Do Free Word Order Languages Need More Treebank Data? Investigating Dative Alternation in German, English, and Russian

Daniel Dakota, Timur Gilmanov, Wen Li, Christopher Kuzma,
Evgeny Kim, Noor Abo Mokh, Sandra Kübler
Indiana University
Bloomington, IN, USA
{ddakota,timugilm,wl9,ckuzma,egkim,noorabom,skuebler}@indiana.edu

Abstract

We investigate whether non-configurational languages, which display more word order variation than configurational ones, require more training data for a phenomenon to be parsed successfully. We perform a tightly controlled study comparing the dative alternation for English (a configurational language), German, and Russian (both non-configurational). More specifically, we compare the performance of a dependency parser when only canonical word order is present with its performance on data sets when all word orders are present. Our results show that for all languages, canonical data not only is easier to parse, but there exists no direct correspondence between the size of training sets containing free(er) word order variation and performance.

1 Introduction

Parsing morphologically rich languages (MRLs) has received much attention in recent years. Research in these areas shows that parsing MRLs presents challenges that cannot be addressed properly with existing parsing approaches. One of the challenges that MRLs pose to parsing lies with the effects that free(er) word order has on parsing. Non-configurational languages, such as German or Russian, mark grammatical functions using a case system, rather than utilizing word order, as happens in configurational languages such as English. In German, for example, the same proposition can be expressed in different word orders, differing only in the structure of information (see section 3 for examples). This means that the parser will not only need to assign grammatical functions when parsing in order to have a meaningful analysis, but also means that the parser needs to have access to the different word orders in the training set in order to be able to handle the phenomena correctly. This raises the question whether we intrinsically need more training data for non-configurational languages than for configurational ones, independent of the effects of syntactic annotation schemes and other differences between the languages.

In the current paper, we investigate the question of whether non-configurational languages require larger training sizes than configurational ones by comparing English, German, and Russian dependency parsing. Since, to our knowledge, this is the first attempt to look at the correlation between free word order and data size, one task was to set up an experimental design that will allow us to investigate this question while disregarding confounding factors as far as possible. We use a tightly controlled setting for our experiments, focusing on dative alternation, which means that we work with extremely small data sets. For this reason, we first explain the problem in more detail and discuss the confounding factors in section 2. Then, in section 3, we give an overview of the phenomena that we use in our experiments. In section 4, we discuss related work. In section 5, we give an overview of the treebanks used in the experiments, and in section 6, we discuss the experimental setup. We present the results of our experiments in section 7 and conclude in section 8.

2 Problem Statement

We plan to investigate whether more training data is needed for non-configurational languages because of the higher variation in word order. However, it is not entirely obvious how to investigate this question since configurational and non-configurational languages differ not only in this respect but also in many others. For example, languages differ with respect to the amount of morphology represented in word forms. This has a

Proceedings of the 6th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2015), pages 14–20,
Bilbao, Spain, July 23rd 2015.

direct influence on parsing accuracy because the more word forms that exist, the higher the possibility of data sparseness. Seddah (p.c.), for example, has shown that when English and French are trained on the same data sizes, the English results on word forms are comparable to French results on lemmas. In addition to language-internal differences, there are also differences in the corpus text and in the syntactic annotations. Kübler (2005) and Rehbein and van Genabith (2007) have shown that there are large variations between the different annotation schemes of German text.

We perform a comparison of English as a configurational language and German and Russian as the non-configurational languages. Since English and German are close relatives they share many phenomena, thus minimizing differences that may affect parsing accuracy. Although Russian shows many similarities to German (complex case system & case syncretism), it is from a different language family and its case system is more complex than that of German. In order to abstract away from differences in the treebanks, we focus on one relevant phenomenon. By using only sentences that exhibit this phenomenon, we keep the effect of other phenomena to a minimum. For our experiments we focus on dative alternation, a well known phenomenon in linguistics (c.f. e.g., (Bresnan et al., 2007)). Dative alternation was chosen as it is common across many languages, and it occurs frequently in any given language. For examples of dative alternations in the languages used here, see the next section.

