UPPSALA
UNIVERSITET

Correlating language divergence and cross-lingual parsing accuracy

Jesper Segeblad

Uppsala University
Department of Linguistics and Philology
Språkteknologiprogrammet
(Language Technology Programme)
Bachelor's Thesis in Language Technology

May 26, 2015

Supervisor:
Joakim Nivre

Abstract

This thesis investigates the correlation between language divergence and parsing accuracy scores when transferring a delexicalized dependency parser from a source language to a target language through the means of a regression analysis. 10 languages using the Universal Dependencies annotation are included in this study. For each language, a delexicalized dependency parser is trained and used to parse the test data of all 10 languages. A few different measures of language divergence have been explored, including the distribution of dependency relations, the distribution of part-of-speech tags, and distribution of different word orders in common constructions. The main analysis consist of a multiple linear regression model, and the results from this show a correlation between the language divergence variables considered and parsing accuracy in all but two cases. A few simple regression models are presented as well and show that the distribution of dependency relations and part-of-speech tags have a larger correlation with parsing accuracy than the divergence in the word order variables considered.

Contents

1 Introduction

Grammatical analysis of natural language has been the subject of much research in the natural language processing community throughout the years. It is a crucial step in many downstream applications, such as information extraction. In recent years, dependency parsing has attracted a lot of attention, partly due to being a fast and efficient approach whilst not sacrificing descriptive adequacy. And as in most areas of natural language processing, statistical methods are the most popular and yield the best results. However, statistical methods for dependency parsing relies on large amounts of annotated training data. Such data exists for some languages, but not all are so lucky. And as interest in parsing languages that lack such resources grow, the question of how to transfer parsers across languages has gained a lot of attention.

To say that the subject in a clause often is a noun, or that the modifier of a noun is an adjective is not a particularly bold claim. A word can be said to merely be a realization of a particular part-of-speech, and the syntactic structure of a sentence is simply relations between these parts-of-speech. If we also say that these structures are similar between languages, we should be able to transfer these structures. It is however not that simple, since the ordering of these parts-of-speech varies between languages, and what would constitute a syntactic relation in one language may be expressed morphologically in another. However, given two languages that are reasonably similar, directly transferring a syntactic parser between them should be fairly effective.

One thing that causes large problems in the area of cross-lingual research, even when annotated resources exists, is the lack of consistency between different dependency treebanks. This has made it hard to actually evaluate how well the transfer approaches work. For this reason, Universal Dependencies has been developed to provide a consistent way to annotate dependency treebanks.

1.1 Purpose

The purpose of this thesis is to investigate whether the difference between languages affects the parsing results when transferring a delexicalized parser from a source language to a target language. I will work from the hypothesis that target language parsing accuracy increase as the the divergence to the source language decreases.

To test the hypothesis, two things are needed; parsing accuracy scores for a transferred parser and a measured distance between the languages. Several distance metrics will be explored and used in a regression model to measure the correlation.

1.2 Outline

This thesis is structured as follows. Chapter 2 will bring up the necessary background, including previous research in the area of cross-lingual dependency parsing, and the Universal Dependencies project. Chapter 3 will describe parsing experiments and present the result of those. Chapter 4 will describe how similarity between the treebanks and languages are measured, with what metrics and which variables. In chapter 5, the similarity will be used to try to explain the difference in parsing results. Chapter 6 will discuss the results from chapter 5. Finally, chapter 7 will conclude this thesis.

2 Background

In this chapter, the necessary background for this thesis will be presented, including a discussion of why a direct transfer approach can work in practice, as well as a brief presentation of the Universal Dependencies project.

2.1 Cross-lingual parsing

As the unsupervised methods for dependency parsing are far from state-of-the-art, and supervised methods relies on annotated training data, how to transfer parsers to languages that lack such data have attracted a lot of attention. Methods for doing this have often involved parallel corpora to project dependency structures from English to a target language, e.g. Ganchev et al. (2009) and Smith and Eisner (2009). These methods do however add additional processing steps, making them somewhat ineffective.

Zeman and Resnik (2008) experimented on transferring a constituent-based parser between closely related languages, specifically Swedish and Danish. Their approach was to replace the tokens in the treebank they used with POS tags, and later reversing this – thus still adding another processing step, although simpler.

Søgaard (2011) extended the approach of Zeman and Resnik (2008), and like them used treebanks with lexical items removed, to less related languages, experimenting with data in Arabic, Danish, Bulgarian and Portuguese. The approach used was to select sentences in the source language that resembles sentences in the target language for training a dependency parser, and reported accuracy scores between 50 and 75 per cent.

McDonald et al. (2011) could show that a delexicalized dependency parser (a parser that does not rely on lexical features for its classification) can obtain nearly as good result as one also making use of lexical features for unlabeled parsing (i.e. only attaching a token to the correct head, without labeling the arc with a syntactic relation). In an experiment on English data, they reported an unlabeled attachment score of 82.5% for a delexicalized parser, and 89.3% for a parser making use of lexical features as well. On the basis of this, they could argue that the amount of information part-of-speech tags contain is quite significant, and perhaps enough for unlabeled parsing. This led to the idea of direct transfer – to train a delexicalized parser on one language (*source*) and use this to directly parse another language (*target*). They tested this idea on eight languages, among them Swedish and Danish, and a parser trained on Danish proved to produce the worst result when parsing Swedish. One can not argue that this is due to the dissimilarity of the languages, since they are extremely

similar with regards to their syntactic structure. A better explanation for this is the difference in how the treebanks used are annotated.

2.2 Universal Dependencies

Because of the problems with divergent annotation schemes, the idea of creating one that can be shared across languages have sparked interest in the NLP community.

McDonald et al. (2013) proposed such a cross-linguistic annotation scheme and released data using this, the Universal Dependency Treebank (UDT). They made use of the Stanford Dependencies (SD) (De Marneffe et al., 2006) for the syntactic relations and the Google universal part-of-speech (POS) tags (Petrov et al., 2012), and released treebanks for 6 languages annotated with this scheme.

The Universal Dependencies (UD) project (Nivre, 2015) is continuation of UDT, aiming at creating an annotation scheme for dependency treebanks that can be used for a broad set of languages. The first steps towards the UD scheme was de Marneffe et al. (2014), proposing Universal Stanford Dependencies (USD), a taxonomy useful for a wide set of languages. Using SD as a starting-point, 42 relations that were deemed to represent most of the syntactic structure across languages were presented. As the aim is to have a consistent scheme across languages, there has to be an agreement as how to treat constructions that are essentially the same, but show a difference in surface form. One example of this is the treatment of the copula verb *be*. As some languages completely lack this copula, *be* is never the head of a clause, just as in SD (de Marneffe et al., 2014). The USD taxonomy is the one used in the UD scheme, with the exception of three removed relations (*nfincl*, *relcl* and *ncmod*) and one added (*acl*) (Nivre, 2015). As pointed out in both de Marneffe et al. (2014) and Nivre (2015), it is important to also be able to capture constructions that are specific to a particular language. In the UD scheme, these language specific relations are extensions of some universal relation in the form *universal:extension*.

Apart from the universal dependency relations, UD also includes universal POS tags and universal morphological features. The Universal POS tags are based upon the Google universal POS tags, which originally consist of 12 tags, but have been expanded to 17 tags for the UD scheme (Nivre, 2015). The universal morphological features are based upon Interset (Zeman, 2008) and consist 17 different features, and UD guidelines allows language specific features as well.

