In Search of the Best Method for Sentence Alignment in Parallel Texts*

Alexandr Rosen

Institute of Theoretical and Computational Linguistics,
Faculty of Philosophy and Arts, Charles University, Prague
alexandr.rosen@ff.cuni.cz

Abstract After a brief account of a parallel corpus project involving many diverse languages and a summary of two previous evaluations of sentential alignment tools, results are presented from tests of three automatic aligners on English-Czech and French-Czech literary and legal texts, clean and noisy. The results confirm that an alignment tool may perform well on one type of texts and fail on another type, and indicate that near-to-perfect alignment is possible when tools providing high precision are combined with manual checking, where the proofreader can focus only on those parts of the text that were either not aligned at all, or that were aligned less reliably. Further gains in precision are shown to be feasible when alignments proposed by multiple aligners are intersected.

1 Introduction

Once we have a text and its translation, is there a way to match corresponding sentences reliably and without too much human intervention? This question has been asked before, e.g, by Langlais et al. (1998), Véronis & Langlais (2000) and, most recently, by Singh & Husain (2005). The answers do not point to a single all-purpose method. Different contexts may require different solutions and their choice should be based on a careful consideration of properties of the text pair and ways of using the result. The factors include structural distance between the two texts (how free or literal the translation is), typological distance between the two languages, size of the texts (a critical issue for statistical methods), acceptable error rate in terms of precision and recall, and acceptable amount of manual checking. Given the task to provide sentence alignment tools for a number of diverse language pairs and text genres with the obvious desideratum to reach a near-to-perfect result, an opportunistic mix is inevitable.

In Sect. 2 we provide background information on our parallel corpus project including over 20 languages with Czech as the pivot. Given a wide range of languages, distributed setting is required as linguists knowledgeable of specific language pairs are necessarily involved in the whole process of text acquisition, pre-processing, alignment and checking of the alignment results. At the

* The work reported here is supported by the Czech Ministry of Education, grant no. MSM 0021620823.

same time, common shared procedures, tools, text formats and other resources are needed for the results to be integrated into a single corpus, maintained and queried by a parallel corpus manager. The solution to this challenge aims at maximising synergy effects of the large team of linguists as experts on the individual languages, and the main coordinator, providing project management and software infrastructure.

It is alignment that largely determines the usefulness of a parallel corpus, Sect. 3 deals with this issue, listing some candidate automatic alignment methods and providing data from previous evaluations. Sect. 4 presents results of testing three aligners on available texts, comparing them with previous evaluations. Based on the results, we argue for a strategy for integrating highly reliable automatic alignment with a minimum amount of human intervention. Finally, in Sect. 5 we explore options for combining results of different aligners to obtain maximum precision as a suitable step preceding manual checking of alignment results. This seems to be the least painful way to achieve minimum error rate for all sentences, and thus a corpus with highly reliable alignment. Sect. 6 summarises conclusions and suggests what should be done next.

2 The project *InterCorp*

This parallel corpus project[1] is not unique in involving a larger number of languages: a portion of the Uppsala and Oslo's universities' OPUS project[2] (Tiedemann & Nygaard 2004) includes 60 languages, and the Acquis Communautaire parallel corpus,[3] compiled at the European Commission's Joint Research Centre at Ispra (Italy), includes 20 languages (Erjavec et al. 2005). Still, there are at least three aspects that make it different: distributed setup, preference for a balanced choice of text types, and a fair amount of texts with manually checked alignment.

The project is based upon an idea of integrating expertise and efforts of a number of project participants into a common shared resource, providing them with the necessary infrastructure and complying with their preferences: although for some languages it may not be easy to acquire enough texts, preference is given to balance rather then quantity, with literary texts and – at least in the initial stages – Czech originals the priority.

Due to the substantial involvement of a large number of participants, a distributed mode of pre-processing is inevitable: the current institutional participants consist of twelve departments and institutes, two of them outside Charles University, each responsible for at least one language pair. There are at least 20 such pairs, all of them including Czech as the pivot language, the other languages being as diverse as Arabic and Chinese. Guidance, coordination and

[1] See `https://trnka.ff.cuni.cz/ucnk/intercorp/`, only Czech version is available at the time of writing.

[2] `http://logos.uio.no/opus/`

[3] `http://www.fi.muni.cz/~zizka/Langtech/`

support are provided by the main coordinator, the Institute of the Czech National Corpus.

