Linear Regression Based Alignment of Parallel Texts Using Homograph Words

António Ribeiro[1], **Gabriel Pereira Lopes**[2] and **João Tiago Mexia**[3]

Abstract. About 15% of the vocabulary found in large texts of the Official Journal of the European Communities is the *same* in its various official languages. If we take, for example, the Portuguese–Spanish pair, the rate rises to more than 30% since these are similar languages and, for the opposite reason, it drops to about 10% for the pair Portuguese–German. This is a wealthy source of information for parallel texts alignment that should not be left unused.

Bearing this in mind, this paper describes a language independent method that makes use of those words, which are homograph for a pair of languages, in order to align parallel texts. This work was originally inspired and extends work done by Pascale Fung & Kathleen McKeown, and Melamed. In order to filter out words that may cause misalignment, we use confidence bands of linear regression analysis instead of statistically unsupported heuristics. We do not get 100% text alignment precision mostly due to term order policies in the different languages. The parallel segments obtained have an average length of four words for case law texts.

1 INTRODUCTION

If we are aiming at building bilingual databases of equivalent expressions (typical translations) either for cross-language information retrieval (for web applications, for example), machine translation or bilingual lexicography, we should be able to make this an automatic language independent task. We can no longer afford to waste human time and effort building manually these ever changing databases or design language specific applications to solve this problem. It becomes quite clear in the European Community context where, at this moment, eleven official languages are already in use let alone the ones to come as new member states arrive. Thousands of pages are translated daily into the eleven languages.

Parallel texts (texts that are mutual translations) are valuable sources of information for these information extraction tasks as they provide the typical usage of equivalent expressions. However, they are not of much use unless a computational system may find which piece of text in one language corresponds to which piece in the other language. In order to achieve this goal, they must be *aligned* first, i.e. the various text pieces must be put into correspondence. This is usually done by finding *correspondence points* – sequences of characters with the same form (*homograph*, e.g numbers, names, punctuation marks) or even known translations.

Term translations have been used as correspondence points in [5] for alignment of English–Chinese. Orthographic cognates (see [14]) were also added in [8]. However, the problem is that both approaches use statistically unsupported heuristics to filter candidate correspondence points.

A method to filter candidate correspondence points using confidence bands of linear regression lines is proposed in [12]. The points of this line were generated from homograph words which occur with the same frequency in parallel texts. This work extends previous work reported in [11] where only hapaxes were used as candidate correspondence points. Both approaches avoid heuristic filters and the authors claim 100% alignment precision but the linear regression analysis provides a small number of reliable correspondence points. For the first approach they report an average of about 100 points, leading to segments ranging from 70 words to 12 pages for large texts.

In this paper, we will extend the work in [12] by defining a *recursive* algorithm for alignment of parallel texts. We will also use linear regression lines built from candidate correspondence points generated from homograph words which occur with the same frequency *within parallel text segments*. So, we define the initial parallel text segment based on the lengths of the original parallel texts, find the reliable correspondence points using confidence bands (see [9]) and apply this very same algorithm recursively to each sub-segment. In the end, we are able to get over 100 times more correspondence points on average (a quadratic increase over [12]) and alignment precisions close to 100%.

The following section will briefly discuss some related work. In section 3, we will describe the corpus used, outline the method and show some results. Section 4 evaluates, compares them and shows some of the persisting misalignment problems. Finally, we present the conclusions and future work in the last sections.

2 BACKGROUND

There have been two mainstream approaches to parallel text alignment. One assumes that translations have proportional sizes; the other tries to use lexical information in the parallel texts to

[1] Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Departamento de Informática, Quinta da Torre, P-2825-114 Monte da Caparica, Portugal, email: ambar@di.fct.unl.pt

[2] Address as above, email: gpl@di.fct.unl.pt

[3] Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Departamento de Matemática, Quinta da Torre, P-2825-114 Monte da Caparica, Portugal

generate candidate correspondence points. All in all, both use some notion of correspondence points.

In early work, parallel texts were aligned using sentences which had a proportional number of words and characters (see [1] and [6]). However, these algorithms tended to break down since they needed clearly marked sentence boundaries. But [2] showed that text alignment was still possible by exploiting orthographic cognates (see [14]). In order to avoid noisy points, an *empirically* estimated search space was used to filter them out. In [7] two sentences were aligned if the number of correspondence points associating them was greater than an empirically defined threshold. Those correspondence points were generated from words with *similar distributions*, i.e. if they occurred in the same sentences. In [3] noisy points were filtered out by deleting frequent words.