In the first release (v1, January 2015), there are 10 treebanks utilizing the UD guidelines, which all have been used in this thesis: Czech (cs), German (de), English (en), Spanish (es), Finnish (fi), French (fr), Irish Gaelic (ga), Hungarian (hu), Italian (it), and Swedish (sv). These are all divided into training, development and test sets. The training and test sets are used in this thesis.

The Universal Dependencies treebank uses the CoNLL-U format, a re-worked version of the CoNLL-X format (Buchholz and Marsi, 2006), with the main difference being the information in some of the fields. As described in Buchholz and Marsi (2006), the CoNLL-X format contains all the sentences of the treebank in a single text file with each token on a new line. Each line

```
1       This    _       DET     DT      _       2       det     _       _
2       item    _       NOUN    NN      _       6       nsubj   _       _
3       is      _       VERB    VBZ     _       6       cop     _       _
4       a       _       DET     DT      _       6       det     _       _
5       small   _       ADJ     JJ      _       6       amod    _       _
6       one     _       NOUN    NN      _       0       root    _       _
7       and     _       CONJ    CC      _       6       cc      _       _
8       easily  _       ADV     RB      _       9       advmod  _       _
9       missed  _       VERB    VBN     _       6       conj    _       _
10      .       _       PUNCT   .       _       6       punct   _       _
```

Figure 2.1: Example of an English sentence in the CoNLL-U format.

consists of ten fields separated by a tab character, and sentences are separated by a blank line. Underscore is used for fields that lack values.

The most important difference for the purpose of this thesis is that the field for coarse-grained POS tags now is used for the universal POS tags, described above. An example of a sentence in the CoNLL-U format is provided in figure 2.1. The columns of particular interest for the purpose of this thesis:

- **[1]** The index of the token
- **[4]** The universal POS tag
- **[7]** The head of the current token
- **[8]** The dependency relation the token has to its head

3 Parsing experiments

The first step towards the goal of correlating parsing accuracy with language divergence is to get accuracy scores for parsers transferred from source to target languages. The parsing experiments are conducted using a direct transfer approach, similar to that used by McDonald et al. (2011).

The parser used for these experiments is MaltParser, a data-driven dependency parser-generator (Nivre et al., 2006). This choice is made on the basis of the flexibility of the system, as well as being completely language independent, allowing to easily train parsers for all of the languages considered. Further, it has been shown to produce good results even with relatively small sets of training data (Nivre et al., 2007), which in this case is somewhat relevant due to the fact that some of the data sets are fairly small. One of the greatest advantages with the flexibility is defining the feature model, making it easy to generate a delexicalized parsing model. Where, for example, Zeman and Resnik (2008) removed all lexical items from the treebank for training, the approach used here is to simply delexicalize the feature model used for training, making minimal changes to the actual data.

3.1 Method

These experiments are conducted as follows: train a parsing model on the training split of the data (the language on which it was trained will be referred to as the *source*), parse the test data of all 10 languages, including the source (the parsed language will be referred to as the *target*), and finally evaluating this parse against the gold standard. This is done for all 10 languages. The training and test splits used are the default splits of the UD treebanks.

Since the UD guidelines do allow for language specific dependencies, there exists some discrepancies between the data sets with regards to the dependency relations. Some manipulation of the data was performed to get a consistent set of relations across all languages. Fortunately, since these language specific relations are in practice an extension of some specific universal relation, they can be mapped into the general dependency relation of which it is a subset. For example, Swedish have the language specific relation *nmod:agent* where *nmod* is the universal relation, and *agent* is the extension. This relation is then mapped to the general *nmod* relation, and the mapping is done by truncating at the colon.

MaltEval (Nilsson and Nivre, 2008) is used to evaluate the parser performance.

3.2 Parser configuration

Even though MaltParser allows for a great number of optimization options, configurations of the parser are kept to a minimum, only making changes to the feature model used for inducing the parsing models, and leaving all other parameters to the default. The main reason behind this is that the same model should be used across all languages. A configuration optimized for one language might not be the best for another, and to find a configuration that is reasonably well balanced for all languages is a lengthy process given all the configuration options.

The algorithm used for parsing is the arc-eager algorithm, a transition-based algorithm as described in Nivre (2008). The important thing to note about this is the data structures it utilizes: a stack of partially processed tokens (on which they are candidates for getting a dependency relation), and an input buffer from which the tokens of the sentence are pushed on to the stack. Further, it does not handle non-projective trees, and no method for handling those types of trees are utilized.

The feature model used can be found in figure 3.1. As can be seen, the only features relied on are the universal POS tags (in the feature model referred to as `CPOSTAG`) and dependency relations (`DEPREL`). *Stack* refers to the stack of partially processed tokens, and *Input* refers to the input buffer. *InputColumn* refers to features of the input string and *OutputColumn* refers to features from the output generated by the parser, i.e. the partially built dependency structure. This means that the universal POS tag of the two first tokens on the stack, and the four first tokens in the input buffer are used as features. The dependency relations of the first token of the stack, and the first in the input buffer are also used. *Merge* and *Merge3* combines two and three features into one, respectively.

The reasoning behind using this very sparse set of features is that the features should be the same across all languages.

3.3 Results

All parsing results are reported in labeled attachment score (LAS), which is the percentage of tokens that have both the correct head and correct dependency label, excluding punctuation tokens (Buchholz and Marsi, 2006).

As can be seen in table 3.1, the results are not on par with the state-of-the-art, which is expected due to the parser configuration presented earlier, with a sparse set of features. The direct transfer approach does not produce very good results for many of the language pairs, with the worst source language being Irish.

The lowest source to source result is for Finnish. This is something that can not be attributed to a sparsity of training data, since the training part of the Finnish data is larger than the Swedish for example. A better explanation for this is that Finnish is a morphologically rich language (MRL), and parsing such languages are simply harder than those that are morphologically poorer (Tsarfaty et al., 2010).

Even though introducing morphological features would most likely improve the results for Finnish in particular, and the others that can be considered

```
InputColumn(CPOSTAG, Stack[0])
InputColumn(CPOSTAG, Stack[1])
InputColumn(CPOSTAG, Input[0])
InputColumn(CPOSTAG, Input[1])
InputColumn(CPOSTAG, Input[2])
InputColumn(CPOSTAG, Input[3])
OutputColumn(DEPREL, Stack[0])
OutputColumn(DEPREL, ldep(Stack[0]))
OutputColumn(DEPREL, rdep(Stack[0]))
OutputColumn(DEPREL, ldep(Input[0]))
Merge(InputColumn(CPOSTAG, Stack[0]), OutputColumn(DEPREL, Stack[0]))
Merge(InputColumn(CPOSTAG, Input[0]), OutputColumn(DEPREL, ldep(Input[0])))
Merge3(InputColumn(CPOSTAG, Stack[1]), InputColumn(CPOSTAG, Stack[0]),
        InputColumn(CPOSTAG, Input[0]))
Merge3(InputColumn(CPOSTAG, Stack[0]), InputColumn(CPOSTAG, Input[0]),
        InputColumn(CPOSTAG, Input[1]))
Merge3(InputColumn(CPOSTAG, Input[0]), InputColumn(CPOSTAG, Input[1]),
        InputColumn(CPOSTAG, Input[2]))
Merge3(InputColumn(CPOSTAG, Input[1]), InputColumn(CPOSTAG, Input[2]),
        InputColumn(CPOSTAG, Input[3]))
Merge3(InputColumn(CPOSTAG, Stack[0]), OutputColumn(DEPREL,ldep(Stack[0])),
        OutputColumn(DEPREL, rdep(Stack[0])))
```

