

Predicting Success in Machine Translation

Alexandra Birch **Miles Osborne** **Philipp Koehn**

a.c.birch-mayne@sms.ed.ac.uk miles@inf.ed.ac.uk pkoehn@inf.ed.ac.uk

School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB, UK

Abstract

The performance of machine translation systems varies greatly depending on the source and target languages involved. Determining the contribution of different characteristics of language pairs on system performance is key to knowing what aspects of machine translation to improve and which are irrelevant. This paper investigates the effect of different explanatory variables on the performance of a phrase-based system for 110 European language pairs. We show that three factors are strong predictors of performance in isolation: the amount of reordering, the morphological complexity of the target language and the historical relatedness of the two languages. Together, these factors contribute 75% to the variability of the performance of the system.

1 Introduction

Statistical machine translation (SMT) has improved over the last decade of intensive research, but for some language pairs, translation quality is still low. Certain systematic differences between languages can be used to predict this. Many researchers have speculated on the reasons why machine translation is hard. However, there has never been, to our knowledge, an analysis of what the actual contribution of different aspects of language pairs is to translation performance. This understanding of where the difficulties lie will allow researchers to know where to most gainfully direct their efforts to improving the current models of machine translation.

Many of the challenges of SMT were first outlined by Brown et al. (1993). The original IBM Models were broken down into separate translation

and distortion models, recognizing the importance of word order differences in modeling translation. Brown et al. also highlighted the importance of modeling morphology, both for reducing sparse counts and improving parameter estimation and for the correct production of translated forms. We see these two factors, reordering and morphology, as fundamental to the quality of machine translation output, and we would like to quantify their impact on system performance.

It is not sufficient, however, to analyze the morphological complexity of the source and target languages. It is also very important to know how similar the morphology is between the two languages, as two languages which are morphologically complex in very similar ways, could be relatively easy to translate. Therefore, we also include a measure of the family relatedness of languages in our analysis.

The impact of these factors on translation is measured by using linear regression models. We perform the analysis with data from 110 different language pairs drawn from the Europarl project (Koehn, 2005). This contains parallel data for the 11 official language pairs of the European Union, providing a rich variety of different language characteristics for our experiments. Many research papers report results on only one or two languages pairs. By analyzing so many language pairs, we are able to provide a much wider perspective on the challenges facing machine translation. This analysis is important as it provides very strong motivation for further research.

The findings of this paper are as follows: (1) each of the main effects, reordering, target language complexity and language relatedness, is a highly significant predictor of translation performance, (2) individually these effects account for just over a third of

the variation of the BLEU score, (3) taken together, they account for 75% of the variation of the BLEU score, (4) when removing Finnish results as outliers, reordering explains the most variation, and finally (4) the morphological complexity of the source language is uncorrelated with performance, which suggests that any difficulties that arise with sparse counts are insignificant under the experimental conditions outlined in this paper.

2 Europarl

In order to analyze the influence of different language pair characteristics on translation performance, we need access to a large variety of comparable parallel corpora. A good data source for this is the Europarl Corpus (Koehn, 2005). It is a collection of the proceedings of the European Parliament, dating back to 1996. Version 3 of the corpus consists of up to 44 million words for each of the 11 official languages of the European Union: Danish (da), German (de), Greek (el), English (en), Spanish (es), Finnish (fi), French (fr), Italian (it), Dutch (nl), Portuguese (pt), and Swedish (sv).

In trying to determine the effect of properties of the languages involved in translation performance, it is very important that other variables be kept constant. Using Europarl, the size of the training data for the different language pairs is very similar, and there are no domain differences as all sentences are roughly trained on translations of the same data.

3 Morphological Complexity

The morphological complexity of the language pairs involved in translation is widely recognized as one of the factors influencing translation performance. However, most statistical translation systems treat different inflected forms of the same lemma as completely independent of one another. This can result in sparse statistics and poorly estimated models. Furthermore, different variations of the lemma may result in crucial differences in meaning that affect the quality of the translation.

