicate morphological complexity of a language?." *EPL (Europhysics Letters)*, 93.2 (2011): 28005.

Liu, Haitao, and Chunshan Xu. "Quantitative typological analysis of Romance languages." *Poznań Studies in Contemporary Linguistics PsiCL*, 48 (2012): 597-625.

Liu, Haitao, and Jin Cong. "Language clustering with word co-occurrence networks based on parallel texts." *Chinese Science Bulletin*, 58.10 (2013): 1139-1144.

Liu, Haitao, Richard Hudson, and Zhiwei Feng. "Using a Chinese treebank to measure dependency distance." *Corpus Linguistics and Linguistic Theory*, 5.2 (2009): 161-174.

Liu, Haitao, and Wenwen Li. "Language clusters based on linguistic complex networks." *Chinese Science Bulletin*, 55.30 (2010): 3458-3465.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale.* Klincksieck, Paris.

Petrov, Slav, Dipanjan Das, and Ryan McDonald. "A universal part-of-speech tagset." *arXiv preprint arXiv:1104.2086*, (2011).

Sanguinetti M, Bosco C. "Building the multilingual TUT parallel treebank". *Proceedings of The Second Workshop on Annotation and Exploitation of Parallel Corpora* 2011 Sep 15 (p. 19).

Song, Jae Jung. *Linguistic typology: Morphology and syntax*. Routledge, 2014.

Zeman, Daniel. "Reusable Tagset Conversion Using Tagset Drivers." *LREC*. 2008.

Appendix A. *Selected Language Data*

Our study is based on the UD 2.0 treebanks of 43 languages combining 67 corpora.

As an example, we provide a table with the (alphabetically) first functions of rounded DDD data per language:

name	acl	advcl	advmod	amod	appos	aux
Arabic	3,37	9,87	3,42	1,39	3,43	-1,05
Bulgarian	5,07	2,73	-1,33	-1,09	2,58	-1,32
Catalan	5,51	7,41	-1,24	0,89	5,26	-1,45
Czech	5,58	1,72	-1,22	-0,97	4,83	-2,14
Old Church Slavonic	2,37	0,02	-0,97	0,66	1,63	0,79
Danish	5,42	5,15	-0,24	-0,63	2,59	-2,31
German	9,9	7,47	-1,84	-1,17	2,29	-4,54
Greek	4,25	4,01	-1,04	-1,08	5,67	-1,14
English	3,48	2,4	-0,93	-1,16	4,07	-1,58
Spanish	4,94	6,11	-1,16	0,7	3,45	-1,5
Estonian	2,07	3,39	-0,63	-1,04	2,84	-1,98
Basque	-1,83	-0,03	-1,93	0,43	4	0,78
Persian	7,81	-4,98	-5,66	0,95	2,81	-1,64
Finnish	1,4	2,24	-0,56	-1,19	2,96	-1,66
French	3,72	4,59	-1,17	0,65	3,2	-1,46
Irish	3,13	8,37	1,88	1,3	4,59	0
Galician	4,33	5,07	-1,06	0,78	5,14	-1,31
Gothic	3,35	1,04	-1,09	0,17	2,34	0,96
Ancient Greek	4,6	-0,52	-1,91	0,37	3,66	-1,73
Hebrew	4,53	2,83	-0,33	1,8	4,15	-1,96
Hindi	3,73	-5,67	-2,35	-1,32	0	1
Croatian	4,55	2,99	-1,48	-1,2	2,34	-1,54
Hungarian	8,67	4,22	-2,26	-1,39	3,67	0
Indonesian	3,81	4,65	-1,15	1,25	3,7	-1,33
Italian	3,84	2,46	-1,51	0,53	4,98	-1,32
Japanese	-6,35	0	-8,99	-1,43	0	1,76
Korean	-1,55	-5,22	-3,26	-1,08	-6,52	0
Latin	3,55	0,85	-2,33	0,1	3,5	0,55
Latvian	3,41	1,52	-1,5	-1,42	5,67	-1,11
Dutch	5	4,39	-1,67	-1,07	2,27	-2,62
Norwegian	3,77	3,71	-0,67	-0,94	4,79	-1,77
Polish	4,7	1,85	-1,13	-0,34	1,7	0,05
Portuguese	4,37	3,76	-1,29	0,46	3,68	-1,43
Romanian	4,13	3,37	-1,21	1	4,95	-1,21
Russian	4,19	3,07	-1,17	-1,05	2,31	-0,89
Slovak	4,57	1,73	-1,14	-1,06	3,68	-0,64
Slovenian	5,77	1,04	-1,28	-1,17	3,35	-2,35
Swedish	3,66	3,06	-0,64	-1,07	5,6	-1,95
Turkish	-2,46	0	-1,05	-1,9	2,11	1,35
Ukrainian	4,06	2,15	-1,28	-1,19	2,22	-0,65
Urdu	5,84	-3,73	-6,4	-1,43	0	1
Vietnamese	0	-3,61	-0,66	1,18	3,83	-0,77
Chinese	-4,88	-8,17	-2,5	-2,18	1,5	-2,67

The unabridged data used in this paper is available on https://gerdes.fr/papiers/2017/dependencyTypology/

code	Language	tokens
ar	Arabic	233, 712
ar_nyuad	Arabic	670, 612
bg	Bulgarian	123, 178
ca	Catalan	417, 453
cs	Czech	1, 174, 076
cs_cac	Czech	426, 274
cs_cltt	Czech	22, 000
cu	Old Church Slavonic	39, 394
da	Danish	80, 351
de	German	245, 524
el	Greek	47, 343
en	English	194, 428
en_lines	English	58, 223
en_partut	English	34, 195
es	Spanish	377, 020
es_ancora	Spanish	443, 951
et	Estonian	29, 051
eu	Basque	82, 516
fa	Persian	113, 699
fi	Finnish	152, 583
fi_ftb	Finnish	118, 747
fr	French	349, 973
fr_partut	French	16, 328
fr_sequoia	French	53, 635
ga	Irish	11, 627
gl	Galician	105, 844
gl_treegal	Galician	13, 819
got	Gothic	37, 931
grc	Ancient Greek	161, 184
grc_proiel	Ancient Greek	171, 524
he	Hebrew	127, 018
hi	Hindi	262, 007
hr	Croatian	161, 533
hu	Hungarian	27, 607
id	Indonesian	82, 588
it	Italian	254, 058
it_partut	Italian	38, 768
ja	Japanese	149, 147
ko	Korean	43, 921
la	Latin	15, 978
la_ittb	Latin	254, 683
la_proiel	Latin	134, 030
lv	Latvian	38, 476
code	Language	tokens
nl	Dutch	170, 665
nl_lassysmall	Dutch	73, 373
no_bokmaal	Norwegian	243, 529
no_nynorsk	Norwegian	240, 917
pl	Polish	63, 236
pt	Portuguese	196, 032
pt_br	Portuguese	260, 983
ro	Romanian	177, 755
ru	Russian	78, 025
ru_syntagrus	Russian	872, 362
sk	Slovak	79, 704
sl	Slovenian	113, 498
sl_sst	Slovenian	16, 389
sv	Swedish	65, 954
sv_lines	Swedish	56, 661
tr	Turkish	37, 167
uk	Ukrainian	11, 312
ur	Urdu	99, 024
vi	Vietnamese	25, 979
zh	Chinese	103, 614