Figure 3.1: The feature model used for inducing the parsing models.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	**0.629**	0.449	0.425	0.463	0.377	0.425	0.283	0.314	0.465	0.398
DE	0.453	**0.662**	0.447	0.53	0.319	0.501	0.324	0.378	0.536	0.435
EN	0.424	0.505	**0.724**	0.581	0.353	0.57	0.371	0.356	0.593	0.491
ES	0.425	0.45	0.462	**0.712**	0.295	0.604	0.405	0.335	0.668	0.427
FI	0.391	0.336	0.374	0.294	**0.564**	0.285	0.161	0.356	0.26	0.358
FR	0.408	0.485	0.49	0.636	0.298	**0.675**	0.373	0.32	0.653	0.435
GA	0.203	0.263	0.224	0.411	0.115	0.385	**0.61**	0.157	0.449	0.233
HU	0.288	0.356	0.283	0.3	0.321	0.289	0.152	**0.62**	0.291	0.246
IT	0.422	0.441	0.43	0.624	0.275	0.602	0.382	0.326	**0.753**	0.419
SV	0.391	0.474	0.42	0.473	0.323	0.475	0.376	0.404	0.516	**0.636**

Table 3.1: Results from the parsing experiments reported in LAS. The rows represents the source language on which the parser was trained and the columns represent the target language used for testing. Bold numbers indicate that the source and target are the same.

an MRL, we do not want to introduce features that are divergent across the languages – only rely on those features that are common to them all. Even though there are universal morphological features as described earlier, they are kept in the same column as the language specific ones meaning that features may vary across languages.

To get a clearer picture of the cross-lingual results and how they relate to the source to source result, they are also presented as relative to the source. This means that the source to source result is considered the highest achievable

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	1	0.714	0.676	0.736	0.599	0.599	0.450	0.499	0.739	0.633
DE	0.684	1	0.675	0.801	0.482	0.757	0.489	0.571	0.810	0.657
EN	0.586	0.698	1	0.802	0.488	0.787	0.512	0.492	0.819	0.678
ES	0.597	0.632	0.649	1	0.414	0.848	0.569	0.471	0.938	0.600
FI	0.693	0.596	0.663	0.521	1	0.505	0.285	0.631	0.461	0.635
FR	0.604	0.719	0.726	0.942	0.441	1	0.553	0.474	0.967	0.644
GA	0.333	0.431	0.367	0.674	0.189	0.631	1	0.257	0.736	0.382
HU	0.465	0.574	0.456	0.484	0.518	0.466	0.245	1	0.469	0.397
IT	0.560	0.586	0.571	0.829	0.365	0.799	0.507	0.433	1	0.556
SV	0.615	0.745	0.660	0.744	0.508	0.747	0.591	0.635	0.811	1

Table 3.2: Parsing results relative to the source to source result. The rows represent the source language and columns represent the target language.

with a model trained on that particular language, hence getting a score of 1.

In table 3.2 there are some interesting things to be seen. In particular, parsing results between the Romance languages, French, Italian, and Spanish, are fairly good. Parsing Italian with a model trained on French data amounts to a result almost as good as parsing French with the same model. This does not come as a surprise given that they are three very similar languages. Between the Germanic languages the transfer results are not as impressive. With a model trained on English, accuracy scores for parsing Swedish are lower than for parsing Spanish, or any of the Romance languages. And parsing Czech data with a model trained on German yields a slightly higher accuracy score than parsing English and Swedish with the same model.

4 Measuring distance

From the parsing results presented in the previous chapter some interesting things could be seen, such as a relatively high parsing accuracy when transferring a parser between the Romance languages. But the results between the Germanic languages showed that we can not rely on our intuition when reasoning about how well it might work to directly transfer a parser from a source to target language. German might have more in common with Czech than it has with Swedish and English. This makes it necessary to somehow measure the divergence between the languages, which is what will be explored in this chapter.

Three types of variables have been considered for measuring the distance between the languages: the distribution of POS tags, the distribution of dependency relations and word order. For this part, the training part of respective treebank is used, for two reasons: of the standard splits it is by far the largest, and the test split should only be used for evaluation.

The analysis and quantification of the difference presented in this chapter was done by the means of a program implemented in Java.

4.1 Distribution of dependency relations

As in the parsing experiments, language specific relations are mapped to their respective universal dependency relations before retrieving the frequencies. These frequencies are presented, and later used, as relative to the total number of relations in respective treebank. Since all information regarding this is readily available in the data, the task is merely to count the number of times each relation occurs.

The distribution of the universal relations are presented in 4.1. Some things to note is that the *case* relation have a very low frequency in Finnish and Hungarian, *det* having a low frequency in Czech and Finnish, and *iobj* not being found in the Czech, Finnish and Irish data. Most of these things are expected due to the nature of these languages.

The *case* relation is used for case-marking, such as adpositions. In Finnish and Hungarian such case-marking is often expressed morphologically, and the the low frequency of *det* Czech and Finnish can also be attributed to the use of inflection. Both Finnish and Irish lack indirect objects (*iobj*), which is why no such relations can be found in the data of these two languages. Czech do however have indirect objects, and the lack of *iobj* in the Czech data is due to a bug.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
acl	0.013	0.01	0.016	0.025	0.028	0.027	0.018	0.007	0.024	0.02
advcl	0.005	0.005	0.019	0.01	0.017	0.007	0.014	0.01	0.009	0.018
advmod	0.053	0.054	0.046	0.022	0.07	0.025	0.027	0.057	0.03	0.064
amod	0.118	0.057	0.044	0.054	0.049	0.049	0.03	0.109	0.052	0.047
appos	0.005	0.025	0.008	0.021	0.006	0.017	0.004	0.006	0.004	0.005
aux	0.011	0.013	0.032	0.01	0.016	0.015	-	0.001	0.024	0.027
auxpass	0.007	0.012	0.007	0.004	0.004	0.007	-	-	0.008	-
case	0.105	0.108	0.085	0.161	0.015	0.156	0.123	0.017	0.148	0.108
cc	0.029	0.03	0.033	0.031	0.042	0.027	0.024	0.045	0.027	0.035
ccomp	0.007	0.002	0.012	0.004	0.01	0.003	0.016	0.011	0.007	0.007
compound	0.011	0.004	0.051	0.001	0.018	0.003	0.069	0.013	0.002	0.009
conj	0.045	0.041	0.037	0.04	0.051	0.036	0.028	0.047	0.033	0.044
cop	0.011	0.017	0.022	0.015	0.018	0.014	0.016	0.005	0.011	0.016
csubj	0.003	0.001	0.001	0.002	0.001	≈ 0	0.002	0.004	0.001	0.003
csubjpass	≈ 0	≈ 0	≈ 0	≈ 0	-	-	-	-	≈ 0	≈ 0
dep	0.012	0.003	≈ 0	0.003	≈ 0	0.001	-	-	0.004	0.002
det	0.001	0.159	0.076	0.142	0.017	0.153	0.081	0.129	0.155	0.073
discourse	≈ 0	-	0.003	-	0.001	0.001	≈ 0	≈ 0	≈ 0	-
dislocated	-	-	≈ 0	-	-	-	-	≈ 0	-	0.001
dobj	0.059	0.029	0.049	0.032	0.056	0.037	0.037	0.042	0.033	0.043
expl	-	0.001	0.003	-	-	0.002	-	-	0.011	0.004
foreign	0.001	-	≈ 0	-	-	-	0.001	-	≈ 0	-
goeswith	-	-	0.001	-	-	-	-	≈ 0	-	-
iobj	-	0.004	0.002	0.017	-	0.002	-	0.002	0.001	0.002
list	-	-	0.002	-	-	-	≈ 0	≈ 0	-	-
mark	0.016	0.012	0.035	0.021	0.019	0.01	0.066	0.017	0.01	0.031
mwe	0.002	0.001	0.002	0.007	0.003	0.009	-	-	0.004	0.018
name	0.009	0.026	-	0.017	0.017	0.018	0.008	0.025	0.013	0.003
neg	≈ 0	0.004	0.01	0.004	0.01	0.007	0.003	0.011	0.007	0.007
nmod	0.165	0.113	0.104	0.149	0.19	0.144	0.142	0.15	0.153	0.115
nsubj	0.058	0.058	0.08	0.037	0.08	0.053	0.074	0.066	0.044	0.076
nsubjpass	0.004	0.011	0.006	0.003	-	0.007	-	-	0.007	0.017
nummod	0.02	0.011	0.012	0.017	0.019	0.011	0.016	0.013	0.01	0.019
parataxis	0.001	0.002	0.007	0.004	0.004	0.001	0.001	0.006	0.002	0.002
punct	0.145	0.13	0.115	0.11	0.145	0.112	0.106	0.098	0.117	0.109
remnant	-	-	≈ 0	-	0.001	-	-	0.004	-	-
reparandum	-	-	≈ 0	-	-	-	-	-	-	-
root	0.068	0.053	0.061	0.037	0.075	0.041	0.043	0.101	0.047	0.065
vocative	≈ 0	-	0.001	-	0.001	-	0.001	≈ 0	≈ 0	-
xcomp	0.012	0.003	0.015	0.003	0.018	0.002	0.048	0.007	0.004	0.009