Work on improving MT systems' treatment of morphology has focussed on either reducing word forms to lemmas to reduce sparsity (Goldwater and McClosky, 2005; Talbot and Osborne, 2006) or including morphological information in decod-

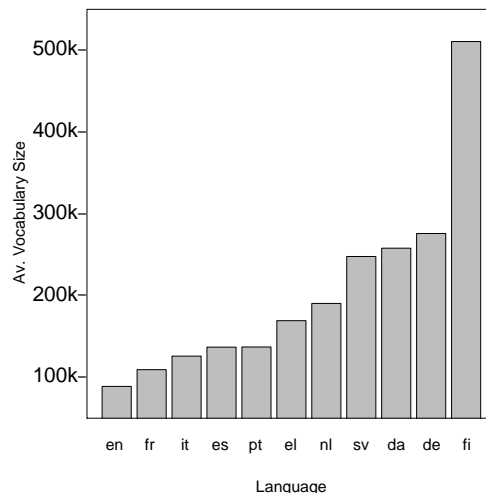


Figure 1. Average vocabulary size for each language.

ing (Dyer, 2007).

Although there is a significant amount of research into improving the treatment of morphology, in this paper we aim to discover the effect that different levels of morphology have on translation. We measure the amount of morphological complexity that exists in both languages and then relate this to translation performance.

Some languages seem to be intuitively more complex than others, for instance Finnish appears more complex than English. There is, however, no obvious way of measuring this complexity. One method of measuring complexity is by choosing a number of hand-picked, intuitive properties called *complexity indicators* (Bickel and Nichols, 2005) and then to count their occurrences. Examples of morphological complexity indicators could be the number of inflectional categories or morpheme types in a typical sentence. This method suffers from the major drawback of finding a principled way of choosing which of the many possible linguistic properties should be included in the list of indicators.

A simple alternative employed by Koehn (2005) is to use vocabulary size as a measure of morphological complexity. Vocabulary size is strongly influenced by the number of word forms affected by number, case, tense etc. and it is also affected by the number of agglutinations in the language. The complexity of the morphology of languages can therefore be approached by looking at vocabulary size.

Figure 1 shows the vocabulary size for all relevant languages. Each language pair has a slightly different parallel corpus, and so the size of the vocabularies for each language needs to be averaged. You can see that the size of the Finnish vocabulary is about six times larger (510,632 words) than the English vocabulary size (88,880 words). The reason for the large vocabulary size is that Finnish is characterized by a rich inflectional morphology, and it is typologically classified as an agglutinative-fusional language. As a result, words are often polymorphemic, and become remarkably long.

4 Language Relatedness

The morphological complexity of each language in isolation could be misleading. Large differences in morphology between two languages could be more relevant to translation performance than a complex morphology that is very similar in both languages. Languages which are closely related could share morphological forms which might be captured reasonably well in translation models. We include a measure of language relatedness in our analyses to take this into account.

Comparative linguistics is a field of linguistics which aims to determine the historical relatedness of languages. Lexicostatistics, developed by Morris Swadesh in the 1950s (Swadesh, 1955), is an approach to comparative linguistics that is appropriate for our purposes because it results in a quantitative measure of relatedness by comparing lists of lexical cognates.

The lexicostatic percentages are extracted as follows. First, a list of universal culture-free meanings are generated. Words are then collected for these meanings for each language under consideration. Lists for particular purposes have been generated. For example, we use the data from Dyen et al. (1992) who developed a list of 200 meanings for 84 Indo-European languages. Cognacy decisions are then made by a trained linguist. For each pair of lists the cognacy of a form can be positive, negative or indeterminate. Finally, the lexicostatic percentage is calculated. This percentage is related to the proportion of meanings for a particular language pair that are cognates, i.e. relative to the total without indeterminacy. Factors such as borrowing, tradition and

Language	“animal”	“black”
French	animal	noir
Italian	animale	nero
Spanish	animal	negro
English	animal	black
German	tier	schwarz
Swedish	djur	svart
Danish	dyr	sort
Dutch	dier	zwart

Table 1. An example from the (Dyen et al., 1992) cognate list.

taboo can skew the results.

