Word Order Typology through Multilingual Word Alignment

Robert Östling
Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
robert@ling.su.se

Abstract

With massively parallel corpora of hundreds or thousands of translations of the same text, it is possible to automatically perform typological studies of language structure using very large language samples. We investigate the domain of word order using multilingual word alignment and high-precision annotation transfer in a corpus with 1144 translations in 986 languages of the New Testament. Results are encouraging, with 86% to 96% agreement between our method and the manually created WALS database for a range of different word order features. Beyond reproducing the categorical data in WALS and extending it to hundreds of other languages, we also provide quantitative data for the relative frequencies of different word orders, and show the usefulness of this for language comparison. Our method has applications for basic research in linguistic typology, as well as for NLP tasks like transfer learning for dependency parsing, which has been shown to benefit from word order information.

1 Introduction

Since the work of Greenberg (1963), word order features have played a central role in linguistic typology research. There is a great deal of variation across languages, and interesting interactions between different features which may hint at cognitive constraints in the processing of human language. A full theoretical discussion on word order typology is beyond the scope of this paper, but the interested reader is referred to e.g. Dryer (2007) for an overview of the field.

This study uses multilingual word alignment (Östling, 2014) and high-precision annotation projection of part-of-speech (PoS) tags and dependency parse trees to investigate five different word order properties in 986 different languages, through a corpus of New Testament translations. The results are validated through comparison to relevant chapters in the World Atlas on Language Structures, WALS (Dryer and Haspelmath, 2013), and we find a very high level of agreement between this database and our method.

We identify two primary applications of this method. First, it provides a new tool for basic research in linguistic typology. Second, it has been shown that using these word order features leads to increased accuracy during dependency parsing model transfer (Täckström et al., 2013). These benefits can now be extended to hundreds of more languages. The quantified word order characteristics computed for each of the 986 languages in the New Testament corpus, including about 600 not in the WALS samples for these features, are available for download.[1]

2 Related work

Using parallel texts for linguistic typology has become increasingly popular recently, as massively parallel texts with hundreds or thousands of languages have become easily accessible through the web (Cysouw and Wälchli, 2007; Dahl, 2007; Wälchli, 2014). Specific applications include data-driven language classification (Mayer and Cysouw, 2012) and lexical typology (Wälchli and Cysouw, 2012). However, unlike our work, none of these authors developed automatic methods for studying syntactic properties like word order, nor did they utilize recent advances in the field of word alignment algorithms.

[1] http://www.ling.su.se/acl2015-wordorder.zip

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 205–211,
Beijing, China, July 26-31, 2015.

3 Method

The first step consists of using supervised systems for annotating the source texts with Universal PoS Tags (Petrov et al., 2012) and dependency structure in the Universal Dependency Treebank format (McDonald et al., 2013). For PoS tagging, we use the Stanford Tagger (Toutanova et al., 2003) followed by a conversion step from the Penn Treebank tagset to the "universal" PoS tags using the tables published by Petrov et al. Next, we use the MaltParser dependency parser (Nivre et al., 2007) trained on the Universal Dependency Treebank using MaltOptimizer (Ballesteros and Nivre, 2012).

The corpus is then aligned using the multilingual alignment tool of Östling (2014). This model learns an "interlingua" representation of the text, in this case the New Testament, to which all translations are then aligned independently. An interlingua sentence $\boldsymbol{e}$ is assumed to generate the corresponding sentences $\boldsymbol{f}^{(l)}$ for each of the L languages through a set of alignment variables $\boldsymbol{a}^{(l)}$ for each language. This can be seen as a multilingual extension of the IBM model 1 (Brown et al., 1993) with Dirichlet priors (Mermer and Saraçlar, 2011), where not only the alignment variables are hidden but also the source $\boldsymbol{e}$. The probability of a sentence and its alignments (in L languages) under this model is

$$P(\boldsymbol{a}^{(1\ldots L)},\boldsymbol{f}^{(1\ldots L)}|\boldsymbol{e}) = \prod_{l=1}^{L}\prod_{j=1}^{J} p_t(f_j^{(l)}|e_{a_j^{(l)}}) \cdot \prod_{i=1}^{I} p_c(e_i) \quad (1)$$

where the translation distributions p_t are assumed to have symmetric Dirichlet priors and the source token distribution p_c a Chinese Restaurant Process prior. Given the parallel sentences $\boldsymbol{f}^{(1\ldots L)}$, then $\boldsymbol{a}^{(1\ldots L)}$ and $\boldsymbol{e}$ are sampled using Gibbs sampling. The advantage of this method is that the multi-source transfer can be done once, to the interlingua representation, then transferred in a second step to all of the 986 languages investigated. It would be possible to instead perform 986 separate multi-source projection steps, but at the expense of having to perform a large number of bitext alignments.