Table 4.1: The distribution of dependency relations in the treebanks. Relations marked with a hyphen indicates that the relation does not exist in the data, and relations marked with ≈ 0 have a frequency less than 0.001.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
ADJ	0.115	0.072	0.061	0.056	0.071	0.056	0.05	0.126	0.067	0.087
ADP	0.1	0.11	0.09	0.164	0.015	0.16	0.153	0.017	0.152	0.122
ADV	0.051	0.047	0.052	0.029	0.076	0.034	0.019	0.076	0.039	0.075
AUX	0.015	0.022	0.039	0.014	0.02	0.022	-	0.001	0.028	-
CONJ	0.034	0.03	0.033	0.032	0.04	0.025	0.032	0.045	0.026	0.042
DET	0.015	0.126	0.089	0.14	-	0.154	0.088	0.131	0.155	0.058
INTJ	≈ 0	-	0.003	-	0.001	0.001	≈ 0	0.001	≈ 0	≈ 0
NOUN	0.242	0.178	0.173	0.181	0.28	0.18	0.27	0.224	0.196	0.241
NUM	0.032	0.027	0.02	0.028	0.027	0.025	0.014	0.017	0.017	0.02
PART	0.005	0.007	0.027	≈ 0	-	0.002	0.066	0.002	≈ 0	0.011
PRON	0.046	0.048	0.078	0.031	0.065	0.044	0.037	0.033	0.041	0.059
PROPN	0.069	0.114	0.063	0.094	0.061	0.079	0.04	0.063	0.052	0.017
PUNCT	0.149	0.13	0.116	0.11	0.149	0.112	0.106	0.149	0.117	0.111
SCONJ	0.018	0.006	0.015	0.018	0.018	0.007	0.016	0.017	0.01	0.014
SYM	-	-	0.003	0.003	0.003	0.001	≈ 0	-	0.001	-
VERB	0.108	0.083	0.134	0.095	0.174	0.09	0.097	0.097	0.098	0.143
X	-	0.001	0.004	0.005	0.001	0.007	0.011	0.001	0.001	0.001

Table 4.2: The distribution of the universal part-of-speech tags in respective treebank. Tags marked with a hyphen does not exist in the data, and tags marked with ≈ 0 have a frequency less than 0.001.

4.2 Distribution of part-of-speech tags

As the dependency relations, the part-of-speech tag distribution is presented and used as relative to the total number of tokens. The method here is the same as the one used for dependency relations: simply count the number of times each POS tag occurs in the data.

The distribution of POS tags can be found in table 4.2. Some things can be noted on this. Many of the low frequencies are expected, such as *SYM* (used for symbols that are not punctuation marks), and interjections (*INTJ*). Finnish and Hungarian have a low frequency of adpositions (*ADP*), for the same reason as the low frequency of the dependency relation *case* – it is often expressed through inflection. Irish do not have auxiliary verbs (*AUX*) which is why this tag can not be found in the Irish data, while the lack of them in the Swedish is because they have not been separated from regular verbs in the release used here, which also explains the relatively high frequency of verbs in Swedish.

4.3 Word order

Word order variables are inspired by features from The World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), though in some cases a bit simplified. Just as the distribution of dependency relations and POS tags, these variables are considered a frequency distribution, and not as taking on a discrete value depending on which order is the most common.

CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
2871	3894	2520	2324	1984	5044	77	217	1625	1086

Table 4.3: Number of transitive clauses found for each language with described method.

The word order variables considered and how they are denoted throughout this chapter:

- Subject, verb and object (SVO) (Dryer, 2013a)
- Subject and verb (SV) (Dryer, 2013b)
- Object and verb (OV) (Dryer, 2013c)
- Adjective and noun (AdjN) (Dryer, 2013e)
- Ad positions (Prep) (Dryer, 2013d)
- Numeral and noun (NumN) (Dryer, 2013f)

Dryer (1997) has argued against the subject-object-verb word order typology, partially on the basis of it being a relatively rare construction, and have argued for using the binary subject-verb and verb-object typological features instead. With the method described for finding those types of constructions later in this chapter, one could raise some concerns regarding this. But with the large amounts of data available, we can still get a reasonable set of points for a calculating a frequency distribution. As can be seen in table 4.3 there are only two instances where one could argue that the sample size is too small to draw any definite conclusions, Irish and Hungarian, due to these languages having the smallest amount of data available. For the purpose here however, the sets are considered enough.

As previously described, each token in the treebank is kept on a separate line with information on the index, POS tag, dependency relation to the head, and the index of the head. Because of this format, retrieving information on the word order for the binary variables is a fairly simple process. Each token in the corpora is matched against the prerequisites (the dependency relation of the token and POS tag of the head, or in the case of ad positions, simply the POS tag). If a match is found, the index of the token and its head are extracted and compared against each other. The prerequisites for the binary word order variables can be found in table 4.4.

Getting the order of words in transitive clauses is somewhat more complex. For each sentence, the predicate is found by finding the dependent of `root` and checking whether it is a verb, due to the way that copulas are handled. Next, if both a subject and an object have this predicate as head, their indexes are saved and a three-way comparison between the subject, object and verb indexes are made.

The five variables in table 4.5 are binary variables and only one of the orderings is presented. The opposite order (for example verb-subject or postpositions) is 1 minus the presented value. All values for each language in table 4.6 are a distribution on its own.

	Dependency	Head
SV	nsubj	VERB
VO	dobj	VERB
AdjN	amod	NOUN
NumN	nummod	NOUN
Prep	ADP	any

Table 4.4: The prerequisites for the binary word order variables. Dependency denotes the dependency relation of the token, or in the case of adpositions, the POS tag of the token. Head denotes the POS tag of the tokens head.