Some participants have already built parallel corpora of various sizes and fashions, using *ParaConc* as the segmentation, alignment and search tool (Barlow 1999, 2002),[4] and they continue to do so within the project. The challenge is to reconcile distributed pre-processing with the need to store, maintain and access the corpus at one place at the final stage. Thus, a battery of tools take care of the smooth transition between the 'local' format required by *ParaConc*, *MS Word* and other PC-based software and the canonical format adopted for the common shared corpus, making sure that – in the worst case when an electronic source is not available – a paper document goes through OCR, proofreading, conversion to tagged text, segmentation into paragraphs and sentences, sentential alignment and alignment checking, ending up in the XML format with stand-off alignment annotation.

3 Alignment

In most cases, reliable alignment of sentences is a necessary condition for a useful parallel corpus. Indeed, a parallel corpus is only as good as its alignment. In order to minimise the amount of manual checking, it is worthwhile to search for the best methods of automatic alignment.

The default alignment tool is an implementation of Church and Gale's algorithm (Gale & Church 1991a), integrated with *Paraconc*. The obvious question is whether there is a better alternative.

There are some published reports on comparative evaluation of sentential alignment. In *ARCADE*, a major project (Langlais et al. 1998, Véronis & Langlais 2000), a number of important issues are brought up, but today the choice of evaluated tools would probably be different. Six systems were tested on French-English texts of various types (over 1M words per language), including an abridged translation of Jules Verne's novel *From the Earth to the Moon*. Interestingly, this was a pitfall for all systems except one, which was based on a combination of techniques including sentence length, recognition of cognates (identical or similar strings) and bilingual lexicon look-up.

More recently, results of another detailed evaluation were reported by Singh & Husain (2005) (henceforth S&H). S&H aimed for systematic evaluation of four aligners on different text types. They used a mix of 21 samples from three different English-Hindi corpora, systematically varied in terms of size and noise (sentences added at random from other corpora). Due to practical constraints, only 1:1 links were considered. Three of the four systems have also been used in our evaluation, so results presented by S&H are examined more closely below.

Two of the four systems are based on methods matching most likely sentences by comparing their lengths, either in words (Brown et al. 1991) – hence-

[4] `http://www.athel.com/para.html`

forth **Brn** – or characters (Gale & Church 1991b) – **GC**.[5] Both systems are quite fast and language-independent, but they assume some fixed points: **Brn** expects at least some sentences to be previously aligned, while **GC** requires identification and alignment of paragraphs ("hard regions") across the texts.[6] The other two systems use word correspondences: Melamed (1997) – **Mmd** – gives better results with a bilingual dictionary, although cognates such as punctuation, numbers and similar words may suffice,[7] while Moore (2002) – **Mre** – generates word correspondences from input texts by combining length-based pre-alignment of sentences with a stochastic method (IBM Translation Model 1), the correspondences are used subsequently to improve the initial pre-alignment. In the available implementation **Mre** proposes 1:1 links only.[8]

The results are measured in recall, precision and F-measure, computed for the purpose of alignment evaluation in the usual way as in Fig. 1.[9] Overall, the best results are achieved by **Mre** in precision (92.9) and **GC** in recall (84.3). On noisy texts, **Mre** compares with **GC** even better in precision (92.2 and 91.5, compared to 84.1 and 84.9). For 'clean' texts, precision of **GC** is better (98.7 vs. 95.1). **Mmd** scores worst, possibly due to inadequate tuning to the language pair, while **Brn** is marginally worse than **GC**.[10] On the other hand, **Mre** shows marked improvements the more input it gets. With 10,000 sentences it wins on both clean and noisy texts in precision (100 and 98.4) and on noisy texts in recall (89.2). Rather surprisingly, it fails on an easy corpus sample with short sentences (precision 66.8), as opposed to more difficult samples (100 and 99.5).[11] The lessons learnt from the previous evaluations can be summarised as follows:

1. Quality of alignment depends to a large extent on properties of the input: on its formatting complexity – the presence of elements other than running text (graphics, tables, notes), on "structural distance" between the original and its translation (a scale from literal to free translation), on the amount of "noise" (such as omissions or segmentation differences/errors due to preprocessing), on typological distance between the two languages (important

[5] Probably the most popular alignment tool, dubbed *vanilla aligner*. For an implementation see `http://nl.ijs.si/telri/Vanilla/`.