From the annotated source texts, PoS and dependency annotations are transferred to the interlingua representation. Since alignments are noisy and low recall is acceptable in this task, we use an aggressive filtering scheme: dependency links must be transferred from at least 80% of source texts in order to be included. For PoS tags, which are only used to double-check grammatical relations and should not impact precision negatively, the majority tag among aligned words is used. Apart from compensating for noisy alignments and parsing errors, this method also helps to catch violations against the direct correspondence assumption (Hwa et al., 2002) by filtering out instances where different source texts use different constructions, favoring the most prototypical cases. Each word order feature is coded in terms of dependency relations, with additional constraints on the parts of speech that can be involved. For instance, when investigating the order between nouns and their modifying adjectives we look for an AMOD dependency relation between an ADJ-tagged and a NOUN-tagged word, and note the order between the adjective and the noun. This method rests on the assumption that translation equivalents have the same grammatical functions across translations, which is not always the case. For instance, if one language uses a passive construction where the source texts all use the active voice, we would obtain the wrong order between subject and object.

To summarize, our algorithm consists of the following steps:

1. Compute an interlingua representation of the parallel text, as well as word alignments linking it to each of the translations.

2. Annotate a subset of translations with PoS tags and dependency structure.

3. Use multi-source annotation projection from this subset to the interlingua representation, including only dependency links where the same link is projected from at least 80% of the source translations.

4. Use single-source annotation projection from the interlingua representation to each of the 986 translations.

5. For each construction of interest, and for each language, count the frequency of each ordering of its constituents.

4 Evaluation

We evaluate our method through comparison to the WALS database (Dryer and Haspelmath,

SOV	SVO	OSV	OVS	VSO	VOS
Polynesian (Hawaiian, Maori)					
3	31	2	2	70	3
6	26	5	4	76	18
Sinitic (Mandarin, Hakka)					
54	235	6	0	3	5
18	84	1	2	5	3
Turkic (Kara-Kalpak, Kumyk)					
114	2	8	7	0	0
89	1	12	11	4	1

Table 1: Number of transitive clauses with a given order of subject/object/verb, according to our algorithm, for six languages (from three families).

2013), by manual analysis of selected cases, and by cluster analysis of the word order properties computed for each language by our method.

4.1 Data and methodology

A corpus of web-crawled translations of the New Testament was used, comprising 1144 translations in 986 different languages. Of these, we used five English translations as source texts for annotation projection. Ideally more languages should be used as sources, but since we only had access to complete annotation pipelines for English and German we only considered these two languages, and preliminary experiments using some German translations in addition to the English ones did not lead to significantly different results. A typologically more diverse set of source languages would help to identify those instances in the text which are most consistently translated across languages, in order to reduce the probability that peculiarities of the source language(s) will bias the results.

In order to evaluate our method automatically, we used data from the WALS database (Dryer and Haspelmath, 2013) which classifies languages according to a large number of features. Several features concern word order, and we focused on five of these (listed in Table 2). Only languages which are represented both in the New Testament corpus and the WALS data were used for the evaluation. In addition, we exclude languages for which WALS does not indicate a particular word order. This might be due to e.g. lacking adpositions altogether (which makes the adposition/noun order of that language undefined), or because no specific order is considered dominant.

The frequencies of all possible word orders for a feature are then counted, and for the purpose of evaluation the most common order is chosen as the algorithm's output. Although the relative frequencies of the different possible word orders are discarded for the sake of comparability with WALS, these frequencies are themselves an important result of our work and tell a much richer story of the word order properties (see Table 1 and Figure 1).

Counting the number of instances (token frequency) of each word order is the most straightforward way to estimate the relative proportions of each ordering, but the results are biased towards the behavior of the most frequent words, which often have idiosyncratic, non-productive features. Therefore, we also compute the corresponding statistics where each type is counted only once for each word order it participates in, disregarding its frequency. The type-based counts should better capture the behavior of productive patterns in the language. For the purpose of this study, we define the type of our relations as follows:

- **adjective-noun**: the form of the adjective
- **adposition-noun**: the forms of both adposition and noun
- **verb-(subject)-(object)**: the form of the verb

For instance, given the following three sentences: "we see him," "I see her" and "them I see", we would increase the count by one for SVO order and for OVS order, because these are the orders in which the verb *see* has been observed to participate.