However, by focusing exclusively on dative alternation, we severely restrict the size of the data set with which we can work. Thus, the parsing results per se are not very meaningful and should not be considered out of context. Additionally, we still face the problem that there are differences in the annotation schemes, which in turn affect parsing results. This means it is also not possible to compare directly across languages. E.g., an F-score for German cannot be directly compared to the F-score for English of the corresponding experiment. Instead, we need to consider a setup in which we first restrict the data sets to sentences in which the dative alternation is present in its canonical word order. We will call such sentences "canonical sentences" and sentences of other word orders "non-canonical sentences." Thus, our baseline per language consists of a training and test set containing only canonical sentences. Here we start with a small training set, which is then increased incrementally (where data size allows). The learning curve can show us how much data is required for the parser to "acquire" the phenomenon. In the second step, we then look at training and test sets with all word orders. Our results can be interpreted as (the beginnings of) learning curves, and the curves can be compared between cases, giving us an indication as to whether the scenario encompassing all word orders is a more difficult task. If we abstract away from the individual numbers and only look at the visual representations, the learning curves also allow us to make a comparison across languages. Another possibility to compare across languages would be using *tedeval* (Tsarfaty et al., 2012). However, tedeval requires a common label set or only an unlabeled evaluation in a cross-language setting, which is not useful in our case since a considerable degree of relevant information in the trees of the non-canonical languages is encoded in grammatical functions.

3 Phenomena

We concentrate on dative alternation. This construction involves ditransitive verbs which select for two arguments in addition to the subject. In non-configurational languages, these objects tend to be noun phrases (NPs), one marked as dative and one as accusative. In configurational languages the arguments are either expressed as two NPs in fixed order or as an NP and a PP.

In English, the dative alternation follows the pattern for configurational languages. We follow standard assumptions (e.g., (Chomsky, 1975)) and consider the case using an NP and a PP as the canonical case. We use a linguistic definition of the canonical case, rather than a data-driven one, such as by Bresnan et al. (2007), to avoid circular reasoning. An example of the alternation is shown in (1).

(1) a. The woman gives the book to the man. (canonical)
    b. The woman gives the man the book.

German, in contrast, expresses the alternation via two NPs in different cases. Thus, all possible alternations of the nominative NP (subject), the dative NP, and the accusative NP are possible, with nominative, dative, accusative being the canonical order (Lenerz, 1977), cf. the German translations

of (1) in (2).

(2) a. Die Frau$_{nom}$ gibt dem Mann$_{dat}$ das Buch$_{acc}$. (canonical)
b. Die Frau gibt das Buch dem Mann.
c. Das Buch gibt die Frau dem Mann.
d. Das Buch gibt dem Mann die Frau.
e. Dem Mann gibt das Buch die Frau.
f. Dem Mann gibt die Frau das Buch.

Russian, similar to German, expresses the dative alternation using two NPs in different cases allowing for all possible combinations of nominative, dative, and accusative NPs, as shown in (3) (Russian translations of (1)). The canonical case in Russian, however, is the same as in German (Kallestinova, 2007).

(3) a. Женщина$_{nom}$ даёт мужчине$_{dat}$ книгу$_{acc}$. (canonical)
b. Женщина даёт книгу мужчине.
c. Мужчине даёт женщина книгу.
d. Книгу даёт мужчине женщина.
e. Книгу мужчине даёт женщина.
f. Мужчине книгу даёт женщина.

Our hypothesis is that the more complex alternation in non-configurational languages, typical of a MRL, should require more training data relative to the corresponding canonical case and relative to English.

4 Related Work

To our knowledge, there exists no prior work on the interrelation of word order freedom and data set size. There are however approaches that concentrate on parsing specific phenomena. Most work in this area has concentrated on parsing coordinations (e.g., (Hogan, 2007; Kübler et al., 2009a; Kurohashi and Nagao, 1994)). Other work has focused on improving PP attachment in a parser (e.g., (Agirre et al., 2008; Foth and Menzel, 2006)). We are not aware of any work that focuses on parsing the dative alternation, but this phenomenon has been used as a feature in parsing (Chan et al., 2010) or acquired automatically (Sasano et al., 2013). Additionally, there are test suites for parsing German (Kübler et al., 2009b; Maier et al., 2014), but none of these cover dative alternations.