[6] In fact, a "hard region" can be larger than one paragraph. With some loss in speed, it could be a chapter or even a book. Similarly, a "soft region" can be larger than a sentence – this way paragraphs may be aligned instead of sentences.

[7] `http://nlp.cs.nyu.edu/GMA/`

[8] `http://research.microsoft.com/research/downloads/default.aspx`

[9] *correct links* = number of correct links among those proposed by the aligner, *reference links* = number of links in correctly aligned texts (the gold standard), *test links* = number of all links proposed by the aligner. *F-measure* combines recall and precision into a single measure. For a discussion of these measures in the context of alignment see, e.g., Véronis & Langlais (2000) and Melamed et al. (2003).

[10] **Mmd** with appropriate tuning and a Czech-English lexicon was successfully used before on a large set of English-Czech data, see `http://ufal.mff.cuni.cz/pdt/Corpora/Czech-English/`.

[11] In the readme file that comes with **Mre** code a minimum of 10,000 sentence pairs is recommended for reliable estimation of a statistical word-translation model.

$$recall = \frac{correct\ links}{reference\ links}$$

$$precision = \frac{correct\ links}{test\ links}$$

$$F\text{-}measure = 2\,\frac{recall\ \times\ precision}{recall\ +\ precision}$$

Figure 1. Measures for evaluating alignment

especially for methods based on searching for *cognates* in the two texts), and – at least for some alignment methods – on the input size.

2. Alignment methods differ in their sensitivity to such properties.[12] Some methods can be trained or supplied with additional resources to handle difficult texts in a specific language pair, but it requires additional effort and/or availability of such resources. It seems that there is no single best all-purpose way to sentence alignment.
3. As can be expected, word-correspondence methods fare better on noisy texts, but even standard sentence-length-based methods turn out to yield satisfactory results.
4. Counting the number of correctly aligned sentence pairs as the evaluation result is not always a fair measure: sentence boundaries may not have been detected correctly (often there is no unanimous way to segment a text into sentences anyway), and a sentence pair where one sentence is a partial translation of the other should not be treated on par with a totally unrelated pair. Thus, alignments of sentences in *ARCADE* were measured also in terms of words and characters. However, for the practical purpose of building a parallel corpus, the "strict" measure in terms of alignment links seems to be sufficient, or even preferable.
5. When correct alignment (gold standard) is available, both precision and recall can be obtained: selecting a method maximising precision may be the right move for some tasks, while the opposite may be needed for other tasks.

To answer our original question concerning an optimal choice of (a mix of) tools and procedures that would be best suited to a specific text type and language pair, with minimum manual checking and the goal of a near-to-perfect result, the inevitable conclusion would be that with various text types and diverse languages there is probably no universal solution. Instead, a new choice must be made each type a significantly new input occurs, based on experience and experimentation.

[12] S&H make this a key point of their report.

4 Comparison

Although we could not compete with the previous evaluation projects on the level of methodology and systematic exploration of text versions, we decided to conduct a smaller scale evaluation of our own. We were interested in trying out candidate tools on our data, including Czech and at least two other languages. We used three aligners (**GC**, **Mmd** and **Mre**) from the set of four introduced in the previous section, some with additional resources or in a slightly modified version:

Mmd$^+$ – Same as **Mmd**, with a 106K-entries English-Czech lexicon.[13]

Mre* – Same as **Mre**, with some words in the input truncated by a character or two.[14]

Mre$^+$ – Same as **Mre**, with more input data (the previously mentioned English-Czech lexicon and an English-Czech pre-aligned corpus of 830K/731K words[15]).

The systems were tested on a rather opportunistic set of text samples for which hand-corrected alignment was available.[16] Nevertheless, the set at least partially reflects the needs of the project: the samples consist mostly of fiction, two language pairs are represented, and one of the sample includes substantial noise.

AC – This is the sample with the highest noise. It consists of 46 documents (in each language) from the English-Czech part of Acquis Communautaire[17] (roughly 1% of the total number, eliminating those that did not contain usable data). All omissions and mismatches in segmentation were retained. As in the full corpus, the segments aligned are paragraphs rather than sentences, which, however, does not make too much difference as most paragraphs in these legal texts consist of a single sentence.

1984 – George Orwell's novel in English and Czech. This is the most orderly sample, with just a few omissions in the Czech part.[18]

FR7 – Seven French fiction/essay books with Czech translations.[19] The sample does not include any information about paragraph boundaries.