A portion of the Dyen et al. (1992) data set is shown in Table 1 as an example. From this data a trained linguist would calculate the relatedness of French, Italian and Spanish as 100% because their words for “animal” and “black” are cognates. The Romance languages share one cognate with English, “animal” but not “black”, which means that the lexicostatic percentage here would be 50%, and no cognates with the rest of the languages, 0%.

We use the Dyen lexicostatic percentages as our measure of language relatedness or similarity for all bidirectional language pairs except for Finnish, for which there is not data. Finnish is a Finno-Ugric language and is not part of the Indo-European language family and is therefore not included in the Dyen results. We were not able to recreate the conditions of this study to generate the data for Finnish - expert linguists with knowledge of all the languages would be required. Excluding Finnish would have been a shame as it is an interesting language to look at, however we took care to confirm which effects found in this paper still held when excluding Finnish. Not being part of the Indo-European languages means that its historical similarity with our other languages is very low. For example, English would be more closely related to Hindu than to Finnish. We therefore assume that Finnish has zero similarity with the other languages in the set.

Figure 2 shows the symmetric matrix of language relatedness, where the width of the square is proportional to the value of relatedness. Finnish is the language which is least related to the other languages and has a relatedness score of 0%. Spanish-Portuguese is the most related language pair with a

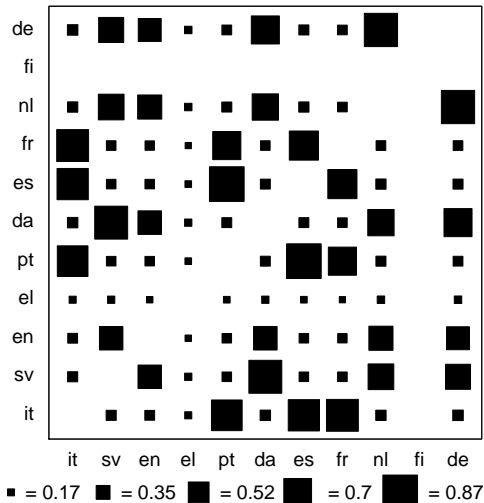


Figure 2. Language relatedness - the width of the squares indicates the lexicostatical relatedness.

score of 0.87%.

A measure of family relatedness should improve our understanding of the relationship between morphological complexity and translation performance.

5 Reordering

Reordering refers to differences in word order that occur in a parallel corpus and the amount of reordering affects the performance of a machine translation system. In order to determine how much it affects performance, we first need to measure it.

5.1 Extracting Reorderings

Reordering is largely driven by syntactic differences between languages and can involve complex rearrangements between nodes in synchronous trees. Modeling reordering exactly would require a synchronous tree-substitution grammar. This representation would be sparse and heterogeneous, limiting its usefulness as a basis for analysis. We make an important simplifying assumption in order for the detection and extraction of reordering data to be tractable and useful. We assume that reordering is a binary process occurring between two blocks that are adjacent in the source. This is similar to the ITG constraint (Wu, 1997), however our reorderings are not dependent on a synchronous grammar or a derivation which covers the sentences. There are also similarities with the Human-Targeted Transla-

tion Edit Rate metric (HTER) (Snover et al., 2006) which attempts to find the minimum number of human edits to correct a hypothesis, and admits moving blocks of words, however our algorithm is automatic and does not consider inserts or deletes.

Before describing the extraction of reorderings we need to define some concepts. We define a *block* A as consisting of a source span, $A_{\bar{s}}$, which contains the positions from A_{smin} to A_{smax} and is aligned to a set of target words. The minimum and maximum positions (A_{tmin} and A_{tmax}) of the aligned target words mark the block’s target span, $A_{\bar{t}}$.

A reordering r consists of the two blocks r_A and r_B , which are adjacent in the source and where the relative order of the blocks in the source is reversed in the target. More formally:

$$r_{A_{\bar{s}}} < r_{B_{\bar{s}}}, \quad r_{A_{\bar{t}}} > r_{B_{\bar{t}}}, \quad r_{A_{smax}} = r_{B_{smin}} - 1$$

A consistent block means that between A_{tmin} and A_{tmax} there are no target word positions aligned to source words outside of the block’s source span $A_{\bar{s}}$. A reordering is consistent if the block projected from $r_{A_{smin}}$ to $r_{B_{smax}}$ is consistent.