While the effect of training set size has been recognized as an important factor, we are not aware of any explicit investigations across languages. But the factor was noted in the shared tasks on dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007).

There is considerable work on parsing these three languages. We focus here on approaches that are state of the art across languages, which we consequently use in our experiments. For dependency parsing, we use the MATE parser (Bohnet, 2010). For this parser, labeled accuracy scores (LAS) for German are close to the English results: 90.33 for English and 88.06 for German (Bohnet, 2010). The best achieved LAS for Russian is 82.3, based on MaltParser (Nivre et al., 2008).

5 Data

We collected dative alternations from English, German, and Russian as representatives of fixed-order and free-order languages. We discard sentences that contain other complex phenomena, such as coordinations that scope over higher constituents rather than base phrases.

English: For English we used the Penn Treebank (Marcus et al., 1993), more specifically the conversion to dependencies by *pennconverter* (Johansson and Nugues, 2007). The Penn Treebank consists of approximately 50,000 sentences. First, we used the TIGERSearch (König et al., 2003) queries described below to extract the alternations in the constituent version of the Penn Treebank. Then we extracted the same sentences from the dependency version.

For the canonical order, we extracted sentences with an NP and a PP, as in (1-a), and we only used cases where the PP is annotated as "DTV" (ditransitive) as grammatical function. Only 10 sentences of the alternation in which the PP precedes the NP were found in the Penn Treebank. Since this is a marked phenomenon, we did not include these sentences. For the non-canonical order, we extracted sentences that have two NP nodes under the same VP node. After collecting sentences with dative alternations, passive sentences and sentences with movement were filtered out manually.

German: For German, we used the dependency version of the TüBa-D/Z treebank, version 9 (Telljohann et al., 2004), which covers approx. 85 000 sentences. However, we extracted ditransitive sentences from the constituent version via TIGERSearch queries and subsequently used the

corresponding dependency versions. TüBa-D/Z annotates phrase structures using a topological fields model (Höhle, 1986), which allows for an effective extraction of the dative alternation. The alternations were extracted via two sub-cases. The first case covers V2 clauses, where the initial field dominates one NP while the middle field dominates the other two. Precedence was then set in the middle field to obtain the six varieties. In the second case (subordinate clauses), all objects including the subject were assumed to be under the same field node. Precedence was then set to extract the six different orders. The annotation scheme does not distinguish between true reflexives and reflexive objects. For this reason, all dative alternation sentences containing a reflexive were left in the training data.

Russian: For Russian, we used the dependency treebank SynTagRus[1] (Nivre et al., 2008), which includes a comprehensive morphological and syntactic annotation (Boguslavsky et al., 2009) in the form of dependencies. SynTagRus covers approx. 62 000 sentences. To extract the alternations, we first converted the corpus to CoNLL dependency format and then obtained all sentences from the corpus that contained dependency relations between a verb and two NPs, one in dative and one in accusative case. This means, we do include pro-drop sentences.

An overview of the resulting data sets is shown in Table 1.

Alternation	No.	Total
English		
NP NP	640	
NP PP	358	998
German		
NP_{nom} NP_{dat} NP_{acc}	1 030	
NP_{nom} NP_{acc} NP_{dat}	176	
NP_{acc} NP_{nom} NP_{dat}	113	
NP_{acc} NP_{dat} NP_{nom}	65	
NP_{dat} NP_{nom} NP_{acc}	245	
NP_{dat} NP_{acc} NP_{nom}	11	1 640
Russian		
V NP_{dat} NP_{acc}	293	
V NP_{acc} NP_{dat}	130	
NP_{dat} V NP_{acc}	124	
NP_{acc} V NP_{dat}	62	
NP_{acc} NP_{dat} V	51	
NP_{dat} NP_{acc} V	30	690

Table 1: Overview of the extracted sentences.

