Extracting Portuguese–Spanish Word Translations from Aligned Parallel Texts

António Ribeiro, Gabriel Lopes
Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia – Departamento de Informática
Quinta da Torre, P-2825-114 Monte da Caparica, Portugal
ambar@di.fct.unl.pt, gpl@di.fct.unl.pt

João Mexia
Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia – Departamento de Matemática
Quinta da Torre, P-2825-114 Monte da Caparica, Portugal

Abstract. This paper describes a method for extracting Portuguese–Spanish word translation equivalents from aligned parallel texts. This method uses the standard loglikelihood statistics to measure the similarity between two words. Parallel texts are aligned using a simple method that extends previous work by Pascale Fung & Kathleen McKeown and Melamed. In contrast, the method in this paper does not use statistically unsupported heuristics to filter reliable correspondence points. Instead, it provides the statistical support those authors could not claim by using confidence bands of linear regressions. The points of the linear regression line are generated from the positions of homograph words which occur with the same frequency in parallel text segments. With this alignment method, we are able to extract word translation equivalents (about 90 of the best 100 are correct equivalents).

1 *Introduction*

If we aim at building bilingual databases of equivalent expressions (typical translations) either for cross-language information retrieval (e.g. web applications), machine translation, bilingual lexicography or terminology research, 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. This problem is quite clear in the European Union where eleven official languages are already in use at the moment let alone the ones to come as new member states join in. Everyday, thousands of pages are translated into these languages.

Parallel texts (texts that are mutual translations) are valuable sources of information for bilingual lexicography. However, they are not of much use unless a computational system may find which piece of text in one language corresponds to which piece of text in the other language. In order to achieve this, they must be *aligned* first, i.e. smaller pieces of text must be put into bijective correspondence. This is usually done by finding *correspondence points* – sequences of characters with the same form in both texts (*homographs*, e.g. numbers, proper names, punctuation marks), equivalent forms or even previously known translations.

Pascale Fung & Kathleen McKeown (1997) present an alignment algorithm that uses term translations as correspondence points between English and Chinese texts. Melamed (1999) aligns texts using correspondence points taken either from orthographic cognates (Michel Simard *et al.*, 1992) or from a seed translation lexicon. However, both approaches use statistically unsupported heuristics to filter noisy points. The former approach considers a candidate correspondence point reliable as long as, among some other constraints, "[...] it is not too far away from the diagonal [...]" (Pascale Fung & Kathleen McKeown, 1997, p.72) of a rectangle whose sides sizes are proportional to the lengths of the texts in each language (henceforth, the 'golden translation diagonal'). The latter approach uses various heuristic filtering parameters (Melamed, 1999, pp. 115-116): maximum point ambiguity level (measures how ambiguous a point is for alignment), point dispersion (measures how well the points fit to a linear interpolation) and angle deviation (measures how much the angle formed by a cluster of points deviates from the “golden translation diagonal” angle).

Although all the heuristics found in previous work may be intuitively quite acceptable and may significantly improve the results, they are just heuristics. It is true that parallel texts alignment precision is higher when reliable correspondence points are found. They provide the basic means for extracting information from parallel texts, namely, translation equivalents. Still, as far as we have learned from previous work, current methods have *repeatedly used statistically unsupported heuristics* to filter out noisy points.

António Ribeiro *et al.* (2000a, c) propose a method to align parallel texts based on the occurrence of homograph words which occur with the same frequency in parallel text segments, using the *statistically* defined confidence bands of the “golden translation diagonal”. The method is completely statistics-based using no heuristic filters as in previous work (Pascale Fung & Kathleen McKeown 1997; Michel Simard & Pierre Plamondon 1998; Melamed 1999). We will use this method to align Portuguese–Spanish parallel texts and the likelihood similarity measure (Ted Dunning 1993) to extract bilingual word lexicons[1] based on co-occurrence frequencies (Philippe Langlais & Marc El-Bèze, 1999).

In the following section we will briefly discuss some related work on text alignment[2]. The alignment method is described in section 3 and the translation equivalents extractions is described in section 4. Finally, we evaluate the results in section 5 and present the conclusions and future work.