The following algorithm detects reorderings and determines the dimensions of the blocks involved. We step through all the source words, and if a word is reordered in the target with respect to the previous source word, then a reordering is said to have occurred. These two words are initially defined as the blocks A and B . Then the algorithm attempts to grow block A from this point towards the source starting position, while the target span of A is greater than that of block B , and the new block A is consistent. Finally it attempts to grow block B towards the source end position, while the target span of B is less than that of A and the new reordering is inconsistent.

See Figure 3 for an example of a sentence pair with two reorderings. Initially a reordering is detected between the Chinese words aligned to “from” and “late”. The block A is grown from “late” to include the whole phrase pair “late last night”. Then the block B is grown from “from” to include “Beijing” and stops because the reordering is then consistent. The next reordering is detected between “arrived in” and “Beijing”. We can see that block A attempts to grow as large a block as possible and block

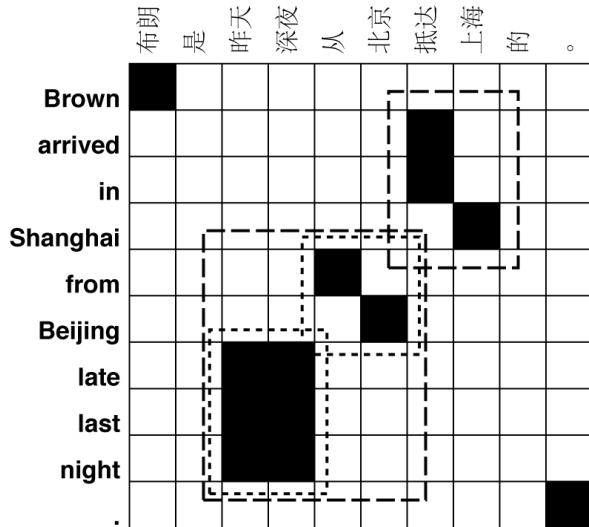


Figure 3. A sentence pair from the test corpus, with its alignment. Two reorderings are shown with two different dash styles.

B attempts to grow the smallest block possible. The reorderings thus extracted would be comparable to those of a right-branching ITG with inversions. This allows for syntactically plausible embedded reorderings. This algorithm has the worst case complexity of $O(\frac{n^2}{2})$ when the words in the target occur in the reverse order to the words in the source.

5.2 Measuring Reordering

Our reordering extraction technique allows us to analyze reorderings in corpora according to the distribution of reordering widths. In order to facilitate the comparison of different corpora, we combine statistics about individual reorderings into a sentence level metric which is then averaged over a corpus.

$$RQuantity = \frac{\sum_{r \in R} |r_{A_s}| + |r_{B_s}|}{I}$$

where R is the set of reorderings for a sentence, I is the source sentence length, A and B are the two blocks involved in the reordering, and $|r_{A_s}|$ is the size or span of block A on the source side. $RQuantity$ is thus the sum of the spans of all the reordering blocks on the source side, normalized by the length of the source sentence.

	RQuantity
Europarl, auto align	0.620
WMT06 test, auto align	0.647
WMT06 test, manual align	0.668

Table 2. The reordering quantity for the different reordering corpora for DE-EN.

5.3 Automatic Alignments

Reorderings extracted from manually aligned data can be reliably assumed to be correct. The only exception to this is that embedded reorderings are always right branching and these might contradict syntactic structure. In this paper, however, we use alignments that are automatically extracted from the training corpus using GIZA++. Automatic alignments could give very different reordering results. In order to justify using reordering data extracted from automatic alignments, we must show that they are similar enough to gold standard alignments to be useful as a measure of reordering.

5.3.1 Experimental Design

We select the German-English language pair because it has a reasonably high level of reordering. A manually aligned German-English corpus was provided by Chris Callison-Burch and consists of the first 220 sentences of test data from the 2006 ACL Workshop on Machine Translation (WMT06) test set. This test set is from a held out portion of the Europarl corpus.