	SV	VO	AdjN	NumN	Prep
CS	0.664	0.668	0.958	0.798	0.961
DE	0.676	0.549	0.998	0.99	0.952
EN	0.966	0.961	0.971	0.893	0.929
ES	0.795	0.904	0.241	0.883	0.983
FI	0.819	0.7	0.999	0.757	0.136
FR	0.967	0.761	0.27	0.925	0.951
GA	0.143	0.786	0.009	0.625	0.8
HU	0.743	0.469	0.999	1	0.006
IT	0.738	0.855	0.313	0.77	0.983
SV	0.818	0.945	0.994	0.839	0.874

Table 4.5: Distribution of binary word order features. They are presented with as one preceding the other.

	SVO	SOV	VSO	VOS	OSV	OVS
CS	0.54	0.079	0.078	0.079	0.028	0.195
DE	0.443	0.132	0.292	0.068	0.026	0.04
EN	0.981	0	0	0	0.018	0
ES	0.92	0.044	0.003	0.009	0	0.023
FI	0.885	0.016	0.01	0.002	0.02	0.067
FR	0.783	0.204	0.002	0	0.003	0.009
GA	0.026	0	0.961	0	0	0.013
HU	0.415	0.276	0.032	0.088	0.074	0.115
IT	0.813	0.04	0.002	0.02	0.002	0.122
SV	0.774	0	0.165	0.007	0.02	0.033

Table 4.6: The distribution of different word orders in transitive clauses.

As can be seen from the results in table 4.5 and table 4.6, most languages show some degree of flexibility in most of the variables. These distributions also seem to reflect what we know of the structure of these languages reasonably well; English almost exclusively placing subjects before the verb, German having a relatively free word order, and Irish being strictly VSO. Not surprisingly, the order of numeral and noun, and adpositions are the variables where there

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0.117	0.102	0.123	0.071	0.123	0.131	0.128	0.116	0.083
DE	0.117	0	0.077	0.031	0.128	0.021	0.119	0.082	0.028	0.057
EN	0.102	0.077	0	0.092	0.077	0.078	0.06	0.1	0.074	0.036
ES	0.123	0.031	0.092	0	0.148	0.014	0.107	0.114	0.025	0.063
FI	0.071	0.128	0.077	0.148	0	0.14	0.118	0.076	0.138	0.082
FR	0.123	0.021	0.078	0.014	0.14	0	0.109	0.111	0.01	0.052
GA	0.131	0.119	0.06	0.107	0.118	0.109	0	0.126	0.11	0.087
HU	0.128	0.082	0.1	0.114	0.076	0.111	0.126	0	0.108	0.097
IT	0.116	0.028	0.074	0.025	0.138	0.01	0.11	0.108	0	0.049
SV	0.083	0.057	0.036	0.063	0.082	0.052	0.087	0.097	0.049	0

Table 4.7: The Jensen-Shannon divergence calculated from the distribution of dependency relations.

are least divergence between languages to be found. As pointed out in Dryer (2013f), most European languages tend to place the numeral before the noun. And most languages tend to have either prepositions or postpositions, to not have a dominant order is quite rare (Dryer, 2013d). Only two of the languages here are postpositional: Finnish and Hungarian.

4.4 Quantifying language difference

To measure how different the languages are with respect to the previously presented variables, the Jensen-Shannon divergence (JSD) is calculated over the frequency distributions, giving divergence values between the languages. Since eight different distributions were presented, we get eight different divergence values for each language pair.

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(P \parallel M) + \frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(Q \parallel M) \tag{1}$$

P and Q are two frequency distributions. The function D_{KL} is the Kullback-Leibler divergence (KL), of which the Jensen-Shannon divergence is a symmetrized and smoothed version, and the distribution M is the average of the two distributions , i.e. $M = \frac{1}{2}(P + Q)$. The Jensen-Shannon divergence is always non-negative and bounded by 1 given that the base 2 logarithm is used (Lin, 1991), which is the case here.

$$\mathrm{D}_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \tag{2}$$

The main reason for choosing JSD over KL is the fact that it is smoothed, meaning that P and Q do not have to be absolutely continuous with respect to each other. D_{KL} is undefined if $P(x) > 0$ and $Q(x) = 0$, which is not the case for JSD. This is quite important, mostly for dependency relations, since some relations do not exist for some languages, e.g. Irish lacking indirect objects.

The Jensen-Shannon divergence calculated from the distribution of dependency relations can be found in table 4.7. Between the Romance languages,

where the direct transfer approach worked reasonably well, the divergence is low. Irish show a relatively high divergence to all languages, and transferring a parser from Irish did not yield very high attachment scores. These observations may give some hints as to what expect from the test of correlation. The results from the other divergence calculations can be found in appendix A.

5 Correlating parsing accuracy and distance

In the two previous chapters the preliminaries needed for testing the hypothesis, that target language parsing accuracy increase as divergence to the source language decrease, have been presented: parsing results and a number of measures of divergence. This chapter will explain how the test of correlation is done, and presents the results from this.

5.1 Method

To test the hypothesis, a linear regression model is used, using the least squares method as implemented in the statistical software package R. Parsing results are set as the response variable, and the divergence variables, distributional and word order, are set as the explanatory. This regression analysis is done for each of the 10 source languages. The data used for the analysis is the parsing results as achieved by the direct transfer from the source language, and the divergence in the variables presented in the previous chapter between the source and target language. The parsing results used are those that are relative to the source. Source to source results are also included in the model and thus we have to include the divergence for a source to source comparison as well, which is zero since the source is equal to the source.

The hypothesis is tested with both simple linear regression models, meaning that only one variable is used as explanatory, and multiple linear regression models, combining several explanatory variables. We can not expect that only one of the divergence variables is enough to explain the varying transfer results, as none of them on their own captures the divergence between the languages, which is why a multiple linear regression model is used for the main analysis.

For the multiple regression analysis, I will start from the assumption that six variables will explain the parsing results: the distribution of dependency relations (dep), the distribution of POS tags (pos), the divergence in subject-object-verb word order (svo), the divergence in subject-verb word order (sv), the divergence in verb-object word order (vo), and the divergence in adjective-noun word order (adj). This is based on that these show the largest amount of divergence across the languages. The two other word order variables, numeral-noun (num) and adpositions (adp), are left out from the first analysis, but are added to see whether they can improve the fit of the model. In other words: three different multiple regression models are used.

As the hypothesis states, parsing accuracy should increase as the divergence decreases. Thus, the null hypothesis is that the distance between source language

and target language does not affect target parsing result.

5.2 Results

The main, and most important, analysis is the multiple linear regression where the response variable is the parsing result and our explanatory variables are the divergence in the distribution of dependency relations, POS tags, and word order. Some intermediate results presented as well, in the form of simple linear regression results with one of the variables as the explanatory.

The simple linear regression is presented with the intercept, slope, and R^2. For each of the multiple linear regressions, adjusted R^2 and the *p-value* is reported. The reason for using adjusted R^2 is that R^2 always increases as more variables are added to the model (Chatterjee and Simonoff, 2013). The *p-value* and standardized beta coefficients for the explanatory variables are reported as well. When referring to languages in this section it is the source language that is meant.