In cases where there are multiple translations into a particular language, information is aggregated from all these translations into a single profile for the language. This is problematic in some cases, such as when a very long time separates two translations and word order characteristics have evolved, or simply due to different translators or source texts. However, since the typical case is a single translation per language, and WALS only contains one data point per language, we leave inter-language comparison to future research.

4.2 Results and Discussion

Table 1 shows how the output of our token-based algorithm looks for three pairs of languages selected from different families. The absolute counts vary due to our filtering procedure and differing numbers of translations, but as we might expect

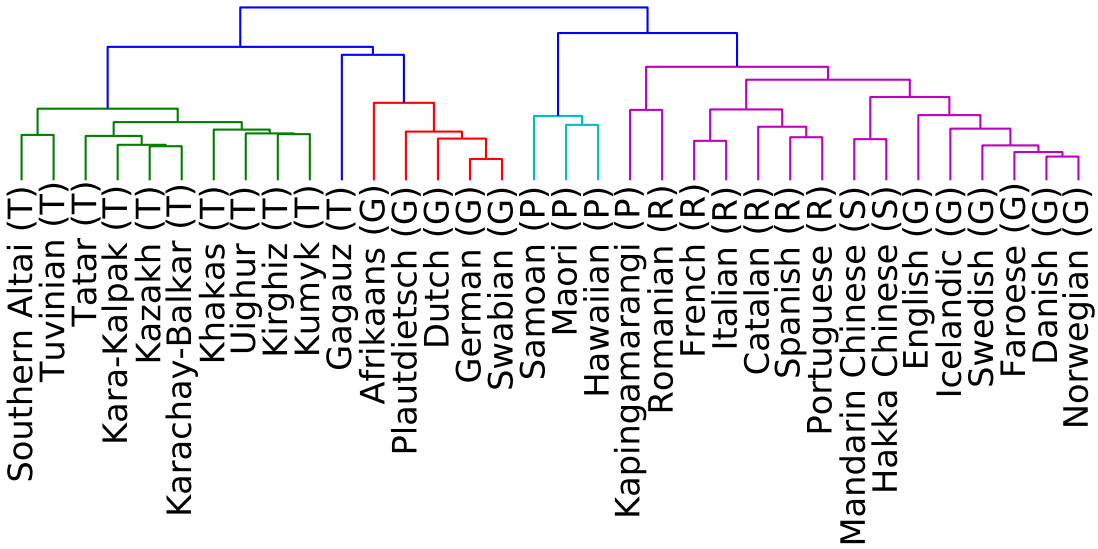

Figure 1: Hierarchical clustering based on word order statistics from our algorithm. Language families represented are (G)ermanic, (R)omance, (T)urkic, (P)olynesian and (S)initic.

the relative numbers are quite similar within each pair.

As a way of visualizing our data, we also tried performing hierarchical clustering of languages, by normalizing the word order count vectors and treating them (together) as a single 14-dimensional vector. The result confirmed that languages can be grouped remarkably well on basis of these five automatically extracted word order features. A subset of the clustering containing all languages from five language families represented in the New Testament corpus can be found in Figure 1. While the clustering mostly follows traditional genealogical boundaries, it is perhaps more interesting to look at the cases where it does not. The most glaring case is the wide split between the West Germanic and the North Germanic languages, which in spite of their shared ancestry have widely different word order characteristics. Interestingly, English is not grouped with the West Germanic languages, but rather with the North Germanic languages which it has been in close contact with.[2] One can also note that the Sinitic languages, with respect to word order, are quite close to the North Germanic languages.

Table 2 shows the agreement between the algorithm's output and the corresponding WALS chapter for each feature. The level of agreement is high, even though the sample consists mainly of languages unrelated to English, from which the dependency structure and PoS annotations were transferred. The **most common** column gives the ratio of the most common ordering for each feature (according to WALS), which can serve as a naive baseline.

As expected, the lowest level of agreement is observed for WALS chapter 81A, which has a lower baseline since it allows six permutations of the verb, subject and object, whereas all the other features are binary. In addition, this feature requires that *two* dependency relations (subject-verb and object-verb) have been correctly transferred, which substantially reduces the number of relations available for comparison.

The fact that sources sometimes differ as to the basic word order of a given language makes it evident that the disagreement reported in Table 2 is not necessarily due to errors made by our algorithm. Another example of this can be found when looking at the order of adjective and noun in some Romance languages (Spanish, Catalan, Portuguese, French and Italian), which are all classified as having noun-adjective order (Dryer, 2013a). It turns out that adjective-noun order in fact dominates in all of these languages, narrowly when using type counts and by a fairly large margin when using token counts. This result was confirmed by manual inspection, which leads us

[2]One reviewer pointed us to the controversial claim of Emonds (2011), that modern English in fact *is* a North Germanic language, albeit with strong influence from the extinct West Germanic language of Old English.

