Optimal Bilingual Data for French–English PB-SMT

Sylwia Ozdowska and **Andy Way**
National Centre for Language Technology
Dublin City University
Glasnevin, Dublin 9, Ireland
{sozdowska,away}@computing.dcu.ie

Abstract

We investigate the impact of the original source language (SL) on French–English PB-SMT. We train four configurations of a state-of-the-art PB-SMT system based on French–English parallel corpora which differ in terms of the original SL, and conduct experiments in both translation directions. We see that data containing original French and English translated from French is optimal when building a system translating from French into English. Conversely, using data comprising exclusively French and English translated from several other languages is suboptimal regardless of the translation direction. Accordingly, the clamour for more data needs to be tempered somewhat; unless the quality of such data is controlled, more training data can cause translation performance to decrease drastically, by up to 38% relative BLEU in our experiments.

1 Introduction

Statistical machine translation (SMT) systems are trained on sentence-aligned parallel corpora consisting of translated texts. In the simplest case the translation direction is constant so that one part of the parallel corpus is the translation of the other. In more complex cases, either some texts may have been translated from language A to language B and others the other way round, or more than two languages are involved and both parts were translated from one another or several other languages. This is the case of corpora involving European languages, such as the Europarl corpus (Koehn, 2005)[1] or the Acquis Communautaire corpus (Steinberger et al., 2006)[2], which comprise texts coming from institutions of the European Union. They are amongst the largest and most widely used corpora in SMT.

Typically, given a corpus in language A, its version in language B and an SMT system translating from A to B, SMT training assumes A to be the source language (SL) and B to be the target language (TL) irrespective of the original translation direction or languages involved. In other words, it is assumed that the original SL does not matter when training an SMT system which aims to translate from language A to language B.

Following a brief overview of related work (section 2), we investigate the impact of the original SL with regard to French–English translation. Our experimental objective is to compare training configurations which differ in terms of the original SL by measuring French-to-English and English-to-French translation quality of a state-of-the-art phrase-based SMT (PB-SMT) system. We train four different configurations of the same PB-SMT system based on French–English parallel corpora which differ in terms of the original SL (sections 3 and 4) and carry out translation experiments from French into English and from English into French (section 5). We evaluate each output using standard evaluation metrics, compare the results and present our findings (section 6). We then conclude and give some avenues for future work (section 7).

2 Related work

Although it is a big topic of interest in translation studies, directionality seems to have been almost

© 2009 European Association for Machine Translation.

[1] http://www.iccs.inf.ed.ac.uk/\~pkoehn/publications/europarl/
[2] http://wt.jrc.it/lt/Acquis/

Proceedings of the 13th Annual Conference of the EAMT, pages 96–103,
Barcelona, May 2009

totally neglected in SMT research. In the context of SMT, the question of directionality is not addressed directly. Instead, Wu and Wang (2007) propose a method for PB-SMT based on a pivot language to translate between languages for which there exist only small amounts of or no parallel data. They show for instance that good translation quality can be achieved when using Greek as pivot to translate from French into Spanish. In the context of translation studies, Teubert (1996) claims that if a text is translated from language A into languages B and C, then the B and C versions are likely to bear more resemblance to A than to each other. More generally, it seems to be acknowledged that translated texts should not be viewed as bidirectional resources (Bowker, 2003).

Therefore, it seems reasonable to think that there might be a correlation between MT quality from language A to language B and the actual "translational status" of languages A and B in the training corpus and the testset. More precisely, our hypothesis is that using data where A is the original SL and B the TL is likely to be the optimal configuration with regard to MT quality from A to B. Conversely, the case where neither A nor B is the original SL, meaning that both are translated from other languages, is expected to be the suboptimal configuration.

In order to test whether this hypothesis holds true, we perform training on four sub-corpora extracted from the Europarl corpus, namely: a) no criterion is imposed on the original SL, b) the original SL is neither French nor English, c) the original SL is French and d) the original SL is English. We then measure translation accuracy according to a range of automatic MT evaluation metrics.