The automatic alignments were extracted by appending the manually aligned sentences on to the respective Europarl v3 corpora and aligning them using GIZA++ (Och and Ney, 2003) and the grow-final-diag algorithm (Koehn et al., 2003).

5.3.2 Results

In order to use automatic alignments to extract reordering statistics, we need to show that reorderings from automatic alignments are comparable to those from manual alignments.

We first look at global reordering statistics and then we look in more detail at the reordering distribution of the corpora. Table 2 shows the amount of reordering in the WMT06 test corpora, with both manual and automatic alignments, and in the automatically aligned Europarl DE-EN parallel corpus.

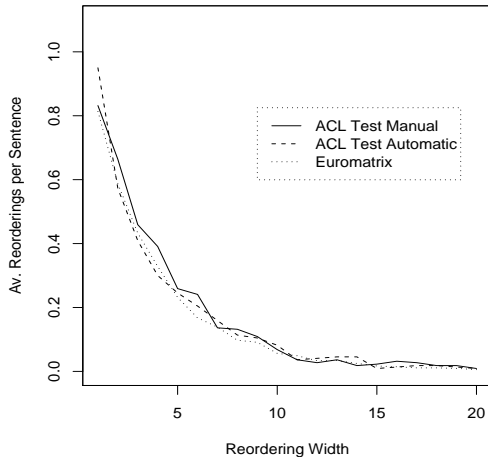


Figure 4. Average number of reorderings per sentence mapped against the total width of the reorderings for DE-EN.

We can see that all three corpora show a similar amount of reordering.

Figure 4 shows that the distribution of reorderings between the three corpora is also very similar. These results provide evidence to support our use of automatic reorderings in lieu of manually annotated alignments. Firstly, they show that our WMT06 test corpus is very similar to the Europarl data, which means that any conclusions that we reach using the WMT06 test corpus will be valid for the Europarl data. Secondly, they show that the reordering behavior of this corpus is very similar when looking at automatic vs. manual alignments.

Although differences between the reorderings detected in the manually and automatically aligned German-English corpora are minor, there we accept that there could be a language pair whose real reordering amount is very different to the expected amount given by the automatic alignments. A particular language pair could have alignments that are very unsuited to the stochastic assumptions of the IBM or HMM alignment models. However, manually aligning 110 language pairs is impractical.

5.4 Amount of reordering for the matrix

Extracting the amount of reordering for each of the 110 language pairs in the matrix required a sampling approach. We randomly extracted a subset of 2000 sentences from each of the parallel training corpora. From this subset we then extracted the av-

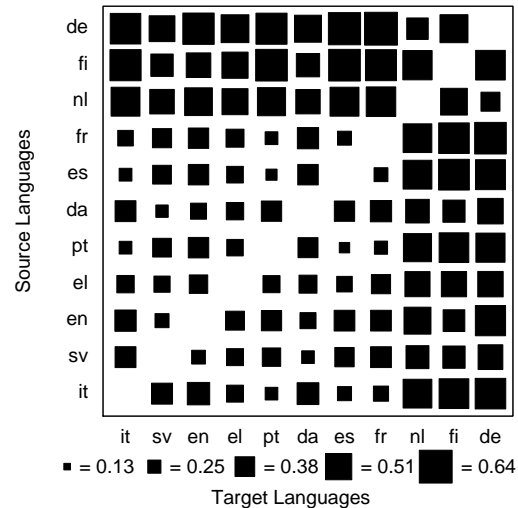


Figure 5. Reordering amount - the width of the squares indicates the amount of reordering or RQuantity.

erage RQuantity.

In Figure 5 the amount of reordering for each of the language pairs is proportional to the width of the relevant square. Note that the matrix is not quite symmetrical - reordering results differ depending on which language is chosen to measure the reordering span. The lowest reordering scores are generally for languages in the same language group (like Portuguese-Spanish, 0.20, and Danish-Swedish, 0.24) and the highest for languages from different groups (like German-French, 0.64, and Finnish-Spanish, 0.61).

5.5 Language similarity and reordering

In this paper we use linear regression models to determine the correlation and significance of various explanatory variables with the dependent variable, the BLEU score. Ideally the explanatory variables involved should be independent of each other, however the amount of reordering in a parallel corpus could easily be influenced by family relatedness. We investigate the correlation between these variables.