2 Related Work

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

In early work by Peter Brown *et al.* (1991) and William Gale & Kenneth Church (1991), sentences were aligned counting words and characters, respectively. The algorithms grouped sequences of sentences till they had proportional sizes. However, these algorithms tended to break down when sentence boundaries were not clearly marked. However, Kenneth Church (1993) showed that *cheap* alignment of text segments was still possible exploiting orthographic cognates (Michel Simard *et al.*, 1992). In order to avoid noisy points, an *empirically* estimated search space was used to filter them out. Martin Kay & Martin Röscheisen (1993) aligned two sentences if the number of correspondence points associating them was greater than an *empirically* defined threshold. Dagan *et al.*(1993) generated correspondence points from pairs of translations whose words frequencies were neither *high* nor *low*. Empirically, they found that these words “caused difficulties” and therefore they were filtered out.

Pascale Fung & Kathleen McKeown (1994) also dropped the requirement for clear sentence boundaries on a case-study for English-Chinese texts. 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. In Pascale Fung & Kathleen McKeown (1997), the algorithm used extracted terms, when possible, to compile a list of reliable pairs of translations. Michel Simard & Pierre Plamondon (1998) generated candidate correspondence points from isolated cognates, i.e. words that were not mistaken for others within an empirically found text window. Some were also filtered if they either lied outside a search space, named a corridor, or were “not in line” with their neighbours.

Melamed (1999) also needed to filter candidate correspondence points obtained from orthographic cognates. He used the following heuristics: a maximum point ambiguity level to filter points outside a search space, a maximum point dispersion to filter points too distant from a line formed by candidate correspondence points and a maximum angle deviation to filter points that make this line slope too much.

[1] It is not the purpose of this paper to handle collocations equivalents.

[2] See António Ribeiro *et al.* (2000c) for more details.

Whatever heuristic is taken, either similar word distributions (Pascale Fung & Kathleen McKeown, 1997), search corridors (Michel Simard & Pierre Plamondon, 1998) or point dispersion and angle deviation (Melamed, 1999), the most reliable points must be filtered to ensure the best possible text alignment. Our assumption is that reliable points have similar characteristics. For instance, they tend to gather somewhere near the "golden translation diagonal". Homographs with equal frequencies in parallel text segments have proven to provide good points (António Ribeiro *et al.* 2000a, b, c).

3 *Correspondence Points Filters*

3.1 Source Parallel Texts

For the purpose of translation equivalents extraction, we used five parallel Portuguese–Spanish texts from The Court of Justice of the European Communities[3], amounting to 18k words (an average of about 4k words or 5 pages per text).

3.2 Generating Candidate Correspondence Points

We generate candidate correspondence points from *homographs which occur with the same frequency in parallel text segments*. As a naive and particular form of cognate words, homographs are likely translations (e.g. *Madrid* in various European languages). These words end up being basically numbers and names. Here are a few examples from parallel Portuguese–Spanish texts: *2002* (numbers, dates), *Euratom* (acronyms), *Carlos* (proper names), *Portugal* (names of countries), *Madrid* (names of cities), *p* (abbreviations), *República* (common vocabulary words).

Actually, comparing the vocabularies of Portuguese and Spanish words found in the texts we used, 36% of the vocabulary is the same. Consequently, about 33% of the words found in those texts are the same. Language similarity favours the number of candidate correspondence points for more homographs are found in the parallel texts. So, why not make use of this *treasure*?

However, since we use homographs, there is the danger of inadvertently using false friends like *oficina* which means 'workshop' in Portuguese and 'office' in Spanish. So, in order to avoid pairing words like these which are not equivalent though homograph, we restricted ourselves to using homographs which occur with the same frequency in parallel text segments. In this way, we are *pre*-selecting words with *similar distributions*. Actually, equal frequency words helped Jean-François Champollion to decipher the Rosetta Stone for there was a name of a King (Ptolemy V) which occurred the same number of times in the "parallel texts" of the stone. Even if two words happen to have the same frequency though they are not equivalent, they end up appearing in different places in the parallel texts. Consequently, they create extreme points that are easily identified by the method described in the next section.