5.2.1 Simple linear regression results

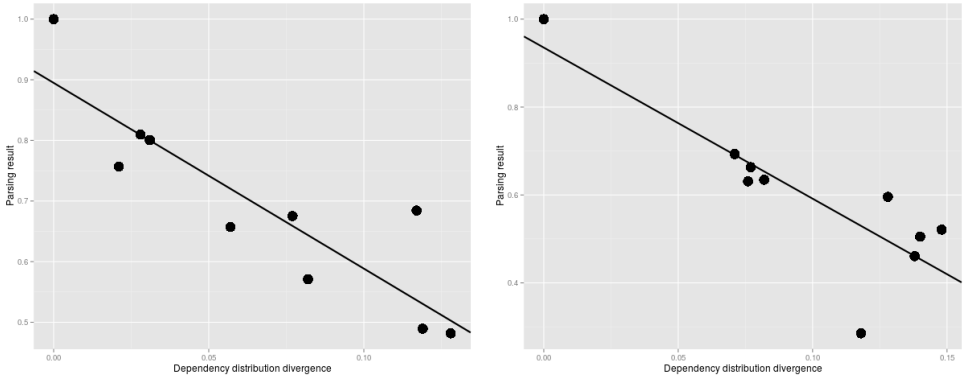
Results from simple linear regressions are presented in table 5.1. In figure 5.1, plots with the divergence in dependency relation distribution as explanatory variable for four languages is presented, chosen because of the relatively good fit to the regression line.

Interestingly enough, the distribution of POS tags and dependency relations have a better fit to the linear model than the word order variables considered, and explains target language parsing results to a greater extent. This can be seen primarily in the R^2 values, where a higher value means that the explanatory variable accounts for more of the variation in the dependent variable.

Consider for example the results for French. When using the dependency divergence variable as explanatory, we get an R^2 value of 0.83, while using the divergence in SV word order as explanatory, the R^2 is as low as 0.104. In the analysis for Hungarian with SVO word order as explanatory, R^2 is quite high compared to the other languages. Still, using POS tag and dependency relation divergence yield a higher value. That the dependency relation and POS tag divergence accounts for more of the variation in the parsing accuracy than the word order variables goes for all languages. However, the dependency and POS tag divergence do not explain all of the variation, confirming that we need to combine variables into a single model.

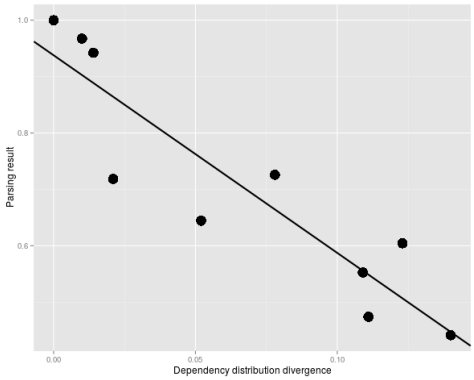
5.2.2 Multiple linear regression results

As we can see in table 5.2 all results from the whole regression but two, Czech and German, are significant at $p < 0.05$. This means that in all but those two cases we can reject the null hypothesis. Looking at the R^2 values, the variables used do explain most of the variation in the target language parsing results, especially in the case of English, Spanish, French, Irish, and Swedish, where the explanatory variables used in this model accounts for between 95 and 98 per cent of the variation in parsing accuracy. The slightly lower R^2 in the case

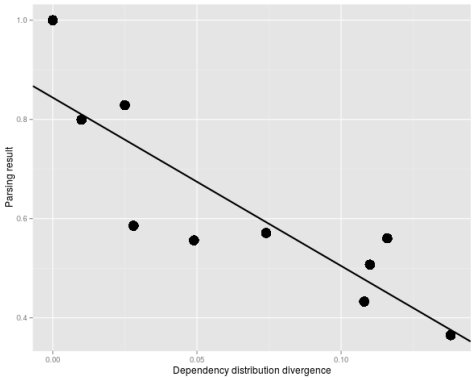


(a) German

(b) Finnish



(c) French



(d) Italian

Figure 5.1: Regression lines plotted for four languages with the relative parsing result on the y-axis and the divergence in dependency distributions on the x-axis.

of Hungarian, Finnish, and Italian suggest that there might be some variable affecting the parsing accuracy that have not been considered here.

The p-values for single variables tells us that not all of the variables are significant in predicting the outcome. However, one should also take into account the phenomenon of *collinearity* that arises when explanatory variables are correlated with each other (Chatterjee and Simonoff, 2013), rendering variables redundant in the presence of others. This can also explain the positive beta coefficients. The hypothesis indirectly states that these should be negative (since parsing accuracy should increase as divergence decrease). And from what we know about our explanatory variables, that languages placing subject before the verb also tend to place the adjective before the noun for example, this is a very likely explanation for the high p-values and positive beta coefficients.

Even though adding the numeral-noun and adposition divergence increases the fit of the model in some cases, it also increases the p-values, making some analyses insignificant as can be seen in table 5.3.

Because of the insignificance of some variables, some additional models were tested in order to find one where all explanatory variables are significant. But as seen in table 5.2, the variables that are insignificant varies between analyses.

	Intercept	Slope	R^2
CS	0.939	-2.766	0.531
DE	0.895	-3.063	0.78
EN	0.896	-3.023	0.31
ES	0.908	-3.295	0.775
FI	0.935	-3.436	0.715
FR	0.937	-3.505	0.83
GA	0.887	-3.997	0.396
HU	0.959	-4.792	0.847
IT	0.844	-3.392	0.75
SV	0.950	-4.035	0.717

(a) Divergence in dependency relation distribution

	Intercept	Slope	R^2
CS	0.946	-4.888	0.609
DE	0.852	-3.499	0.732
EN	0.955	-5.556	0.772
ES	0.856	-3.452	0.737
FI	0.951	-3.155	0.863
FR	0.887	-3.518	0.758
GA	0.829	-4.466	0.61
HU	0.941	-5.945	0.839
IT	0.801	-3.658	0.709
SV	0.917	-3.736	0.461

(b) Divergence POS tag distribution

	Intercept	Slope	R^2
CS	0.736	-0.389	0.239
DE	0.842	-0.724	0.304
EN	0.766	-0.35	0.31
ES	0.717	-0.255	0.119
FI	0.676	-0.450	0.376
FR	0.776	-0.338	0.176
GA	0.779	-0.406	0.221
HU	0.658	-0.623	0.496
IT	0.663	-0.242	0.099
SV	0.775	-0.403	0.202

(c) Divergence in SVO word order

	Intercept	Slope	R^2
CS	0.712	-0.882	0.187
DE	0.736	-0.838	0.163
EN	0.738	-0.454	0.213
ES	0.686	-0.290	0.024
FI	0.652	-1.035	0.381
FR	0.751	-0.36	0.104
GA	0.680	-0.556	0.147
HU	0.562	-1.153	0.282
IT	0.636	-0.327	0.022
SV	0.725	-0.388	0.097

(d) Divergence in SV word order

Table 5.1: Regression results for four analyses.

Because of this, an optimal model that can be used for all languages could not be found.