The requirement for clear sentence boundaries was dropped in [4] on a case-study for English–Chinese. They used vectors that stored distances between consecutive occurrences of a word (DK–vec's) and candidate correspondence points were identified from words with *similar* distance vectors. Noisy points were filtered using some heuristics. In [15] the points were generated from isolated cognates, i.e. words that are not mistaken for others within a text window. Those outside an *empirically* defined search space are filtered. Finally, [8] also uses some empirically defined heuristics to filter candidate correspondence points generated from orthographic cognates.

We all want to find reliable correspondence points for parallel texts alignment. They provide the basic means for extracting reliable information from parallel texts. However, as far as we learned from the above papers, current methods have repeatedly used statistically unsupported heuristics in order to filter out noisy candidate correspondence points. For instance, all mention the "golden translation diagonal" to filter out noisy points. This is the diagonal of a rectangle whose sides are proportional to the lengths of parallel texts. It follows the hypothesis that parallel texts have proportional lengths.

3 FILTERING NOISY CORRESPONDENCE POINTS

3.1 Overview

The basic insight behind our approach is that not all candidate correspondence points are reliable. No matter how we filter correspondence points, either using similar word distributions (see [5] and [7]), search corridors [15], point dispersion [8], angle deviation [8] or some other heuristic, candidate correspondence points must be filtered in order to ensure correct text alignment. Our assumption is that reliable points have similar characteristics. For instance, they tend to gather somewhere near the "golden diagonal". As in [12], we also assume that homograph words with equal frequencies in parallel text segments may offer good points.

3.2 Source Parallel Texts

We worked with a mixed parallel corpus consisting of texts selected at random from the Official Journal of the European Communities [10] and from The Court of Justice of the European Communities[4].

Table 1. Number of words per sub-corpus (average number of words per text appears inside brackets; markups were discarded).[5]

	Sub-corpus			
Language	Written Questions	Debates	Judgements	Total
da	259k (52k)	2,0M (395k)	16k (3k)	2250k
de	234k (47k)	1,8M (368k)	15k (3k)	2088k
el	272k (54k)	1,9M (387k)	16k (3k)	2222k
en	263k (53k)	2,1M (417k)	16k (3k)	2364k
es	292k (58k)	2,2M (439k)	18k (4k)	2507k
fi	---	---	13k (3k)	13k
fr	310k (62k)	2,2M (447k)	19k (4k)	2564k
it	279k (56k)	1,9M (375k)	17k (3k)	2171k
nl	275k (55k)	2,1M (428k)	16k (3k)	2431k
pt	284k (57k)	2,1M (416k)	17k (3k)	2381k
sv	---	---	15k (3k)	15k
Total	2468k (55k)	18,4M (408k)	177k (3k)	21005k

For each language, we included:

- five texts with Written Questions asked by members of the European Parliament to the European Commission with the corresponding answers (average: about 60k words or 100 pages / text);
- five texts with records of Debates (transcripts of spoken discussions) in the European Parliament (average: about 400k words or more than 600 pages / text);
- five texts with Judgements of The Court of Justice of the European Communities (average: about 3k words or 5 pages / text).

In order to reduce the number of possible language pairs from 110 (11 languages×10) to a more manageable size, we decided to take Portuguese as the kernel language of all pairs (10 pairs).

3.3 Generating Candidate Correspondence Points

Homograph words in parallel texts provide good clues for parallel texts alignment. As a naive and particular form of cognate words, they are likely translations (e.g. *Paris* in various European languages). These words end up being mainly numbers and names. Here are a few examples from a parallel Portuguese–German text: *2002* (numbers, dates), *GATT* (acronyms), *Gérard* (proper names), *Portugal* (names of countries), *Dresden* (names of cities), *flow* (foreign words, as in *cash flow*), *ad-hoc* (latin words), *global* (common vocabulary – homograph word).

If we compare the amount of common vocabulary in the selected pairs of parallel texts (see Table 2), we get an average of 10% for pairs with Germanic languages. This number depends on language similarity. For instance, it rises to more than 30% for the Portuguese–Spanish pair.

[4] http://curia.eu.int. The texts are in all official languages of the European Union: Danish (da), Dutch (nl), English (en), Finnish (fi), French (fr), German (de), Greek (el), Italian (it), Portuguese (pt), Spanish (es) and Swedish (sv).