3 Data

3.1 The Europarl corpus

In the experiments we present here, we used an in-house version of the French–English part of the original Europarl corpus.[3] Some manual changes were made to the original files to correct misalignments (*e.g.* extra, empty speaker turns) prior to sentence alignment performed automatically with a technique based on (Gale and Church, 1993). The alignments at sentence level were tagged with information on the original SL.

[3]Thanks to Mary Hearne for providing us with the modified version of the Europarl corpus.

Table 1 gives the spread in terms of number of sentence pairs according to the original SL. It can be seen that out of 1,391,222 French–English sentence pairs appearing in the corpus, only 164,648 were originally translated from French into English and 235,102 the other way round. For 715,090 sentence pairs, the original SL is neither French nor English, meaning that both the French part and the English part of the corpus contain translations from the other 20 source languages represented. Hence translated French and translated English account for at least 50% of the corpus; the original source language is unknown (NONE and EMPTY) for 276,382 sentence pairs.

original SL	sentence pairs
NONE	259540
Enlish	235102
German	201195
French	164648
Dutch	121045
Spanish	84285
Italian	68259
Swedish	56377
Portugese	49183
Greek	43541
Finnish	31334
Danish	25506
EMPTY	16842
Polish	15714
Czech	4613
Hungarian	4589
Slovak	2702
Lithuanian	2034
Latvian	1388
Slovenian	1380
Maltese	996
Estonian	949

Table 1: Repartition according to the original SL in the French–English Europarl corpus

Therefore, the French–English part of the version of the Europarl corpus our experiments are based on is made up of texts where:

- the original SL is French, and hence the English side contains English translated from French;
- or the original SL is English, and hence the French side contains French translated from English;

- or the original SL is neither French nor English, and hence both the French and the English side contains translated French or English.

3.2 Dataset extraction

In order to investigate the influence of the original SL on French–English state-of-the-art PB-SMT, we built four configurations of the same system for each translation direction based on the information on the original SL. Each configuration was built and tested using a French–English dataset (training data and testsets) extracted according to a different criterion as to the original SL. The original SL selection criteria and the contents of the four datasets extracted are described in the following section. The datasets were tokenised and lowercased for the purpose of the experiments. Moreover, only sentence pairs corresponding to a 1-to-1 alignment with lengths ranging from 5 to 40 tokens on both French and English sides were considered. We used 100,000 sentence pairs for training and 500 sentences to test each configuration and measure translation quality.

3.3 Training and test configurations

config-1 No condition is imposed on the original SL, meaning that the French part of the data and its English counterpart contain respectively:

- French translated from ¬English, French translated from English and original French;
- English translated from ¬French, English translated from French and original English.

Table 2 shows the repartition in terms of number of sentence pairs according to the original SL for the training corpus and the testset associated with config-1. It can be seen that both the training corpus and the testset show a similar spread as to the original SL.

config-2 The original SL is neither French nor English, meaning that the French part of the data and its English counterpart contain respectively:

- French translated from ¬English;
- English translated from ¬French.

Table 3 shows the repartition in terms of number of sentence pairs according to the original SL for the training corpus and the testset associated with config-2. Here again the repartition was kept as consistent as possible across the training data and the testset.

original SL	train sentences	test sentences
German	17551	116
English	16635	58
French	15697	47
NONE	12912	98
Dutch	11691	50
Spanish	6260	50
Swedish	4981	22
Italian	3974	22
Portugese	3155	15
Finnish	2772	15
Greek	2458	0
Danish	1914	7

Table 2: Config-1 – training data and testset in terms of original SL

original SL	train sentences	test sentences
German	30467	232
Dutch	21638	115
Swedish	11556	37
Spanish	11265	43
Italian	7497	14
Portugese	5092	23
Finnish	4737	25
Greek	4252	11
Danish	3496	0

Table 3: Config-2 – trainig data and testset in terms of original SL

config-3 The original SL is English, meaning that the French part of the data and its English counterpart contain respectively:

- French translated from English;
- original English.