6 Experimental Setup

We performed a 10-fold cross validation. Each data set was randomized and split into ten equal parts. We then established two versions of the training and test sets: first, we extracted training and test sets that contain only the canonical word order. This was done to establish a baseline of how difficult the syntactic phenomenon is to parse in its "standard" form. The second version contains all canonical and non-canonical examples. In order to determine the learning curve, we set the initial training set size for all training sets to 200 sentences, with increments of 100.

To ensure that the different sizes of the training data are balanced with regard to word order, the sentences are chosen so that the distribution of alternations mirrors the distribution of all dative alternation sentences. For example, since there is a total of 1 640 German sentences, with 1 030 canonical ones, the initial training set of 200 contains that same ratio, i.e., 126 canonical sentences.

For parsing, we use the MATE Parser (Bohnet, 2010). We use gold POS tags as input since we are interested in parsing performance, not in the interaction between POS labeling and parsing, and we also provide gold morphological information for German and Russian.

As discussed above, we are aware of the fact that despite our best efforts, we still face a situation in which we have differences in text types, POS tags, and syntactic annotation, all of which can influence parsing performance. Ideally, we would want to control for these variables as well, but that would mean using an artificial annotation as well as reducing our data set even further. Thus, we advise the reader that the analysis in the following section needs to be approached carefully.

For evaluation, we used the 2009 CONLL shared task scorer[2] and report the labeled attachment score (LAS) and the unlabeled attachment score (UAS) on the whole sentence. We evaluate full sentences rather than individual labels because we assume that the ability to parse the phenomenon includes its correct interpretation in context.

[1] http://ruscorpora.ru/en/

[2] http://ufal.mff.cuni.cz/conll2009-st/scorer.html

	English				German				Russian			
	canonical		all		canonical		all		canonical		all	
Train	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
200	80.74	84.15	78.36	82.07	83.19	85.94	82.57	85.52	69.82	82.22	67.42	79.26
300	83.92	87.04	80.96	84.36	84.69	87.22	83.99	86.66			70.17	81.62
400			82.28	85.62	85.88	88.27	85.61	88.12			71.15	82.02
500			83.37	86.55	86.67	88.90	86.14	88.51			72.55	83.38
600			84.19	87.22	87.28	89.55	86.87	89.11				
700			84.69	87.71	87.68	89.80	87.30	89.51				
800			85.12	88.09	88.08	90.18	87.81	89.78				
900			85.58	88.57	88.56	90.57	88.18	90.27				
1000							88.71	90.71				
1100							88.86	90.82				
1200							89.01	90.97				
1300							89.40	91.23				
1400							89.31	91.14				

Table 2: Results for all languages. Train shows the training set size.

7 Results

The main results of our experiments are shown in Table 2. Note that we have more examples for German than for the other languages, thus the empty cells for those languages. For English and Russian, we only have a rather small set of canonical cases, which does not allow us to look at a learning curve. We also note that LAS for English and German are considerably higher than for Russian, at the same training size. Since the Russian unlabeled attachment score is similar to those for the other languages, we assume that the Russian label set is more difficult to determine, but this needs further investigation. However, since the underlying treebanks have different annotation schemes, we need to be careful not to read too much into this comparison.

More importantly, the results in Table 2 show that for all languages, the results on the data sets that contain all word order variations in training and test are noticeably lower than for the canonical only data sets. Additionally, for English and German, we see that the (beginnings of a) learning curve show a faster increase for the canonical case. For English, this means that we need 2-3 times more data from all word orders to reach the same LAS results as for the canonical case only. For Russian, the ratio is closer to 1.5, and for German around 1.5 for the smallest training set and decreasing to 1.1 for larger sets. These results provide support for our hypothesis that we need more training data when all word orders are included. However, there seems to be no direct correlation between the number of possible word orders and the increase in training set size.

8 Conclusion & Future Work

In our experiments, we investigated whether the larger number of alternations in non-configurational languages requires larger training sets than the fixed word order in configurational languages. We used a tightly controlled study, with English as a configurational language and German and Russian as non-configurational languages, and we restricted the investigation to dative alternations. Our results show that we do need more data if all word orderings are present, but there is no direct correlation between the number of possible orders and the required increase in the training set. However, this is only a first step, and our results need to be analyzed further.