[5] No 'Written Questions' and 'Debates' texts are available for Finnish and Swedish since the respective countries were not still part of the European Union in 1992–4.

Table 2. Average percentages of common vocabulary (homograph words) per pair of parallel texts.

Pair	Sub-corpus			Average
	Written Questions	Debates	Judgements	
pt-es	38%	32%	36%	34%
pt-fi	---	---	19%	19%
pt-sv	---	---	19%	19%
pt-en	19%	10%	20%	13%
pt-fr	19%	11%	22%	13%
pt-it	22%	8%	25%	13%
pt-da	17%	9%	19%	12%
pt-de	15%	9%	19%	11%
pt-el	15%	7%	18%	10%
pt-nl	17%	5%	19%	9%
Average	20%	13%	22%	15%

Furthermore, the number of occurrences of these common vocabulary words in the parallel texts (see Table 3) reaches an average of almost 50% in parallel Portuguese–Spanish texts. For Portuguese–German, this number is about 25% on average. So, why not make use of this *treasure*?

Table 3. Average number of common vocabulary words per pair of parallel texts (average percentage in brackets).

Pair	Sub-corpus			Average
	Written Questions	Debates	Judgements	
pt-da	1,2k (32%)	1,9k (20%)	156 (33%)	1,7k (24%)
pt-de	1,0k (27%)	1,9k (19%)	154 (31%)	1,6k (22%)
pt-el	1,0k (29%)	1,5k (16%)	146 (31%)	1,3k (20%)
pt-en	1,3k (31%)	2,1k (19%)	161 (30%)	1,8k (23%)
pt-es	2,5k (52%)	6,5k (42%)	294 (52%)	5,2k (45%)
pt-fi	---	---	152 (30%)	0,2k (30%)
pt-fr	1,3k (40%)	2,2k (27%)	175 (41%)	1,9k (31%)
pt-it	1,4k (35%)	1,7k (14%)	199 (38%)	1,6k (21%)
pt-nl	1,2k (35%)	1,1k (12%)	149 (35%)	1,1k (19%)
pt-sv	---	---	152 (29%)	0,2k (29%)
Average	1,4k (35%)	2,6k (24%)	174 (35%)	2,2k (27%)

In order to avoid pairing words that are not equivalent though homograph (e.g. *a*, a definite article in Portuguese and an indefinite article in English), we restricted ourselves to using homograph words which occur with the same frequency in both parallel texts segments. In this way, it becomes more likely that they are equivalent. On average, the percentages of these words range from 1% (4k words) for the large texts up to 10% for the small texts (300 words).

For example, for the Written Questions sub-corpus, these words account for about 6% of the total number of words (about 3k words / text). In this way, each pair of texts gives a set of candidate correspondence points which are used to draw a linear regression line. These points are defined using the co-ordinates of the word positions in each of the parallel texts. For example, if the first occurrence of the homograph word *global* occurs at position 52652 in the Portuguese text and at word position 47670 in the German parallel text, then the point co-ordinates are (52652, 47670). Points may adjust themselves well to a linear regression line or may be dispersed around it. In order to filter out extreme points, we apply first a filter based on the histogram of the distances between the expected and real positions. Next, we remove other noisy points using a finer-grained filter based on the confidence bands of the linear regression line. We will elaborate on these statistical filters in the next subsections.

3.4 Eliminating Extreme Points

Points obtained in the first stage from positions of homograph words with equal frequencies are still prone to be noisy.

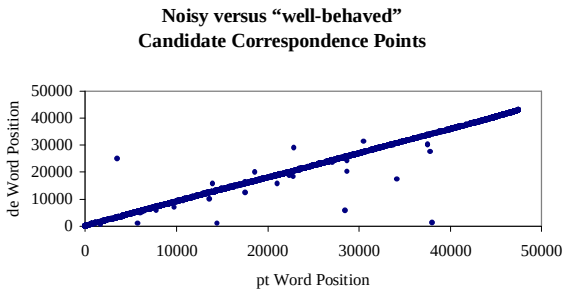

Figure 1. "Well-behaved" points are "in line".

The figure above shows noisy points caused by homograph words whose pairs appear in distant positions. We should feel reluctant to accept these pairings and that is what the first filter does. It filters out those noisy points which are clearly far apart from their *expected positions*. Expected positions are computed from the linear regression line on all points.