	DEP	POS	SVO	SV	VO	ADJ	R^2	p-value
CS	0.610 (0.504)	-1.667 (0.156)	-0.702 (0.312)	0.113 (0.835)	0.111 (0.678)	0.971 (0.090)	**0.621**	**0.168**
DE	-0.330 (0.453)	-0.540 (0.217)	-0.170 (0.743)	-0.149 (0.670)	0.026 (0.927)	-0.005 (0.990)	**0.722**	**0.110**
EN	-0.150 (0.126)	-0.751 (0.001)	-1.133 (0.025)	0.495 (0.145)	0.249 (0.134)	0.370 (0.020)	**0.987**	**0.001**
ES	0.056 (0.767)	-0.508 (0.045)	0.165 (0.536)	-0.539 (0.132)	-0.247 (0.097)	-0.544 (0.015)	**0.970**	**0.004**
FI	-0.676 (0.382)	0.049 (0.956)	-0.406 (0.521)	-0.023 (0.975)	-0.177 (0.471)	-0.210 (0.655)	**0.861**	**0.041**
FR	-0.135 (0.348)	-0.475 (0.030)	0.618 (0.211)	-1.004 (0.095)	-0.194 (0.094)	-0.460 (0.019)	**0.981**	**0.002**
GA	-0.065 (0.641)	-0.216 (0.248)	-0.130 (0.587)	-0.099 (0.562)	0.022 (0.858)	-0.749 (0.018)	**0.952**	**0.009**
HU	-0.633 (0.060)	-0.207 (0.476)	-0.483 (0.293)	0.144 (0.686)	-0.024 (0.881)	0.148 (0.387)	**0.918**	**0.019**
IT	-0.038 (0.909)	-0.441 (0.230)	0.059 (0.855)	-0.382 (0.333)	-0.192 (0.356)	-0.514 (0.070)	**0.904**	**0.024**
SV	-0.770 (0.012)	-0.505 (0.016)	-0.397 (0.360)	-0.360 (0.267)	0.677 (0.026)	0.952 (0.011)	**0.964**	**0.006**

Table 5.2: Regression results for a multiple linear regression with six variables. Standardized beta coefficients are presented for all variables together with the p-value in parentheses. The bold numbers represent adjusted R^2 and *p-value* for the whole regression.

	R^2	**p-value**
CS	0.504	0.336
DE	0.774	0.165
EN	0.988	0.009
ES	0.963	0.028
FI	0.823	0.131
FR	0.972	0.021
GA	0.931	0.052
HU	0.877	0.092
IT	0.958	0.032
SV	0.989	0.008

(a) Model with numeral-noun divergence added.

	R^2	**p-value**
CS	0.604	0.276
DE	0.89	0.083
EN	0.998	0.001
ES	0.961	0.030
FI	0.891	0.083
FR	0.99	0.008
GA	0.949	0.039
HU	0.878	0.092
IT	0.868	0.099
SV	0.970	0.023

(b) Model with adposition divergence added.

Table 5.3: R-squared and p-values for models with ad position and numeral-noun divergence added.

6 Discussion

One of the most interesting thing to note in the simple linear regression results is that the divergence in dependency distribution and POS tags give a better fit than the word order variables presented. It seems more intuitive that there would be a larger correlation between parsing accuracy and word order. One possible explanation for this might be that languages that are different in many respects show similarity in these word order variables, but might show a larger divergence in the two distributional ones. Another explanation might be the feature model used for parsing. Since the only features relied upon are dependency relations and POS tags, the divergence in these distributions in respective language most likely have an impact on target language parsing accuracy. Consider for example a parser trained on Irish data. Since Irish lack indirect objects, this parser would not be able to assign indirect objects their correct label.

As for the multiple regression model used, the results from analysis with Czech and German as source languages are not statistically significant and the null hypothesis can not be rejected in those cases. Two things that they have in common are that they are morphologically rich languages with a relatively free word order, and divergence in morphology is something that most likely affects the accuracy when transferring a parser trained on these languages. Morphological divergence have been not been included in the regression model, and morphological features were not used for training the parser. Apart from the two languages discussed, the analysis with lowest R^2 value was for Finnish as source language, another language with a rich morphology, providing more evidence that this might be an explanation.

Another thing that might have an impact might be the way that the divergence is measured, especially when it comes to the word order. If we for example consider the ordering of subject and verb. Language 1 with relatively fixed ordering might have a distribution of 0.85 SV and 0.15 VS, and language 2 with a freer ordering have a distribution of 0.55 SV and 0.45 VS. The divergence between these are not as large as if language 2 would have a predominantly VS order. The ordering should nonetheless have an impact on the target language results, even if the divergence as measured here is small.

Aside from the results of Czech and German, the variables used explain a great deal, between 86 and 98 per cent, of the varying target language results while still being statistically significant. The lack of consideration for morphological divergence that were discussed can be extended to the analyses that are statistically significant, but have a relatively low R^2 value, such as Hungarian and Finnish. Adding the divergence in numeral-noun and adposition variables to the model did increase the R^2 value for some analyses, but made

a few other analysis results insignificant. This can be expected, since these variables show the least difference between the languages considered.

7 Conclusions

In this thesis I have presented an attempt at correlating target language parsing results with the divergence between target and source language. I have showed that there exists a correlation between target language parsing results and the quantified distance between source and target language in all but two cases, and possible reasons for this have been discussed. The variables that have the largest impact on target language results are different depending on source language. This is not that surprising, since language that are largely different may share common traits.

Although not being the main purpose of this thesis, it has effectively been shown that direct transfer of labeled dependency parsers between similar languages can work reasonably well given a consistent annotation. A completely universal processing of natural language might be far off, but a universal processing of similar languages is not that unrealistic.

A Jensen-Shannon divergence tables

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0.053	0.054	0.072	0.052	0.075	0.085	0.074	0.069	0.041
DE	0.053	0	0.028	0.018	0.131	0.012	0.071	0.059	0.019	0.066
EN	0.054	0.028	0	0.047	0.099	0.035	0.059	0.076	0.038	0.049
ES	0.072	0.018	0.047	0	0.159	0.006	0.063	0.085	0.012	0.072
FI	0.052	0.131	0.099	0.159	0	0.162	0.171	0.096	0.15	0.096
FR	0.075	0.012	0.035	0.006	0.162	0	0.061	0.086	0.007	0.069
GA	0.085	0.071	0.059	0.063	0.171	0.061	0	0.113	0.067	0.05
HU	0.074	0.059	0.076	0.085	0.096	0.086	0.113	0	0.075	0.067
IT	0.069	0.019	0.038	0.012	0.15	0.007	0.067	0.075	0	0.057
SV	0.041	0.066	0.049	0.072	0.096	0.069	0.05	0.067	0.057	0

Table A.1: The Jensen-Shannon divergence calculated from the distribution of part-of-speech tags.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0.097	0.261	0.163	0.127	0.195	0.688	0.07	0.087	0.138
DE	0.097	0	0.338	0.245	0.243	0.227	0.416	0.129	0.23	0.133
EN	0.261	0.338	0	0.047	0.048	0.12	0.91	0.351	0.102	0.109
ES	0.163	0.245	0.047	0	0.026	0.052	0.883	0.234	0.031	0.109
FI	0.127	0.243	0.048	0.026	0	0.098	0.849	0.223	0.026	0.076
FR	0.195	0.227	0.12	0.052	0.098	0	0.895	0.17	0.096	0.191
GA	0.688	0.416	0.91	0.883	0.849	0.895	0	0.796	0.873	0.559
HU	0.07	0.129	0.351	0.234	0.223	0.17	0.796	0	0.178	0.271
IT	0.087	0.23	0.102	0.031	0.026	0.096	0.873	0.178	0	0.123
SV	0.138	0.133	0.109	0.109	0.076	0.191	0.559	0.271	0.123	0