Figure 6 shows the plot of the reordering amount against language similarity. The regression is highly significant and has an R^2 of 0.2347. This means that reordering is correlated with language similarity and that 23% of reordering can be explained by language similarity.

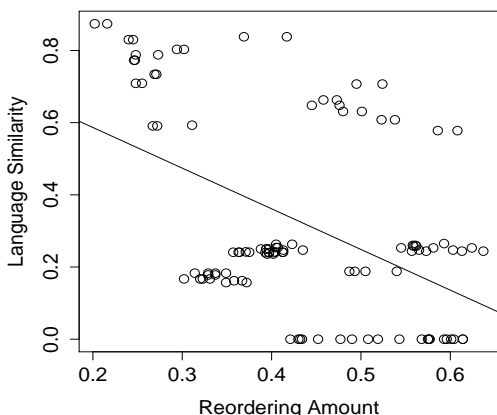


Figure 6. Reordering compared to language similarity with regression.

6 Experimental Design

We used the phrase-based model Moses (Koehn et al., 2007) for the experiments with all the standard settings, including a lexicalized reordering model, and a 5-gram language model. Tests were run on the ACL WSMT 2008 test set (Callison-Burch et al., 2008).

6.1 Evaluation of Translation Performance

We use the BLEU score (Papineni et al., 2002) to evaluate our systems. While the role of BLEU in machine translation evaluation is a much discussed topic, it is generally assumed to be an adequate metric for comparing systems of the same type.

Figure 7 shows the BLEU score results for the matrix. Comparing this figure to Figure 5 there seems to be a clear negative correlation between reordering amount and translation performance.

6.2 Regression Analysis

We perform multiple linear regression analyses using measures of morphological complexity, language relatedness and reordering amount as our independent variables. The dependent variable is the translation performance metric, the BLEU score.

We then use a t-test to determine whether the coefficients for the independent variables are reliably different from zero. We also test how well the model explains the data using an R^2 test. The two-tailed significance levels of coefficients and R^2 are also

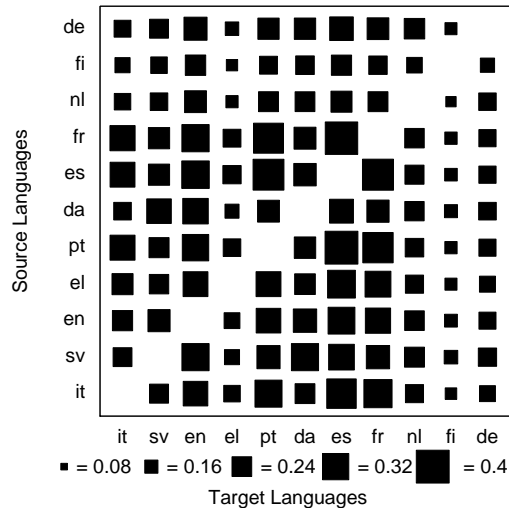


Figure 7. System performance - the width of the squares indicates the system performance in terms of the BLEU score.

Explanatory Variable	Coefficient	
Target Vocab. Size	-3.885	***
Language Similarity	3.274	***
Reordering Amount	-1.883	***
Target Vocab. Size ²	1.017	***
Language Similarity ²	-1.858	**
Interaction: Reord/Sim	-1.4536	***

Table 3. The impact of the various explanatory features on the BLEU score via their coefficients in the minimal adequate model.

given where * means $p < 0.05$, ** means $p < 0.01$, and *** means $p < 0.001$.

7 Results

7.1 Combined Model

The first question we are interested in answering is which factors contribute most and how they interact. We fit a multiple regression model to the data. The source vocabulary size has no significant effect on the outcome. All explanatory variable vectors were normalized to be more comparable.

In Table 3 we can see the relative contribution of the different features to the model. Source vocabulary size did not contribute significantly to the explanatory power of this multiple regression model and was therefore not included. The fraction of the variance explained by the model, or its goodness of fit, the R^2 , is 0.750 which means that 75% of the

variation in BLEU can be explained by these three factors. The interaction of reordering amount and language relatedness is the product of the values of these two features, and in itself it is an important explanatory feature.