Table 4. A sample of the table of distances between the expected and the real positions of some noisy points in Figure 1.

Word	Positions			
	pt	de	de Expected	Distance
The	3546	24885	3546	21681
M	28523	5637	28523	19917
Force	38073	1150	38073	32949

An histogram (Figure 2) of the distances between real and expected word positions helps us to identify those words pairs which are too distant from their expected positions. The noisy pairs are filtered out and we proceed to a finer-grained filter.

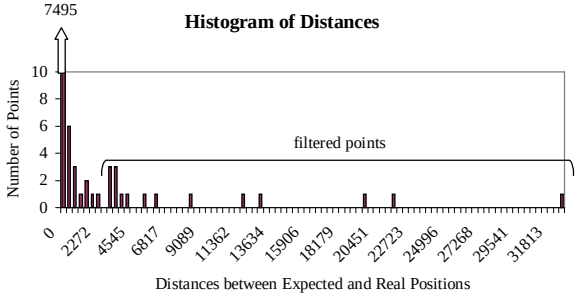

Figure 2. Histogram of the distances between expected and real word positions. For most pairings, the distance is below 400 words.

Since our approach is based on the expected and real word positions, we are even able to identify wrong pairings of homograph words which are not equivalent. Since they are *false friends*, they tend to appear in different places in the parallel texts for they have different meanings as the following example with the homograph German–Portuguese word 'Mais' shows. The German word *Mais* means 'milho' in Portuguese ('corn' in English). Conversely, the Portuguese word *Mais* means 'mehr' in German ('more' in English). The following figure shows parallel text segments with these words in distant word positions. Should those words be paired, they form extreme points.

```
54533|Einfuhr von Mais und Sorghum
58464|importação de milho e do sorgo

193273|Mehr als eine Notwendigkeit
204137|Mais do que uma necessidade
```

Figure 3. Parallel text segments with the homograph word *Mais* which has different meanings in German and Portuguese.

3.5 Linear Regression Line Confidence Bands

Linear regression lines define confidence bands which help us to identify reliable points, i.e. points which belong to that linear regression line with a great confidence level (95%). The band is wider in the extremes of the linear regression line and narrower in the middle, where the confidence is lower.

We start from the linear regression line defined by the points filtered using the histogram technique described in the previous section. We compute the confidence bands of the linear regression line (see [9] for details) to filter out points lying outside, since they are credited as too unreliable for alignment. Then, for each sub-segment defined by the remaining "well-behaved" correspondence points, we recursively re-apply the alignment algorithm. In this way, we are able to do a local identification of candidate correspondence points and to filter noisy points.

Here is a summary of the recursive alignment algorithm:

1. Take two parallel texts A and B;
2. Define the texts' beginnings – the point (0,0) – and the texts' ends – the point (length of text A, length of text B) – as the extremes of the initial parallel text segment;
3. Consider as candidate correspondence points those points defined by homograph words which occur with the same frequency within the parallel text segment;
4. Filter out extreme points using the Histogram technique;
5. Filter out points which lie outside the confidence bands of the linear regression line;
6. For each sub-segment defined by two consecutive points, repeat steps 3 to 6.

4 EVALUATION

We ran the previous algorithms on the parallel texts of 10 language pairs as described in section 3.2. With the current implementation, which is not streamlined, and on a Pentium II 366 MHz with 64MB, the algorithm takes about seven hours to align each pair of "Debates" parallel texts (400k words) and less than one minute for the "Judgements" (3k words).

We compared the results with the ones reported in [12] and found a quadratic increase in the number of correspondence points, on average (Table 5).

Table 5. Average number of final correspondence points (gain inside brackets compared to the results in [12])

	Sub-corpus			
Pair	Written Questions	Debates	Judgements	Average
pt-da	14k (101×)	20k (374×)	1k (11×)	18k (195×)
pt-de	14k (123×)	22k (225×)	1k (14×)	19k (207×)
pt-el	14k (106×)	16k (139×)	1k (15×)	15k (146×)
pt-en	14k (167×)	23k (231×)	1k (20×)	20k (259×)
pt-es	22k (365×)	23k (421×)	2k (15×)	22k (290×)
pt-fi	---	---	1k (14×)	1k (14×)
pt-fr	17k (104×)	30k (270×)	1k (7×)	26k (162×)
pt-it	16k (133×)	21k (209×)	1k (41×)	19k (230×)
pt-nl	14k (116×)	12k (155×)	1k (21×)	12k (148×)
pt-sv	---	---	1k (16×)	1k (16×)
Average	16k (134×)	24k (265×)	1k (14×)	21k (223×)