In this way, each pair of texts gives a set of candidate correspondence points from which we draw a line based on linear regression. Points are defined using the co-ordinates of the word positions in each parallel text. For example, if the first occurrence of the homograph word *Reino* occurs at position 2988 in the Portuguese text and at word position 3065 in the Spanish parallel text, then the first point co-ordinates are (2988,3065). Points may fit well to the 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.

3.3 Eliminating Extreme Points

Points obtained from the positions of homographs with equal frequencies in the whole parallel texts are prone to be noisy.

In Figure 1, there are noisy points because their respective homograph words appear in positions quite apart, e.g. the word *último* in pt word position 940 (Point A) was paired with the es word position 2810:

[3] Homepage: http://curia.eu.int.

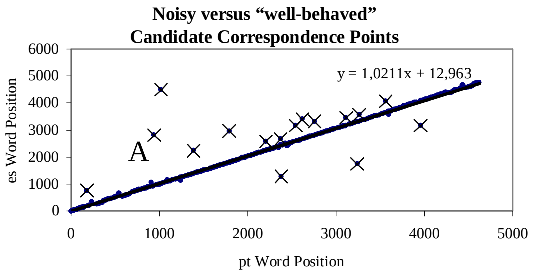

Figure 1: Noisy candidate correspondence points (marked with an ×) versus "well-behaved" candidate correspondence points "in line". The linear regression equation is on the top right corner.

However, this word was *expected somewhere much earlier* in the Spanish text. We should feel reluctant to accept these pairings and that is what the first filter does. It filters out those points which are clearly quite far apart from their *expected* positions to be considered as reliable correspondence points.

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

		Positions		
Word	pt	es	es Expected	Distance
940	último	2810	973	1837
1793	mediante	2965	1844	112
2371	para	2668	2434	234
2381	definidos	1287	2444	1157
3240	vez	1754	3321	1567

Expected distances are computed from the linear regression line equation $y = ax + b$, where a is the line slope and b is the Y-axis intercept (the value of y when x is 0), substituting x for the Portuguese word position. For Figure 1, the expected word position for the word *último* at pt word position 940 is 1,0211 × 940 + 12,963 = 973 and the distance between its expected and real positions is | 973 – 2810 | = 1837.

If we draw a histogram ranging from the smallest to the largest distance, we get:

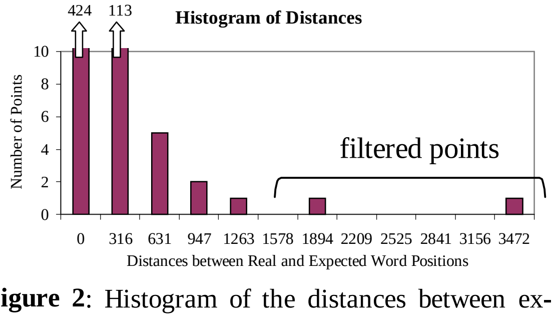

Figure 2: Histogram of the distances between expected and real word positions.

We use the Sturges rule to build this histogram (see Histograms in Samuel Kotz *et al.* 1982). The number of classes (bars or bins) is given by 1 + $\log_2$n, where *n* is the total number of points and the classes size is given by (maximum distance – minimum distance) / number of bins. For example, for Figure 2, we have 547 points and the distances between expected and real positions range from 0 to 3472. Thus, the number of classes is 1 + $\log_2 3472 \cong 10.1 \to 11$ and the classes size is (3472 – 0) / 11 ≅ 315.6. The first class ranges [0 ; 315.6[, the second [315.6 ; 631.3[and so forth.

With this histogram, we are able to identify those words which are too far apart from their expected positions. In Figure 2, the *gap* in the histogram makes clear that there is a discontinuity in the distances between expected and real positions. So, we are confident that extreme points are above 1578. We filter them out of the candidate correspondence points set and proceed to the next filter.