To make sure that our regression is valid, we need to consider the special case of Finnish. Data points where Finnish is the target language are outliers. Finnish has the lowest language similarity with all other languages, and the largest vocabulary size. It also has very high amounts of reordering, and the lowest BLEU scores when it is the target language. The multiple regression of Table 3 where Finnish as the source and target language is excluded, shows that all the effects are still very significant, with the model's R^2 dropping only slightly to 0.68.

The coefficients of the variables in the multiple regression model have only limited usefulness as a measure of the impact of the explanatory variables in the model. One important factor to consider is that if the explanatory variables are highly correlated, then the values of the coefficients are unstable. The model could attribute more importance to one or the other variable without changing the overall fit of the model. This is the problem of multicollinearity. Our explanatory variables are all correlated, but a large amount of this correlation can be explained by looking at language pairs with Finnish as the target language. Excluding these data points, only language relatedness and reordering amount are still correlated, see Section 5.5 for more details.

7.2 Contribution in isolation

In order to establish the relative contribution of variables, we isolate their impact on the BLEU score by modeling them in separate linear regression models.

Figure 8 shows a simple regression model over the plot of BLEU scores against target vocabulary size. This figure shows groups of data points with the same target language in almost vertical lines. Each language pair has a separate parallel training corpus, but the target vocabulary size for one language will be very similar in all of them. The variance in BLEU amongst the group with the same target language is then largely explained by the other factors, similarity and reordering.

Figure 9 shows a simple regression model over the plot of BLEU scores against source vocabulary size.

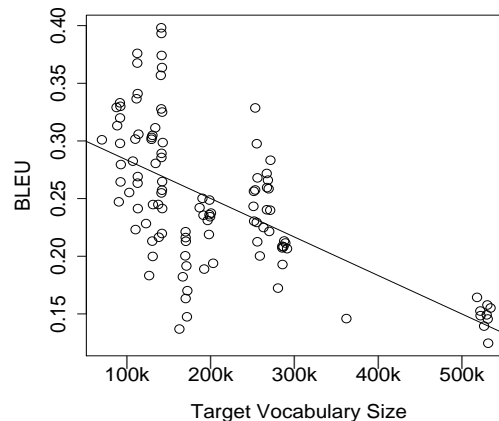


Figure 8. BLEU score of experiments compared to target vocabulary size showing regression

This regression model shows that in isolation source vocabulary size is significant ($p < 0.05$), but that this is due to the distorting effect of Finnish. Excluding results that include Finnish, there is no longer any significant correlation with BLEU. The source morphology might be significant for models trained on smaller data sets, where model parameters are more sensitive to sparse counts.

Figure 10 shows the simple regression model over the plot of BLEU scores against the amount of reordering. This graph shows that with more reordering, the performance of the translation model reduces. Data points with low levels of reordering and high BLEU scores tend to be language pairs where both languages are Romance languages. High BLEU scores with high levels of reordering tend to have German as the source language and a Romance language as the target.

Figure 11 shows the simple regression model over the plot of BLEU scores against the amount of language relatedness. The left hand line of points are the results involving Finnish. The vertical group of points just to the right, are results where Greek is involved. The next set of points are the results where the translation is between Germanic and Romance languages. The final cloud to the right are results where languages are in the same family, either within the Romance or the Germanic languages.

Table 4 shows the amount of the variance of BLEU explained by the different models. As these

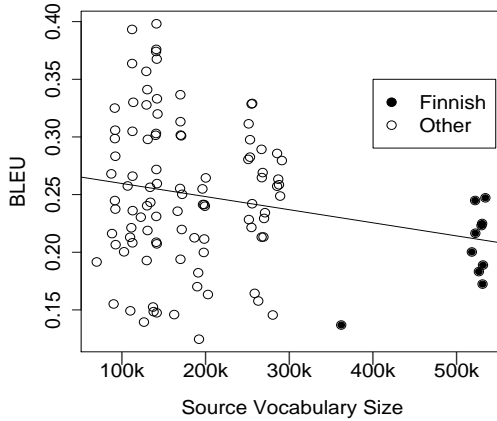


Figure 9. BLEU score of experiments compared to source vocabulary size highlighting the Finnish source vocabulary data points. The regression includes Finnish in the model.