The gain is especially significant in the large texts where we got more than 260 times more points, corresponding to an increase from 90 points [12] to an average of 24k points for our recursive algorithm. It uses not only the homograph words which have equal frequencies in the initial segment, but also within each parallel sub-segment. One word may not have the same frequency in the initial parallel text segment, but may turn out to have the same within some parallel sub-segments.

Table 6 shows that about a third of the homograph words in parallel texts are used for alignment.

Table 6. Ratio of the number of Correspondence Points and the number of homograph words.

	Sub-corpus			
Pair	Written Questions	Debates	Judgements	Average
pt-da	76%	26%	62%	31%
pt-de	91%	29%	65%	34%
pt-el	85%	26%	70%	31%
pt-en	81%	31%	70%	35%
pt-es	73%	14%	70%	18%
pt-fi	---	---	55%	55%
pt-fr	73%	29%	71%	32%
pt-it	79%	38%	69%	44%
pt-nl	72%	26%	64%	33%
pt-sv	---	---	66%	66%
Average	78%	25%	67%	30%

On average, we are able to break a text into segments of four up to 20 words. It should be noted, however, that there are still some misalignment problems. We consider a point misaligned when its corresponding words are not within the same segment. Misalignments occur specially when there are large insertions of non-translated text and in the case of term order inversions. This is the reason why the alignment precision does not reach 100% for all parallel texts. The figure below gives a quite clear example for Portuguese–German:

```
328434|? ||    310432|?
    328435|¶ ¶ Ainda uma pergunta sobre a
avaliação da capacidade de produção . A [...] ¶
¶ vamos [A]apoiar as [B]propostas do [C]relator
, tal como vêm [D]expostas no [E]relatório
    310433|¶ ¶ ¶ Eine Frage zur
Kapazitätsberechnung möchte ich noch [...]
werden wir die [B]Vorschläge des
[C]Berichterstatters , wie sie im
328568|Donnelly||    310543|Donnelly
    328569|
    310544|- [E]Bericht des
[F]Wirtschaftsausschusses [D]niedergelegt sind
328569|, ||    310550|,
    328570|da [F]Comissão dos Assuntos
Económicos . ¶ ¶ [G]PRESIDÊNCIA
    310551|[A]unterstützen . ¶ ¶ ¶ [G]VORSITZ
328579|: ||    310557|:
```

Figure 4. Segments alignment (bold lines show points co-ordinates; letters inside square brackets indicate translation equivalents).

Although the words in bold are correctly paired, the segments are misaligned (see the figure below).

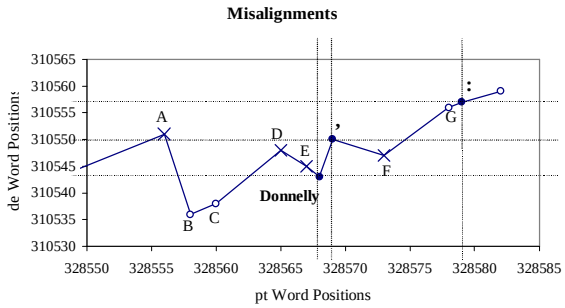

Figure 5. Misaligned segments (× – misaligned points; ● – correspondence points; ○ – translation equivalents).

Points D and E are misaligned because they are in the segment preceding *Donnelly* in Portuguese and in the subsequent one in German. Still, points B and C are correctly paired but point A lies *two* segments ahead in the German text. This has much to do with the languages term order policies. In the case of point A, the word *unterstützen* was placed in the end of the sentence, while its Portuguese translation equivalent *apoiar* was placed after the subject. So, when the alignment becomes more fine-grained, the chances of misalignment increase.

All in all, with the current alignments, we are already able to retrieve some correct translation equivalents by selecting short frequent parallel texts segments. Here are a few examples for Portuguese-German:

Table 7. Some short frequent translation equivalents.