---

[13] The lexicon we used is a GNU/FDL project, available from `http://slovnik.zcu.cz/`.

[14] This was actually due to the fact that the **Mre** perl scripts as downloaded from the Microsoft pages ignored the Czech locale setting. We are grateful to Bob Moore, the author of the program, and Pavel Pecina for their kind assistance in solving this issue. The reason the faulty version is still mentioned is that with less input data it actually produced better results than the corrected version.

[15] `http://ufal.mff.cuni.cz/pdt/Corpora/Czech-English/`

[16] Except for one sample (AC) that was checked and corrected by the author.

[17] See Sect. 2 for details.

[18] This sample was produced and hand-corrected within the project *Multext-East*, see `http://nl.ijs.si/ME/`.

[19] For this hand-corrected sample I owe thanks to Martin Svášek.

Quantitative data on the samples, including hand-corrected alignment counts, are given in Table 1. The percentage of 1:1 links provides a rough measure of the difficulty of the sample – the more such links, the easier the sample.

Table 1. Size of the samples

Text	Cz words	L2 words	Cz segments	L2 segments	All links	1:1 links
AC	62,010	74,986	3,025	2,699	2,685	89%
1984	99,099	121,661	6,756	6,741	6,657	97%
FR7	289,003	337,226	21,936	21,746	21,207	95%

Table 2 gives counts of all types of links (n:n) for all samples and aligners.[20] The counts are compared in terms of recall, precision and F-measure.

Table 2. All links

	Reference	Test	Correct	Recall	Precision	F-measure
AC						
GC	2700	2683	2225	82.4	82.9	82.7
Mmd$^+$	2700	2686	2492	**92.3**	92.8	**92.5**
Mre	2700	2313	2218	82.1	95.9	88.5
Mre$^+$	2700	2375	2308	85.5	**97.2**	91.0
1984						
GC	6657	6633	6446	**96.8**	97.2	97.0
Mmd$^+$	6657	6606	6287	94.4	95.2	94.8
Mre	6657	6167	6110	91.8	99.1	95.3
Mre*	6657	6370	6320	94.9	99.2	97.0
Mre$^+$	6657	6441	6402	96.2	**99.4**	**97.8**
F7						
GC	21207	20868	19427	**91.6**	93.1	**92.3**
Mre	21207	19512	18801	88.7	**96.4**	**92.3**
Mmd	21207	21057	16161	76.2	76.7	76.4

As expected, the two aligners using lexical anchors perform significantly better on noisy texts (AC) than the length-based aligner **GC**, the difference reaching 10 and more percentage points in all measures. Interestingly, on AC, **Mre**$^+$ is better than **GC** even in recall, although it outputs 1:1 links only. On the other hand, **GC** has better recall on the more orderly texts 1984 and F7, but it still lags

[20] Originally, a part of F7 (one of the novels, about one seventh of the total F7 size) was used for testing **Mmd** only. Surprisingly, the results were comparable to those obtained for **Mmd** on English-Czech samples, where additional resources were available. The unconfirmed explanation may be that this specific novel was very easy to align.

behind **Mre**$^+$ in precision. Actually, the relatively good performance of **GC** on F7 is surprising, given that the system expects "hard regions" to be paragraphs, rather than whole books, as was the case here. On F7, **Mmd** clearly suffers from the lack of resources and tuning.[21] The aggregate F-measure distributes its favour rather fairly among all aligners, still pointing twice to **Mre**/**Mre**$^+$. Tables 3, 4, and 5 rank the aligners by recall, precision, and F-measure and precision, respectively.

Table 3. Ranking for recall (all links)

Rank	AC	1984	F7
1.	92.3 Mmd$^+$	96.8 GC	91.6 GC
2.	85.5 Mre$^+$	96.2 Mre$^+$	88.7 Mre
3.	82.4 GC	94.9 Mre*	76.2 Mmd
4.	82.1 Mre	94.4 Mmd$^+$	
5.		91.8 Mre	

Table 4. Ranking for precision (all links)

Rank	AC	1984	F7
1.	97.2 Mre$^+$	99.4 Mre$^+$	96.4 Mre
2.	95.9 Mre	99.2 Mre*	93.1 GC
3.	92.8 Mmd$^+$	99.1 Mre	76.7 Mmd
4.	82.9 GC	97.2 GC	
5.		95.2 Mmd$^+$	