To evaluate the performance of config-3 for French-to-English translation, we use a portion of the French part of the data (*i.e.* French translated from English) as test and the English part (*i.e.* original English) as reference. English-to-French translation evaluations are based on the same portion of the data; this time, the English part (*i.e.* original English) is used as test and the French part (*i.e.* French translated from English) as reference.

config-4 The original SL is French, meaning that the French part of the data and its English counterpart contain respectively:

- original French;
- English translated from ¬French.

To evaluate the performance of config-4 for French-to-English translation, we use a portion of the French part of the data (*i.e.* original French) as test and the English part (*i.e.* English translated from French) as reference. English-to-French translation evaluations are based on the same portion of the data; this time, the English part (*i.e.* English translated from French) is used as test and the French part (*i.e.* original French) as reference.

In addition to each individual 500-sentence testset, we also constructed one unique testset of 2000 sentences by merging the individual tests. The composition in terms of original SL of the 2000-sentence testset is given in Table 4. Overall evaluations in both translation directions are carried out based on this testset. For French-to-English, the French part is used as test and the English part as reference. For English-to-French, the latter is used as test and the former as reference.

original SL	test sentences
English	558
French	547
German	348
Dutch	165
NONE	98
Spanish	93
Swedish	59
Finnish	40
Portugese	38
Italian	36
Greek	11
Danish	7

Table 4: Test-2000 – repartition according to the original SL

4 Tools

4.1 Alignment and translation

All translation experiments are carried out using standard state-of-the-art techniques. Sentence pairs are first word-aligned using GIZA++ implementation of IBM model 4 in both source-to-target and target-to-source translation directions (Brown et al., 1993; Och and Ney, 2003) for each training set. After obtaining the intersection of these directional alignments, alignments from the union are also inserted; this insertion process is heuristics-driven (Koehn et al., 2003). Once the word alignments are finalised, all word- and phrase-pairs which are consistent with the word alignment and which comprise at most 7 words are extracted. Phrase-pairs are extracted by standard PB-SMT techniques using the Moses system (Koehn et al., 2007). A 5-gram language model is trained with SRILM (Stolcke, 2002) on the English side of the training data for French-to-English translation experiments and on the French side of the training data for English-to-French translation experiments. Finally decoding is carried out with Moses.

4.2 Minimum error rate training

Due to time constraints, we do not perform minimum error rate training (MERT) although it is now well established as a standard technique in PB-SMT (Och and Ney, 2003). Our experimental objective is to compare the relative performance of four configurations of the same system for each translation direction which differ only according to the conditions imposed on the original SL when selecting the dataset they are trained and tested on. We are not interested in the absolute performance each of these configurations achieves individually as far as the experiments presented here are concerned. Although carrying out MERT would probably have led to an increase in translation quality achieved with the different configurations that are tested, we have no reason to think that it would have resulted in a radical change as to their relative performance. However, this assumption needs to be confirmed by further experiments, which are currently ongoing (cf. footnote 4).

4.3 Evaluation

The results of the translation output are evaluated using three standard automatic evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and METEOR (Banerjee and Lavie, 2005).

5 Experiments

As described in the previous sections, we built four different configurations of the same system for two translation directions, French-to-English and English-to-French, and carried out translation

experiments. We considered the relative merits to PB-SMT of using data of which the source part actually corresponds to the original SL, meaning that the original translation direction and the translation direction to handle are consistent, *vs.* data where this condition is partially met or not met at all. We also considered the extent to which these relative merits depend on whether the translation direction is French-to-English or English-to-French.

For each translation direction, the evaluation of the different configurations was carried out in three different ways:

- in the first place, each configuration was evaluated against one 500-sentence testset selected according to the same criterion as to the original SL as the data it was trained on; therefore, the four testsets used at this stage are different from one another;
- then, each configuration is evaluated against each of the other three testsets; in other words, each configuration is evaluated against testsets where there is no or little overlap in terms of the original SL with the data it was trained on;
- finally, each configuration is evaluated against the unique 2000-sentence testset resulting from the union of all four individual testsets.

6 Results

In the following sections, we present the results and discuss the associated trends first for French-to-English and then for English-to-French. The highest scores are highlighted in bold; the lowest scores are in italics.