3.4 Linear Regression Line Bands

Confidence bands of linear regression lines (Thomas Wonnacott & Ronald Wonnacott, 1990) help us to identify reliable points, i.e. points which belong to the regression line with a great confidence level (99.9%). The band is wider in the extremes of the linear regression line and narrower in the middle, where the alignment uncertainty is usually higher.

The confidence band is the *error* admitted at an x co-ordinate of a linear regression line. A point (x,y) is considered outside a linear regression line with a confidence level of 99.9% if its y co-ordinate does not lie within [$ax + b - error(x); ax + b + error(x)$], where $ax + b$ is the linear regression line equation and $error(x)$ is the error admitted at the x co-ordinate. The upper and lower limits of the interval are given by the following equation (see Thomas Wonnacott & Ronald Wonnacott 1990, p. 385):

$$y = (ax+b) \pm t_{0.005} s \sqrt{\frac{1}{n} + \frac{(x-\overline{X})^2}{\sum_{i=1}^{n}(x_i-\overline{X})^2}}$$

where:

- $t_{0.005}$ is the *t*-statistics value for a 99.9% confidence interval. We will use the *z*-

statistics instead since $t_{0.005} = z_{0.005} = 3.27$ for large samples of points (above 120);

- n is the number of points;
- s is the standard deviation from the expected value $\hat{y}$ at x (see Thomas Wonnacott & Ronald Wonnacott, 1990, p. 379):

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})}{n-2}}, \text{ where } \hat{y} = ax + b$$

- $\overline{X}$ is the average value of the various x_i:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We start from the points filtered using the histogram technique described in the previous section and build a new linear regression line. Next, we compute the confidence bands using the formulae above 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 *Translation Equivalents*

In order to extract the word translation equivalents from the aligned parallel texts, we used the likelihood similarity measure (Ted Dunning 1994). This measure has already been used for this purpose by Philippe Langlais & Marc El-Bèze (1999) for the extraction of English–French word translation equivalents and provides a good alternative to the sparse data good alternative to the sparse data problem of the specific mutual information.

We start by building a contingency table, like Table 2, for each pair of Portuguese–Spanish words. The table stores the *number of aligned segments* that contain (a) both words (*Comissão* and *Comisión*), (b) the Portuguese word but not the Spanish word, (c) the Spanish word but not the Portuguese word and (d) neither word:

Table 2: Contingency table for the pair *Comissão–Comisión*. *n* is the number of segments. The Portuguese word occurs in 23 segments and the Spanish word occurs in 25. Both words co-occur 16 times.

n: 1671	*Comisión* (25)	× *Comisión*
Comissão (23)	(a) 16	(b) 7
× *Comissão*	(c) 9	(d) 1639

From these tables, we calculate the loglike similarity measure for each pair of words and select the best ranked pair as the translation (Ted Dunning 1993, p. 71):

$$2\left[\log_e L(p_1,k_1,n_1) + \log_e L(p_2,k_2,n_2) \right.$$
$$\left. - \log_e L(p,k_1,n_1) - \log_e L(p_1,k_2,n_2)\right]$$

where:

$$\log_e L(p,n,k) = k\log_e p + (n-k)\log_e(1-p)$$

$$k_1 = a,\ k_2 = c,\ n_1 = a + b,\ n_2 = c + d$$

$$p_1 = \tfrac{k_1}{n_1} = \tfrac{a}{a+b},\ \ p_2 = \tfrac{k_2}{n_2} = \tfrac{c}{c+d} \text{ and}$$

$$p = \tfrac{k_1+k_2}{n_1+n_2} = \tfrac{a+c}{n}$$

which, simplifying, gives:

$$= a\log_e a + b\log_e b + c\log_e c + d\log_e c + n\log_e n$$
$$-(a+c)\log_e(a+c) - (a+b)\log_e(a+b)$$
$$-(b+d)\log_e(b+d) - (c+d)\log_e(c+d)\,.$$

5 *Evaluation*

Table 3: Sample of Portuguese–Spanish word translation equivalents. f(pt,es) is the number of times both words co-occur in aligned segments. f(pt) and f(es) are the frequencies of the Portuguese and Spanish words, respectively.