Table 5. Ranking for F-measure (all links)

Rank	AC	1984	F7
1.	92.5 Mmd$^+$	97.8 Mre$^+$	92.3 GC
2.	91.0 Mre$^+$	97.0 GC	92.3 Mre
3.	88.5 Mre	97.0 Mre$^+$	76.4 Mmd
4.	82.7 GC	95.3 Mmd$^+$	
5.		95.3 Mre	

To enable fair comparison with **Mre** and the data in S&H, Table 6 gives corresponding results on 1:1 links. As can be expected, the results are better

[21] Although it did surprisingly well on F1, an easy subset of F7: with 96.7/97.0/96.8 for recall/precision/F-measure it is the winner in the French-Czech category.

than for n:n links in all cells, except for **Mre**'s precision, where they are necessarily identical (the system outputs 1:1 links only). Again, there is no outright winner: **Mre** scores best in recall everywhere and **Mmd** in precision wherever additional resources were available (AC and 1984), while **GC** is marginally better in precision on F7. Taking into account variations in the amount of noise, structural differences, different language pairs and availability of additional resources, the results fall within the range of those reported by S&H.

Table 6. Links 1:1 only

	Reference	Test	Correct	Recall	Precision	F-measure
AC						
GC	2391	2248	2156	90.2	95.9	93.0
Mmd$^+$	2391	2354	2304	96.4	**97.9**	**97.1**
Mre	2391	2313	2218	92.8	95.9	94.3
Mre$^+$	2391	2375	2308	**96.5**	97.2	96.9
1984						
GC	6440	6438	6274	97.4	97.5	97.4
Mmd$^+$	6404	6301	6287	97.6	**99.8**	98.7
Mre	6440	6167	6110	94.9	99.1	96.9
Mre*	6440	6370	6320	98.1	99.2	98.7
Mre$^+$	6440	6441	6402	**99.4**	99.4	**99.4**
F7						
GC	20116	19220	19427	92.6	**96.9**	94.7
Mre	20116	19512	18801	**93.5**	96.4	**94.9**
Mmd	20116	19714	15539	77.2	78.8	78.0

Considering the overall results, conclusions of the previous evaluations seem to be largely confirmed. On noisy texts, **Mmd** and **Mre** fare better than **GC**, while on clean texts, **Mre** and **Mmd** tend to show higher precision than **GC**. Surprisingly, **GC** performs well on F7 without paragraph boundaries (with book as the hard region) and **Mmd** on an easy subset of F7 without bilingual lexicon. Further improvements might be achieved with the two lexically-based methods: **Mre** can be expected to gain further points with more input data and – possibly – lemmatisation, while **Mmd** may profit from creating more cognates by more tuning and better additional resources.

Overall, the results also confirm the conclusion that there is no single best alignment tool for all purposes, and that the success is to a large extent determined by choosing the right tool for a given text. Additionally, the choice might depend on how the automatically aligned texts will be used, and here the tradeoff between recall and precision comes into play.

For some applications, such as machine learning, maximising precision is probably the best strategy if manual checking is not an option. On the other hand, S&H claim that if the result is going to be manually checked before use, it

is desirable to maximise recall: some decrease in precision is not going to make manual checking much more difficult.

This reasoning assumes that all links are going to be checked. On the other hand, if safe links can be identified in the result and only the rest is presented for manual checking, the amount of human effort could be substantially reduced. In this scenario, 100% precision is needed to obtain error-free alignment, but we might be satisfied even with a figure close to it. Recall is of secondary interest.

With precision close to 100%, the "unsafe" links are simply those that the aligner does not propose, they do not even exist as links yet. An alternative, less reliable method of automatic alignment can then be used to suggest links in this more difficult portion of the input.

In the following section, we explore an option to raise precision to make a scenario combining automatic alignment with manual checking more attractive.

5 Joining forces

In order to push precision closer to 100%, a single text pair can be processed by more than one aligner and a correct link defined as one on which all (or most) aligners agree. The set of proposed links would be smaller, but they would be safer: a decrease in recall, an increase in precision.

The results of the three aligners as solo performers, presented in the previous section, were intersected pairwise and all together. For convenience, the top lines of the two tables (7 and 8) give the counts already presented for solo aligners. Only two samples were used (1984 and F7), and **Mmd** – due to its poor performance – was excluded from the test on F7.