6.1 French-to-English

6.1.1 Individual evaluation

The translation quality of each configuration is measured individually against each 500-sentence testset. First, we give the scores (BLEU, NIST and METEOR) which each configuration achieves on its specific testset (Table 5), *i.e.* the testset which meets the same requirements as to the original SL; for instance config-1 is evaluated against test-1, config-2 against test-2, etc.

The results are consistent across all metrics. If we look for example at BLEU, we see a considerable absolute improvement of 0.0956 when moving from config-2, which achieves the lowest score (0.2008), to config-4, which performs best with a score of 0.2964. This might be due to the fact that for config-2 the French and English parts of the data bear less resemblance to each other. Both languages being translated from several other languages, they may present a higher proportion of divergences than if translated directly from one into another, thus making generalisation over the data less efficient. The second best configuration (0.2857) is config-3, *i.e.* the configuration which was trained on a corpus representing the reverse original translation direction, *i.e.* English-to-French. The third best (0.2608) is config-1 which uses data based on various original SL, thus including original French and English as well as translated French and English. Therefore, we conclude that data containing original French and English translated from French is optimal when building a system translating from French into English. Conversely, data comprising exclusively French and English translated from several other languages appears to be suboptimal.[4]

system	BLEU	NIST	METEOR
config-1	0.2608	5.9771	0.5758
config-2	*0.2008*	*5.1531*	*0.4867*
config-3	0.2857	6.4717	0.6082
config-4	**0.2964**	**6.5502**	**0.6162**

Table 5: French-to-English – evaluation on individual 500-sentence testsets

We further analyse how each configuration performs on each individual testset (Table 6). Here again the results are consistent across all metrics, and hence we present the results as measured by only one of the three metrics used in our experiments, BLEU.

system	test-1	test-2	test-3	test-4
config-1	0.2608	*0.2014*	0.2632	**0.2887**
config-2	0.2449	*0.2008*	0.2529	**0.2764**
config-3	0.2519	*0.1991*	**0.2857**	0.2695
config-4	0.2465	*0.1963*	0.2579	**0.2964**

Table 6: French-to-English – evaluation on all four individual 500-sentence testsets (BLEU)

[4]The results obtained for French-to-English by each configuration on its individual testset when MERT is performed confirm the observations made so far. Tests with MERT are currently ongoing for the experiments presented in the remainder of the paper.

We observe that config-3 and config-4 perform best on the testset which presents the same characteristics as the training data in terms of original SL: English as original SL for config-3/test-3 and French as original SL for config-4/test-4. We also note that both config-1 and config-2 achieve the best scores on test-4 rather than on the testsets that present the same characteristics as the training data in terms of the original SL, test-1 and test-2 respectively. On the other hand, all configurations achieve the lowest translation quality when it comes to translating test-2, which contains exclusively non-original French, *i.e.* French translated from languages other than English. A potential explanation for the latter observation may again lie in the resemblance between the source language being translated and the reference. It is probable that the references associated with test-4 bear a higher resemblance/are more faithful to the source since they were originally translated from French, whereas the opposite might be true for the references associated with test-1 and test-2 since only part or none of them was originally translated from French.

6.1.2 Overall evaluation

This time, each configuration is evaluated against the unique 2000-sentence testset resulting from the union of the individual testsets according to the same metrics as used previously (Table 7).

system	BLEU	NIST	METEOR
config-1-2000	**0.2542**	6.4797	0.5646
config-2-2000	*0.2424*	*6.3211*	*0.5525*
config-3-2000	0.2520	**6.5385**	0.5558
config-4-2000	0.2500	6.4331	**0.5681**

Table 7: French-to-English – evaluation on the unique 2000-sentence testset

First of all, we observe that the scores are lower when measured on the 2000-sentence testset in comparison with the individual 500-sentence testsets, for instance 0.2542 *vs.* 0.2964 for the best BLEU score. Moreover, the metrics give conflicting results. Only one score is consistent across all metrics on the one hand, and with the individual evaluations on the other hand: config-2 yields the lowest translation quality, *i.e.* 0.2424 BLEU. This confirms our previous conclusion: using data where both French and English are translated from other languages has a negative effect on MT performance and constitutes the least optimal training configuration.