ACÓRDÃO DO TRIBUNAL DE JUSTIÇA	URTEIL DES GERICHTSHOFES
CE	EG
de Agosto de	. August
Língua do processo	Verfahrenssprache

5 CONCLUSIONS

In this paper we have presented a statistically backed up algorithm to select correspondence points for parallel texts alignment. It is based on confidence bands of linear regression lines. These lines are built from homograph words which occur with equal frequencies in parallel texts segments. Since the algorithm is recursive, it explores reliable correspondence points within each aligned parallel sub-segment. As the alignment becomes more fine-grained, the 100% precision may be degraded by language specific term order policies in small sub-segments. The method is language and character-set independent. It does not assume any a priori language knowledge, text tagging, well defined sentence or paragraph boundaries nor one-to-one translation of sentences. Moreover, it does not use any stop-list nor removes any words from the text except for mark-ups which might lead to biased results. It can be applied to texts with inserted or deleted parts and it is robust to OCR noise or spelling mistakes. Presently, we are able to extract some translation equivalents using the current alignments. Short frequent parallel text segments often provide them quite clearly.

6 FUTURE WORK

We found several problems with term inversions that cause misalignments. This is leading us to analyse them more carefully in order to improve the alignment precision. In the work reported in this paper, we used only homograph words which occur with equal frequencies in the parallel text segments to generate candidate correspondence points. We are planning to extend this to using words which occur with different frequencies within parallel text segments and equal strings of characters in order to define more candidate correspondence points. A method for extracting meaningful multiword units, string patterns and part of speech tags patterns is described in [13]. This will help us to extract automatically real cognates as candidate correspondence points. Translation equivalents extraction will be starting soon, by using similarity and cohesion measures and taking special care with term inversions.

ACKNOWLEDGEMENTS

This research was partially supported by a grant from Fundação para a Ciência e Tecnologia / Praxis XXI. We would like to thank the anonymous referees for their valuable comments on the paper.

REFERENCES

[1] P. Brown, J. Lai and R. Mercer, 'Aligning Sentences in Parallel Corpora', *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, U.S.A., 169–176, (1991).

[2] K. Church, 'Char_align: A Program for Aligning Parallel Texts at the Character Level', *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, U.S.A., 1–8, (1993).

[3] I. Dagan, K. Church and W. Gale, 'Robust Bilingual Word Alignment for Machine Aided Translation', *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, U.S.A., 1–8, (1993).

[4] P. Fung and K. McKeown, 'Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping', *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, U.S.A., 81–88, (1994).

[5] P. Fung and K. McKeown, 'A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups', *Machine Translation*, **12**, 53–87, (1997).

[6] W. Gale and K. Church, 'A Program for Aligning Sentences in Bilingual Corpora', *Computational Linguistics*, **19**, 75–102, (1993).

[7] M. Kay and M. Röscheisen, 'Text-Translation Alignment', *Computational Linguistics*, **19**, 121–142, (1993).

[8] I. Melamed, 'Bitext Maps and Alignment via Pattern Recognition', *Computational Linguistics*, **25**, 107–130, (1999).

[9] R. Plackett, *Principles of Regression Analysis*, Oxford University Press, Oxford, U.K, 1960.

[10] Office des Publications Officielles des Communautés Européennes, *Multilingual Corpora for Co-operation*, Disk 2 of 2, ELRA, Paris, France, 1997.

[11] A. Ribeiro, G. Lopes and J. Mexia, 'Using Confidence Bands for Alignment with Hapaxes', *Proceedings of the International Conference on Artificial Intelligence (IC'AI 2000)*, CSREA Press, U.S.A., (2000).

[12] A. Ribeiro, G. Lopes and J. Mexia, 'Selecting Homograph Words in Parallel Texts for Alignment', Technical Report, Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Departamento de Informática, Monte da Caparica, Portugal, (2000).

[13] J. da Silva, G. Dias, S. Guilloré and J. Lopes, 'Using Localmaxs algorithms for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units', *Progress in Artificial Intelligence – Lecture Notes in Artificial Intelligence*, P. Barahona and J. Alferes, Springer-Verlag, Berlin, Germany, **1695**, 113–132, (1999).

[14] M. Simard, G. Foster and P. Isabelle, 'Using Cognates to Align Sentences in Bilingual Corpora', *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation TMI-92*, Montreal, Canada, 67–81, (1992).

[15] M. Simard and P. Plamondon, 'Bilingual Sentence Alignment: Balancing Robustness and Accuracy', *Machine Translation*, **13**, 59–80, (1998).