Table 2: Agreement between WALS and our results, on languages present in both datasets. The relative frequency of the most common ordering is given for comparison. **Types** is the agreement using type-based counts (see text for details), whereas **Tokens** uses token-based counts.

Feature	Languages	Types	Tokens	Most common
81A: Subject, Object, Verb (Dryer, 2013e)	342	85.4%	85.7%	SOV: 43.3%
82A: Subject, Verb (Dryer, 2013d)	376	89.4%	90.4%	SV: 79.8%
83A: Object, Verb (Dryer, 2013c)	387	96.4%	96.4%	VO: 54.8%
85A: Adposition, Noun Phrase (Dryer, 2013b)	329	94.8%	95.1%	Prep: 50.4%
87A: Adjective, Noun (Dryer, 2013a)	334	85.9%	88.0%	AdjN: 68.9%

to search further for an explanation for the discrepancy.[3] The Universal Dependency Treebank (McDonald et al., 2013) version 2 contains subcorpora in French, Italian, Spanish and Brazilian Portuguese. In all of these, noun-adjective order is dominant, which casts further doubts on our result. The key difference turns out to be the genre: whereas the modern texts used for the Universal Dependency Treebank have mainly noun-adjective order, we used our supervised annotation pipeline to confirm that the French translations of the New Testament indeed are dominated by adjective-noun order. This should serve as a warning about extrapolating too far from results obtained in one very specific genre, let alone in a single text.

5 Conclusions and future directions

The promising results from this study show that high-precision annotation transfer is a realistic way of exploring word order features in very large language samples, when a suitable parallel text is available. Although the WALS features on word order already use very large samples (over a thousand languages), using our method with the New Testament corpus contributes about 600 additional data points per feature, and adds quantitative data for all of the 986 languages contained in the corpus.

There are many other structural properties of languages that could be investigated with high-precision annotation transfer in massively parallel corpora, not just regarding word order but also within in domains such as negation, comparison and tense/aspect systems. While there are limits to the quality and types of answers obtainable, our work demonstrates that for some problems it is possible to obtain quick, quantitative answers that can be used to guide more traditional and thorough typological research.

On the technical side, the alignment model used is based on a non-symmetrized IBM model 1, and more elaborate methods for alignment and annotation projection could potentially lead to more accurate results. Preliminary results however indicate that adding a HMM-based word order model akin to Vogel et al. (1996) actually leads to somewhat reduced agreement with the WALS classification, because the projections become biased towards the word order characteristics of the source language(s), in our case English. This indicates that using the less accurate but also less biased IBM model 1 is in fact an advantage, when aggressive high-precision filtering is used.

Acknowledgments

Thanks to Calle Börstell, Östen Dahl, Francesca Di Garbo, Joakim Nivre, Janet B. Pierrehumbert, Jörg Tiedemann, Bernhard Wälchli, Mats Wirén and the anonymous reviewers for helpful comments and discussions.

[3] Thanks to Francesca Di Garbo for helping with this.

References

Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: An optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 58–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF - Language Typology and Universals*, 60(2):95–99.

Östen Dahl. 2007. From questionnaires to parallel corpora in typology. *STUF - Language Typology and Universals*, 60(2):172–181.

Matthew S. Dryer and Martin Haspelmath. 2013. *The World Atlas of Language Structures Online*. `http://wals.info`.

Matthew S. Dryer. 2007. Word order. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, volume I, chapter 2, pages 61–131. Cambridge University Press.

Matthew S. Dryer. 2013a. Order of adjective and noun. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013b. Order of adposition and noun phrase. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013c. Order of object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013d. Order of subject and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013e. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Joseph Emonds. 2011. English as a North Germanic language: From the Norman conquest to the present. In Roman Trušník, Katarína Nemčoková, and Gregory Jason Bell, editors, *Proceedings of the Second International Conference on English and American Studies*, pages 13–26, Zlín, Czech Republic, September.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Massachusetts.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 392–399, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, EACL 2012, pages 54–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Coşkun Mermer and Murat Saraçlar. 2011. Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 182–187, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135, 6.

Robert Östling. 2014. Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 123–127, Gothenburg, Sweden, April. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bernhard Wälchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50(3):671–710.

Bernhard Wälchli. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Linguistic Variation in Text and Speech*, number 28 in Linguae & Litterae, pages 355–393. De Gruyter.