Looking at the other scores, we can see that if we ignore NIST, then config-1 outperforms config-3. If we ignore METEOR, then config-3 outperforms config-4. There is a trend towards config-1 and config-3 being the best two configurations when translation is performed on a testset that mixes original French and French translated from English as well as other languages. In this respect, going back to Table 6, the following detailed observations can be drawn:

test-1: config-1>config-3>config-4>config-2
test-2: config-1>config-2>config-3>config-4
test-3: config-3>config-1>config-4>config-2
test-4: config-4>config-1>config-3>config-2

Config-1 outperforms config-3 on 3 out of 4 testsets. Config-3 outperforms config-4 on 3 out of 4 testsets. In at least one case — config-1 — the optimal results are obtained when there is an overlap in the contents of the training data and the testset in terms of original SL.

6.2 English-to-French

6.2.1 Individual evaluation

We now look at the opposite translation direction, *i.e.* English-to-French. The results are presented in Table 8. This time, config-3 is the one which matches the current translation direction since it is based on French translated from English and original English. To confirm the conclusions for French-to-English, config-3 should perform best.

system	BLEU	NIST	METEOR
config-1	0.2615	5.9315	0.5624
config-2	*0.1969*	*4.9954*	*0.4777*
config-3	0.2965	6.3787	0.5910
config-4	**0.3201**	**6.7205**	**0.6161**

Table 8: English-to-French – evaluation on individual 500-sentence testsets

As for French-to-English, scores are consistent across all evaluation metrics. Unexpectedly, the relative ranking turns out to be exactly the same as for French-to-English. Config-4 yields the highest translation quality (0.3201 BLEU) although in this case training was performed on a corpus the content of which represents the reverse translation direction with respect to the tested translation direction, meaning that the English part consists of

texts translated from French which is thus the original SL. Config-3 is second best. As previously, config-2 achieves the lowest score, *i.e.* 0.1969 BLEU. According to BLEU, there is an absolute increase of 0.1232 in performance when moving from config-2 to config-4, which corresponds to 38% relative increase. We also note that English-to-French translation yields better overall results than French-to-English on the same testset, 0.3201 BLEU *vs.* 0.2964 BLEU, which is unusual.

The performance of each configuration on each individual testset is shown in Table 9. The situation is similar as for French-to-English. Here again, config-3 and config-4 perform best on the testset which presents the same characteristics as the training data in terms of the original SL, whereas config-1 and config-2 yield the highest results on test-3 which contains original English. As previously, the lowest translation quality is obtained when translating test-2, which contains only English translated from other languages than French. Therefore, the results for English-to-French confirm the findings for the opposite translation direction.

system	test-1	test-2	test-3	test-4
config-1	0.2615	*0.1970*	**0.2814**	0.2661
config-2	0.2523	*0.1969*	**0.2731**	0.2602
config-3	0.2514	*0.1971*	**0.2965**	0.2649
config-4	0.2478	*0.2011*	0.2754	**0.3201**

Table 9: English-to-French – evaluation on all four individual 500-sentence testsets (BLEU)

6.2.2 Overall evaluation

Table 10 shows evaluation results on the 2000-sentence testset for English-to-French.

system	BLEU	NIST	METEOR
config-1	0.2517	6.3192	0.5478
config-2	*0.2459*	*6.2242*	*0.5406*
config-3	0.2525	6.3335	**0.5576**
config-4	**0.2616**	**6.4384**	0.5511

Table 10: English-to-French – evaluation on the unique 2000-sentence testset

Part of the observations we can make when looking at this table are similar to those made for the French-to-English experiments: translation quality is generally reduced compared to the evaluations made on the individual 500-sentence testsets, 0.2616 *vs.* 0.3201 BLEU score. Furthermore, the metrics give conflicting results; config-2 gives the lowest translation quality, *i.e.* 0.2459 BLEU, which is the only consistent result as far as all metrics and individual evaluations are concerned.