pt	es	f(pt,es)	f(pt)	f(es)	loglike
de	de	149	188	282	268,1
artigo	artículo	32	35	35	137,8
Regulamento	Reglamento	30	36	41	110,3
zairense	zaireño	23	25	25	107,4
acordo	Acuerdo	35	56	55	104,5
Reino	Reino	16	20	20	70,3
não	no	19	40	34	60,0
Comissão	Comisión	16	23	25	59,9
repartição	reparto	10	11	10	57,8
Abril	abril	6	6	6	39,8

Even though the alignment process may introduce some misalignments, the similarity measure provided a good set of translation equivalents (see Table 3).

The algorithm could find 91 correct equivalents out of the 100 best ranked pairs. The incorrect equivalents were mainly “near misses”, i.e. words which belonged to collocations which are not being taken into account in this work (e.g. *Membros* and *Estados* as in the Portuguese collocation *Estados-Membros* (‘member states’) and the Spanish collocation *Estados miembros*).

The following table summarises the precision values:

Table 4: Translation equivalents precision values for a sample text.

Sample	# Equivalents (% of total)	Precision (wrong equivalents)
best 100	100 (8%)	91% (9)
best 500	500 (40%)	63% (187)
frequency ≥ 10	55 (4%)	82% (10)
frequency ≥ 5	93 (7%)	85% (14)
frequency ≥ 2	341 (27%)	69% (106)
all	1247 (100%)	66% (427)

We expect low precision percentages (around 65%) to rise as collocations are extracted from the parallel texts. However, in the present form, the *best 100* offer the most reliable translation equivalents.

6 Conclusions

In this paper we have presented a method to extract word translation equivalents from Portuguese–Spanish parallel texts aligned with a purely statistics based alignment algorithm. This algorithm selects correspondence points generated from homographs with equal frequencies in parallel text segments.

Confidence bands of linear regression lines help us to identify reliable correspondence points without using empirically found or statistically unsupported heuristics. The filtering of candidate correspondence points we presented in this paper is purely statistical and does not recur to heuristics as in previous work. Also, the alignment is not restricted to sentence or paragraph level for which clearly delimited boundaries markers would be needed. It is made at whatever segment size as long as there are reliable correspondence points. This means that it can result at paragraph, sentence, phrase, term or even word level.

Moreover, the alignment methodology does not depend on the way candidate correspondence points are generated, i.e. although we used homographs with equal frequencies in parallel texts segments, we could have also bootstrapped the process using a small bilingual lexicon to identify equivalents of words or expressions (Dekai Wu 1994; Pascale Fung & Kathleen McKeown 1997; Melamed 1999). This is a particularly good strategy when it comes to distant languages like English and Chinese where the number of homographs is reduced. Aligning languages with such different alphabets requires automatic methods to identify equivalents as Pascale Fung & Kathleen McKeown (1997) presented, increasing the number of candidate correspondence points at the beginning.

As the alignment algorithm is not restricted to paragraphs or sentences, 100% alignment precision may be degraded by language specific term order policies in small segments as Philippe Langlais & Marc El-Bèze (1999) remark. The method is language and character-set independent (see António Ribeiro *et al.* 2000b, for a Portuguese–Chinese case study) and does not assume any a priori language knowledge (namely, small bilingual lexicons), text tagging, well defined sentence or paragraph boundaries nor one-to-one translation of sentences.

The translation equivalents extraction method uses the likelihood similarity measure which has already been used in previous work and the best 100 ranked pairs have a precision above 90%. This is quite good a value if we bear in mind how simple the candidate correspondence points fed into the alignment algorithm are.

7 Future Work

We plan to *re-feed* the alignment algorithm with the best ranked equivalent translations so that more candidate correspondence points may be used for text alignment. Moreover, we will deal with multiword units using a methodology described in Joaquim da Silva *et al.* (1999) in order to extract collocations and their translations.

Although inversions in sentence structures caused some misalignments problems, the bilingual lexicon we got in the end still has a high precision: 91 of the 100 best ranked translation equivalents pairs are correct. This is leading us to analyse them more carefully in order to improve the alignment precision. We will also use other similarity measures to extract translation equivalents in order to test their effects.