Table 7. Merging results on 1984

	Ref.	Test	Correct	Recall	Precision	F-measure
GC	6657	6633	6446	**96.83**	97.18	97.01
Mmd^+	6657	6606	6287	94.44	95.17	94.81
Mre^+	6657	6441	6402	96.17	99.39	**97.76**
GC/Mmd^+	6657	6279	6254	93.95	99.60	96.69
GC/Mre^+	6657	6354	6348	95.36	**99.91**	97.58
Mmd^+/Mre^+	6657	6130	6114	91.84	99.74	95.63
$GC/Mmd^+/Mre^+$	6657	6095	6089	91.47	99.90	95.50

Both samples show the same pattern: F-measure is always better for an aligner in solo mode (**Mre^+** and **Mre**), but a tandem of aligners always wins in precision, reaching 99.91 for **GC/Mre^+** on 1984, with recall still at 95.36. This is an improvement of about 2.7/0.5 percentage points over their solo performance in precision. The gain is even more marked for F7: 3.6 points.

Table 8. Merging results on F7

	Reference	Test	Correct	Recall	Precision	F-measure
GC	21207	20868	19427	**91.61**	93.09	92.34
Mre	21207	19512	18801	88.65	96.36	**92.35**
Mmd	21207	21057	16161	76.21	76.68	76.44
GC/Mre	21207	17728	17661	83.28	**99.62**	90.72

6 Conclusions and future work

1. Several conclusions of previous evaluations have been confirmed: quality of alignment depends to a large extent on properties of the input and alignment methods differ in their sensitivity to such properties. Thus, word-correspondence methods fare better on noisy texts, where sentence-length-based methods give mixed results.
2. Although none of the evaluated aligners was the overall winner, it was **Mre**, especially when supplied with additional resources, that often performed better than its contestants. Again, this is in accordance with a previous evaluation (Singh & Husain 2005). Still, the success is to a large extent determined by choosing the right tool for a given text.
3. Manual checking of alignment results can be done more efficiently with an automatic alignment method preferring higher precision to better recall. With precision close to 100, manual checking can focus only on links where good results are less likely. Such links are not even proposed by the aligner, although a different, less reliable aligner can be used in a step preceding manual checking of the difficult parts of the input.
4. In order to raise precision, sets of links proposed by different aligners can be intersected. Our results show that such a move improves precision by 0.5–3.6 percentage points.

The tests should be extended to more languages, text types and tools,[22] and they would profit from a more rigorous methodology. But the present results already suggest that a near-to-perfect sentential alignment with a small amount of manual checking is a realistic perspective.

[22] *HunAlign*, a tool developed within the *Hunglish* English-Hungarian parallel corpus project, is a hot candidate, see `http://mokk.bme.hu/resources/hunalign`.

Bibliography

Barlow, M. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics*, **4**(1), 319–327.

Barlow, M. (2002). ParaConc: Concordance software for multilingual parallel corpora. In *Language Resources for Translation Work and Research, LREC 2002*, pages 20–24.

Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Meeting of the Association for Computational Linguistics*, pages 169–176.

Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger, R. (2005). Massive multilingual corpus compilation; Acquis Communautaire and totale. In *2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (L&T'05)*, Poznań, Poland. Available at http://www.jrc.cec.eu.int/langtech/.

Gale, W. & Church, K. (1991a). Identifying word correspondance in parallel text. In *Proceedings of the DARPA NLP Workshop.*

Gale, W. A. & Church, K. W. (1991b). A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.

Langlais, P., Simard, M., & Véronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 711–717. Association for Computational Linguistics.

Melamed, I. D. (1997). A portable algorithm for mapping bitext correspondence. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 305–312, Somerset, New Jersey. Association for Computational Linguistics.

Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and Recall of Machine Translation. In *HLT-NAACL*.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK. Springer-Verlag.

Singh, A. K. & Husain, S. (2005). Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.

Tiedemann, J. & Nygaard, L. (2004). The OPUS corpus – parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evalu-*

ation (LREC'04), Lisbon, Portugal.

Véronis, J. & Langlais, P. (2000). Evaluation of parallel text alignment systems: the arcade project. In J. Véronis, editor, *Parallel text processing: Alignment and use of translation corpora*, pages 369–388. Kluwer Academic Publishers, Dordrecht.