Looking at the other scores in Table 10, a different situation to that observed for the French-to-English direction arises. This time, if we ignore METEOR, config-4 outperforms config-3, config-3 outperforms config-1 and config-1 outperforms config-2. In other words, the tendency observed on the 2000-sentence testset is consistent with the scores measured on the individual testsets. This is quite unexpected: better translation quality is achieved although there is no overlap between the training corpus and the testset in terms of original SL. Furthermore, the contents of the training corpus were originally issued in French and translated into English, meaning that they represent the reverse translation direction with respect to the tested translation direction. We see that the detailed results are less clear-cut (more mixed) than for French-to-English upon looking at Table 9. Config-4 outperforms config-3 on 2 testsets out of 4; config-3 outperforms config-1 on 2 testsets out of 4.

7 Conclusions and Future Work

In this paper, we argued that the nature of the original SL should not be neglected as far as bilingual data for PB-SMT training is concerned. We observed that the original SL has a considerable impact on French–English PB-SMT training. First of all, using data where neither French nor English is the original SL, *i.e.* both are translated from several other languages, resulted in a clear-cut absolute decrease in translation quality in all scores, for instance up to 0.1232 in BLEU, and regardless of the translation direction considered. For French-to-English, evaluations on individual testsets showed that using data which contains as original SL the source language being translated proved to be the optimal configuration, leading to up to 0.0956 absolute increase in BLEU. However, overall evaluations on one unique testset indicated a tendency towards preferring data based on various original SLs.

System developers have not paid any attention to date to the role of the human translator in developing bilingual corpora for use as training data in PB-SMT. Our results demonstrate quite clearly

that this attitude has to change. Our findings are especially poignant to those whose mantra is "More data is better data" (cf. (Zollmann et al., 2008)), as again it is clear that what we *really* need is *better quality* data. In order to show more significant improvements in our PB-SMT systems, it appears that we might be better off paying translators to develop language pair-specific material for use as training data. Far from ever being made redundant by SMT systems, the role of the translator is even more crucial than has been acknowledged heretofore, and only closer relations between human translators and system designers are likely to lead to further improvements in translation quality in PB-SMT.

We are replicating the experiments with MERT and plan to work with a fixed language model. We will also scale up our experiments in order to investigate to what extent the observed trends are influenced by the amount of data. We will address two additional questions. Once all direct translations have been used, does it hurt to add data that was indirectly translated via another language? Given a full corpus, is it possible to improve translation quality by filtering out parts corresponding to indirect translations? Finally, we will run tests with different language pairs, particularly with languages from different families, and with different corpora provided that enough data is available.

Acknowledgements

We are grateful to Science Foundation Ireland (`http://www.sfi.ie`) grant 05/IN/1732 for funding this work.

References

Banerjee, S. and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 43th Annual Meeting of the Association of Computational Linguistics (ACL-05)*, Ann Arbor, MI, 65–72.

Bowker, L. 2003. Investigate 'reversible' translation resources: are they equally useful in both translation directions? *Speaking in Tongues: Language across Contexts and Users*, Luis Pérez Gonzáles ed. 201–224.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1993. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Doddington, G. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Human Language Technology: Notebook Proceedings*, San Diego, CA, 128–132.

Gale, W. J., and K. W. Church. 1993. A Program for Aligning Sentences in Parallel Corpora. *Computational Linguistics*, 19(3):75–102.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit X: The Tenth Machine Translation Summit*, Phuket, Thailand, 79–86.

Koehn, P., H. Hoang, A. Birch, Ch. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, 177–180.

Koehn, P., F. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*, Edmonton, Canada, 48–54.

Och, F., and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA, 11–318.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The JRC-Acquis: A multilingual Aligned Parallel Corpus with 20+ Languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2142–2147.

Stolcke, A. 2002. SRILM: an Extensible Language Modeling Toolkit. *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 901–904.

Teubert, W. 1996. Comparable or Parallel Corpora? *International Journal of Lexicography*, 9(3):239–264.

Wu, H., and H. Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Zollmann A., A. Venugopal, F. Och, and J. Ponte. 2008. A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT. In *Coling 2008, The 22nd International Conference on Computational Linguistics, Proceedings*, Manchester, UK, 1145–1152.