For the future, we are planning to add more languages, such as Arabic, which shows an interesting mix with regard to dative alternation: like English, it allows NP NP and NP PP orders, but both variants can occur in either order. We also plan to extend this study to more phenomena, also including those where English only allows one fixed word order while non-configurational languages allow alternations. Finally, we are interested in extending these experiments to human language acquisition.

Acknowledgements

We are grateful to the Laboratory of Computational Linguistics of the Institute of Information Transmission Problems in Moscow for making their SynTagRus Corpus plus dependencies available. We are also grateful to Can Liu for help in an early stage, and to Rex Sprouse and Heike Telljohann for insightful discussions.

References

Eneko Agirre, Timothy Baldwin, and David Martinez. 2008. Improving parsing and PP attachment performance with sense information. In *Proceedings of ACL-08: HLT*, pages 317–325, Columbus, OH.

Igor Boguslavsky, Leonid Iomdin, Svetlana P Timoshenko, and Tatyana I Frolova. 2009. Development of the Russian tagged corpus with lexical and functional annotation. In *Proceedings of the MONDILEX Third Open Workshop on Metalanguage and Encoding Scheme Design for Digital Lexicography*, pages 83–90, Bratislava, Slovakia.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 89–97. Beijing, China.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Language Learning (CoNLL)*, pages 149–164, New York, NY.

Samuel W. K. Chan, Lawrence Y. L. Cheung, and Mickey W. C. Chong. 2010. Tree topological features for unlexicalized parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 117–125, Beijing, China.

Noam Chomsky. 1975. *The Logical Structure of Linguistic Theory*. Springer.

Kilian A. Foth and Wolfgang Menzel. 2006. The benefit of stochastic PP attachment to a rule-based parser. In *Proceedings of COLING/ACL 2006*, pages 223–230, Sydney, Australia.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 680–687, Prague, Czech Republic.

Tilman Höhle. 1986. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*, Tartu, Estonia.

Elena Dmitrievna Kallestinova. 2007. Aspects of word order in Russian. *Theses and Dissertations*.

Esther König, Wolfgang Lezius, and Holger Voormann. 2003. Tigersearch 2.1 user's manual. Technical report, IMS, University of Stuttgart, Germany.

Sandra Kübler, Erhard W. Hinrichs, Wolfgang Maier, and Eva Klett. 2009a. Parsing coordinations. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 406–414, Athens, Greece.

Sandra Kübler, Ines Rehbein, and Josef van Genabith. 2009b. TePaCoC - a corpus for testing parser performance on complex German grammatical constructions. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories*, pages 15–28, Groningen, The Netherlands.

Sandra Kübler. 2005. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 293–300, Borovets, Bulgaria.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534.

Jürgen Lenerz. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen. Studien zur deutschen Grammatik*. Gunter Narr.

Wolfgang Maier, Miriam Kaeshammer, Peter Baumann, and Sandra Kübler. 2014. Discosuite – a parser test suite for German discontinuous structures. In *Proceedings of the Ninth Language Resources and Evaluation Coference (LREC)*, Reykjavik, Iceland.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL 2007 Shared Task. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic.

Joakim Nivre, Igor Boguslavsky, and Leonid Iomdin. 2008. Parsing the SynTagRus treebank of Russian. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 641–648, Manchester, UK.

Ines Rehbein and Josef van Genabith. 2007. Treebank annotation schemes and parser evaluation for German. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 630–639, Prague, Czech Republic.

Ryohei Sasano, Daisuke Kawahara, Sadao Kurohashi, and Manabu Okumura. 2013. Automatic knowledge acquisition for case alternation between the passive and active voices in Japanese. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1213–1223, Seattle, WA.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 2229–2235, Lisbon, Portugal.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2012. Joint evaluation of morphological segmentation and syntactic parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6–10, Jeju Island, Korea.