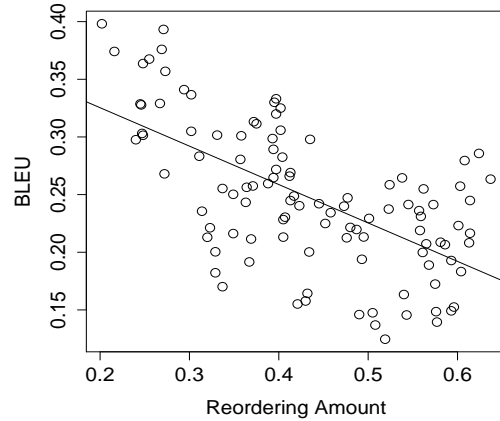


Figure 10. BLEU score of experiments compared to amount of reordering.

Explanatory Variable	R^2
Target Vocab. Size	0.388 ***
Reordering Amount	0.384 ***
Language Similarity	0.366 ***
Source Vocab. Size	0.045 *
Excluding Finnish	
Target Vocab. Size	0.219 ***
Reordering Amount	0.332 ***
Language Similarity	0.188 ***
Source Vocab. Size	0.007

Table 4. Goodness of fit of different simple linear regression models which use just one explanatory variable. The significance level represents the level of probability that the regression is appropriate. The second set of results excludes Finnish in the source and target language.

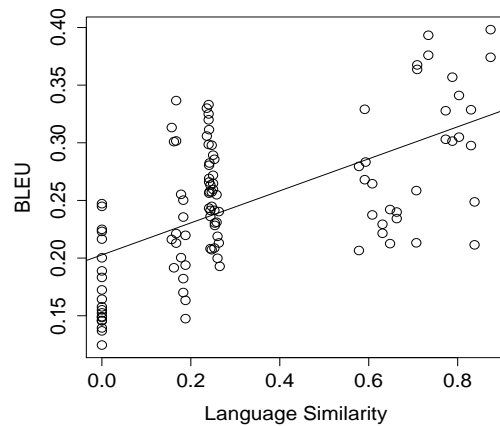


Figure 11. BLEU score of experiments compared to language relatedness.

are simple regression models, with just one explanatory variable, multicollinearity is avoided. This table shows that each of the main effects explains about a third of the variance of BLEU, which means that they can be considered to be of equal importance. When Finnish examples are removed, only reordering retains its power, and target vocabulary and language similarity reduce in importance and source vocabulary size no longer correlates with performance.

8 Conclusion

We have broken down the relative impact of the characteristics of different language pairs on trans-

lation performance. The analysis done is able to account for a large percentage (75%) of the variability of the performance of the system, which shows that we have captured the core challenges for the phrase-based model. We have shown that their impact is about the same, with reordering and target vocabulary size each contributing about 0.38%.

These conclusions are only strictly relevant to the model for which this analysis has been performed, the phrase-based model. However, we suspect that the conclusions would be similar for most statistical machine translation models because of their dependence on automatic alignments. This will be the topic of future work.

References

- Balthasar Bickel and Johanna Nichols, 2005. *The World Atlas of Language Structures*, chapter Inflectional synthesis of the verb. Oxford University Press.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.
- Isidore Dyen, Joseph Kruskal, and Paul Black. 1992. An indoeuropean classification, a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5).
- Chris Dyer. 2007. The 'noisier channel': Translation from morphologically complex languages. In *Proceedings on the Workshop on Statistical Machine Translation*, Prague, Czech Republic.
- Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 127–133, Edmonton, Canada. Association for Computational Linguistics.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics Companion Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):9–51.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.
- M Snover, B Dorr, R Schwartz, L Micciulla, and J Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA*.
- Morris Swadesh. 1955. Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings American Philosophical Society*, volume 96, pages 452–463.
- David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of the Association of Computational Linguistics*, Sydney, Australia.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.