8 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.

9 References

Brown, Peter, Lai, Jennifer and Mercer, Robert (1991) "Aligning Sentences in Parallel Corpora". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, U.S.A., pp. 169–176.

Church, Kenneth (1993) "Char_align: A Program for Aligning Parallel Texts at the Character Level". In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, U.S.A., pp. 1–8.

Dagan, Ido, Church, Kenneth and Gale, William (1993) "Robust Word Alignment for Machine Aided Translation". In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, U.S.A., pp. 1–8.

Dunning, Ted (1993) "Accurate Methods for the Statistics of Surprise and Coincidence". In *Computational Linguistics*, volume 19, number 1, pp. 61–74.

Fung, Pascale & McKeown, Kathleen (1994) "Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping". In *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., pp. 81–88.

Fung, Pascale & McKeown, Kathleen (1997) "A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups". In *Machine Translation*, volume 12, numbers 1–2 (Special issue), pp. 53–87.

Gale, William & Church, Kenneth (1991) "A Program for Aligning Sentences in Bilingual Corpora". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, U.S.A., pp. 177–184 (short version). Also (1993) in *Computational Linguistics*, volume 19, number 1, pp. 75–102 (long version).

Kay, Martin & Röscheisen, Martin (1993) "Text-Translation Alignment". In *Computational Linguistics*, volume 19, number 1, pp. 121–142.

Kotz, Samuel, Johnson, Norman and Read, Campbell (1982) *Encyclopaedia of Statistical Sciences*, John Wiley & Sons, New York Chichester Brisbane Toronto Singapore.

Langlais, Philippe & El-Bèze, Marc (1999) "Alignment de Corpus Bilingues: algorithmes et évaluation". In (1999) *Ressources et Évaluations en Ingénierie de la Langue,* Collection Actualité Scientifique, Aupfel-Uref, Paris, France.

Melamed, I. (1999) "Bitext Maps and Alignment via Pattern Recognition". In *Computational Linguistics*, volume 25, number 1, pp. 107–130.

Ribeiro, António, Lopes, Gabriel and Mexia, João (2000a, in press) "Linear Regression Based Alignment of Parallel Texts Using Homograph Words". In Horn, Werner (ed.) (2000, in press) *ECAI 2000: Proceedings of the 14th European Conference on Artificial Intelligence*, IOS Press, Amsterdam, The Netherlands.

Ribeiro, António, Lopes, Gabriel and Mexia, João (2000b, in press) "Aligning Portuguese and Chinese Parallel Texts Using Confidence Bands". In (2000, in press), *Proceedings of the Sixth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2000) – Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin, Germany.

Ribeiro, António, Lopes, Gabriel and Mexia, João (2000c, in press) "Using Confidence Bands for Parallel Texts Alignment". In (2000, in press) *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*.

da Silva, Joaquim, Dias, Gaël, Guilloré, Sylvie and Lopes, José (1999) "Using Localmaxs algorithms for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units". In Barahona, Pedro & Alferes, José (1999), *Progress in Artificial Intelligence – Lecture Notes in Artificial Intelligence*, number 1695, Springer-Verlag, Berlin, Germany, pp. 113–132.

Simard, Michel, Foster, George and Isabelle, Pierre (1992) “Using Cognates to Align Sentences in Bilingual Corpora”. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation TMI-92*, Montreal, Canada, pp. 67–81.

Simard, Michel & Plamondon, Pierre (1998) “Bilingual Sentence Alignment: Balancing Robustness and Accuracy”. In *Machine Translation*, volume 13, number 1, pp.59–80.

Wu, Dekai (1994) “Aligning a Parallel English–Chinese Corpus Statistically with Lexical Criteria”. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, Las Cruces, New Mexico, U.S.A., pp. 80–87.

Wonnacott, Thomas & Wonnacott, Ronald (1990) *Introductory Statistics*, 5th edition, John Wiley & Sons, New York Chichester Brisbane Toronto Singapore.