Table A.2: The Jensen-Shannon divergence calculated from the distribution of different word orders in transitive clauses.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0	0.124	0.016	0.023	0.125	0.216	0.005	0.005	0.023
DE	0	0	0.117	0.013	0.02	0.118	0.226	0.004	0.003	0.02
EN	0.124	0.117	0	0.055	0.044	0	0.588	0.081	0.083	0.045
ES	0.016	0.013	0.055	0	0.001	0.056	0.335	0.003	0.003	0.001
FI	0.023	0.02	0.044	0.001	0	0.045	0.361	0.006	0.007	0
FR	0.125	0.118	0	0.056	0.045	0	0.59	0.082	0.084	0.046
GA	0.216	0.226	0.588	0.335	0.361	0.59	0	0.283	0.279	0.36
HU	0.005	0.004	0.081	0.003	0.006	0.082	0.283	0	0	0.006
IT	0.005	0.003	0.083	0.003	0.007	0.084	0.279	0	0	0.007
SV	0.023	0.02	0.045	0.001	0	0.046	0.36	0.006	0.007	0

Table A.3: The Jensen-Shannon divergence calculated from the distribution of subject and verb word orders.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0.011	0.115	0.063	0.001	0.008	0.013	0.029	0.036	0.097
DE	0.011	0	0.188	0.122	0.018	0.036	0.046	0.005	0.084	0.166
EN	0.115	0.188	0	0.01	0.097	0.066	0.054	0.245	0.026	0.001
ES	0.063	0.122	0.01	0	0.049	0.027	0.02	0.171	0.004	0.004
FI	0.001	0.018	0.097	0.049	0	0.003	0.007	0.04	0.026	0.081
FR	0.008	0.036	0.066	0.027	0.003	0	0.001	0.066	0.01	0.052
GA	0.013	0.046	0.054	0.02	0.007	0.001	0	0.08	0.006	0.041
HU	0.029	0.005	0.245	0.171	0.04	0.066	0.08	0	0.126	0.221
IT	0.036	0.084	0.026	0.004	0.026	0.01	0.006	0.126	0	0.017
SV	0.097	0.166	0.001	0.004	0.081	0.052	0.041	0.221	0.017	0

Table A.4: The Jensen-Shannon divergence calculated from the distribution of verb and object word orders.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0.016	0.001	0.447	0.019	0.416	0.836	0.02	0.372	0.011
DE	0.016	0	0.01	0.547	0	0.515	0.95	0.001	0.468	0.001
EN	0.001	0.01	0	0.474	0.013	0.442	0.867	0.013	0.397	0.006
ES	0.447	0.547	0.474	0	0.555	0.001	0.108	0.556	0.005	0.533
FI	0.019	0	0.013	0.555	0	0.522	0.957	0	0.475	0.002
FR	0.416	0.515	0.442	0.001	0.522	0	0.124	0.524	0.002	0.501
GA	0.836	0.95	0.867	0.108	0.957	0.124	0	0.959	0.151	0.934
HU	0.02	0.001	0.013	0.556	0	0.524	0.959	0	0.477	0.002
IT	0.372	0.468	0.397	0.005	0.475	0.002	0.151	0.477	0	0.454
SV	0.011	0.001	0.006	0.533	0.002	0.501	0.934	0.002	0.454	0

Table A.5: The Jensen-Shannon divergence calculated from the distribution of adjective and noun word orders.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0.084	0.013	0.01	0.002	0.025	0.027	0.109	0.001	0.002
DE	0.084	0	0.035	0.04	0.107	0.021	0.188	0.005	0.099	0.062
EN	0.013	0.035	0	0	0.023	0.002	0.074	0.056	0.02	0.004
ES	0.01	0.04	0	0	0.02	0.004	0.067	0.061	0.016	0.003
FI	0.002	0.107	0.023	0.02	0	0.04	0.015	0.134	0	0.008
FR	0.025	0.021	0.002	0.004	0.04	0	0.1	0.039	0.035	0.013
GA	0.027	0.188	0.074	0.067	0.015	0.1	0	0.219	0.018	0.043
HU	0.109	0.005	0.056	0.061	0.134	0.039	0.219	0	0.126	0.085
IT	0.001	0.099	0.02	0.016	0	0.035	0.018	0.126	0	0.006
SV	0.002	0.062	0.004	0.003	0.008	0.013	0.043	0.085	0.006	0

Table A.6: The Jensen-Shannon divergence calculated from the distribution of numeral and noun word orders.

	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS	0	0	0.004	0.003	0.588	0	0.048	0.855	0.003	0.019
DE	0	0	0.002	0.006	0.568	0	0.04	0.833	0.006	0.014
EN	0.004	0.002	0	0.013	0.526	0.002	0.027	0.787	0.013	0.006
ES	0.003	0.006	0.013	0	0.641	0.006	0.072	0.912	0	0.036
FI	0.588	0.568	0.526	0.641	0	0.568	0.35	0.057	0.641	0.44
FR	0	0	0.002	0.006	0.568	0	0.04	0.833	0.006	0.014
GA	0.048	0.04	0.027	0.072	0.35	0.04	0	0.586	0.072	0.007
HU	0.855	0.833	0.787	0.912	0.057	0.833	0.586	0	0.911	0.69
IT	0.003	0.006	0.013	0	0.641	0.006	0.072	0.911	0	0.036
SV	0.019	0.014	0.006	0.036	0.44	0.014	0.007	0.69	0.036	0

Table A.7: The Jensen-Shannon divergence calculated from the distribution of adpositions.

Bibliography

Sabine Buchholz and Erwin Marsi. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics, 2006.

Samprit Chatterjee and Jeffrey S Simonoff. *Handbook of regression analysis*, volume 5. John Wiley & Sons, 2013.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of Language Resources and Evaluation Conference*, volume 6, pages 449–454, 2006.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of Language Resources and Evaluation Conference*, 2014.

Matthew S. Dryer. On the six-way word order typology. *Studies in Language*, 21:69–103, 1997.

Matthew S. Dryer. *Order of Subject, Object and Verb*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013a. URL `http://wals.info/chapter/81`.

Matthew S. Dryer. *Order of Subject and Verb*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL `http://wals.info/chapter/82`.

Matthew S. Dryer. *Order of Object and Verb*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013c. URL `http://wals.info/chapter/83`.

Matthew S. Dryer. *Order of Adposition and Noun Phrase*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013d. URL `http://wals.info/chapter/85`.

Matthew S. Dryer. *Order of Adjective and Noun*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013e. URL `http://wals.info/chapter/87`.

Matthew S. Dryer. *Order of Numeral and Noun*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013f. URL `http://wals.info/chapter/89`.

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `http://wals.info/`.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 369–377. Association for Computational Linguistics, 2009.

Jianhua Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics, 2011.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97, 2013.

Jens Nilsson and Joakim Nivre. Malteval: an evaluation and visualization tool for dependency parsing. In *Proceedings of Language Resources and Evaluation Conference*, 2008.

Joakim Nivre. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553, 2008.

Joakim Nivre. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer, 2015.

Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of Language Resources and Evaluation Conference*, volume 6, pages 2216–2219, 2006.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135, 2007.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of Language Resources and Evaluation Conference*, 2012.

David A Smith and Jason Eisner. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 822–831. Association for Computational Linguistics, 2009.

Anders Søgaard. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 682–686. Association for Computational Linguistics, 2011.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. Statistical parsing of morphologically rich languages (spmrl): what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12. Association for Computational Linguistics, 2010.

Daniel Zeman. Reusable tagset conversion using tagset drivers. In *Proceedings of Language Resources and Evaluation Conference*, 2008.

Daniel Zeman and Philip Resnik. Cross-language parser adaptation between related languages. In *IJCNLP*, pages 35–42, 2008.