Bitext Correspondences through Rich Mark-up

Raquel Martínez
Departamento de Sis. Informáticos y Programación, Facultad de Matemáticas
Universidad Complutense de Madrid
e-mail:raquel@eucmos.sim.ucm.es

Joseba Abaitua
Facultad de Filosofía y Letras
Universidad de Deusto, Bilbao
e-mail:abaitua@fil.deusto.es

Arantza Casillas
Departamento de Automática, Universidad de Alcalá de Henares
e-mail:arantza@aut.alcala.es

Abstract

Rich mark-up can considerably benefit the process of establishing bitext correspondences, that is, the task of providing correct identification and alignment methods for text segments that are translation equivalences of each other in a parallel corpus. We present a sentence alignment algorithm that, by taking advantage of previously annotated texts, obtains accuracy rates close to 100%. The algorithm evaluates the similarity of the linguistic and extra-linguistic mark-up in both sides of a bitext. Given that annotations are neutral with respect to typological, grammatical and orthographical differences between languages, rich mark-up becomes an optimal foundation to support bitext correspondences. The main originality of this approach is that it makes maximal use of annotations, which is a very sensible and efficient method for the exploitation of parallel corpora when annotations exist.

1 Introduction

Adequate encoding schemes applied to large bodies of text in electronic form have been a main achievement in the field of humanities computing. Research in computational linguistics, which since the late 1980s has resorted to methodologies involving statistics and probabilities in large corpora, has however largely neglected the existence and provision of extra information from such encoding schemes. In this paper we present an approach to sentence alignment that crucially relies on previously introduced annotations in a parallel corpus. Following (Harris 88), corpora containing bilingual texts have been called "bitexts" (Melamed 97), (Martínez et al. 97).

The utility of annotated bitexts will be demonstrated by the proposition of a methodology that crucially takes advantage of rich mark-up to resolve bitext correspondences, that is, the task of providing correct identification and alignment methods for text segments that are translation equivalencies of each other (Chang & Chen 97). Bitext correspondences provide a great source of information for applications such as example and memory based approaches to machine translation (Sumita & Iida 91), (Brown et al. 93), (Collins et al. 96); bilingual terminology extraction (Kupiec 93), (Eijk 93), (Dagan et al. 94), (Smajda et al. 96); bilingual lexicography (Catizione et al. 93), (Daille et al. 94), (Gale & Church, 91b); multilingual information retrieval (SIGIR 96), and word-sense disambiguation (Gale et al. 92), (Chan & Chen 97). Moreover, the increasing availability of running parallel text in annotated form (e.g. WWW pages), together with evidence that poor mark-up (as HTML) will progressively be replaced by richer mark-up (e.g. SGML/XML), are good enough reasons to investigate methods that benefit from such encoding schemes.

We first provide details of how a bitext sample has been marked-up, with particular emphasis on the recognition and annotation of proper nouns. Then we show how sentence alignment relies on mark-up by the application of a methodology that resorts to annotations to determine the similarity between sentence pairs.

This is the 'tags as cognates' algorithm, *TasC*.

2 Bitext tagging and segmentation

A large bitext has been compiled consisting of a collection of administrative and legal bilingual documents written both in Spanish and Basque, with close to 7 million words in each language. For the experiments, we have worked on a representative subset of around 500,000 words in each language. Several stages of automatic tagging, based on pattern matching and heuristics, were undertaken, rendering different descriptive levels:

- General encoding (paragraph, sentence, quoted text, dates, numbers, abbreviations, etc.).
- Document specific tags that identify document types and define document internal organisation (sections, divisions, identification code, number and date of issue, issuer, lists, itemised sections, etc.).
- Proper noun tagging (identification and categorisation of proper nouns into several classes, including: person, place, organisation, law, title, publication and uncategorised).

This collection of tags (shown in Table 1) reflects basic structural and referential features, which appear consistently at both sides of the bitext. Although the alignment of smaller segments (multi-word lexical units and collocations) will require more expressive tagging, such as part-of-speech tagging (POS), for the task of sentence alignment, this is not only unnecessary, but also inappropriate, since it would introduce undesired language dependent information. The encoding scheme has been based on TEI's guidelines for SGML based mark-up (Ide & Veronis 95).

2.1 Proper noun tagging

As for many other text processing applications, proper noun tagging plays a key role in our approach to sentence alignment. It has been reported that proper nouns reach up to 10% of tokens in text (newswire text (Wakao et al. 96) and (Coates-Stephens 92)) and one third of noun groups (in the *Agence France Presse* flow (Wolinski et al. 95)). We have calculated that proper nouns constitute a 15% of the tokens in our corpus. The module for the recognition of proper nouns relies on patterns of typography (capitalisation and punctuation) and on contextual information (Church 88). It also makes use of lists with most common person, organisation, law, publication and place names. The tagger annotates a multi-word chain as a proper noun when each word in the chain is uppercase initial. A closed list of functional words (prepositions, conjunctions, determiners, etc.) is allowed to appear inside the proper noun chain, see examples in Table 2. A collection of heuristics discard uppercase initial words in sentence initial position or in other exceptional cases.

In contrast with other known classifications (e.g. MUC-6 95), we exclude from our list of proper nouns time expressions, percentage expression, and monetary amount expressions (which for us fall under a different descriptive level). However, on top of organisation, person and location names, we include other entities such as legal nomenclature, the name of publications as well as a number of professional titles whose occurrence in the bitext becomes of great value for alignment.

2.2 Bitext asymmetries

Because our approach to alignment relies on consistent tagging, bitext asymmetries of any type need to be carefully dealt with. For example, capitalisation conventions across languages may show great divergences. Although, in theory, this should not be the case between Spanish and Basque, since officially they follow identical conventions for capitalisation (which are by the way the same as in French), in practise these conventions have been interpreted very differently by the writers of the two versions (lawyers in Spanish and translators in Basque). In the Basque version, nouns referring to organisations *saila* 'Department', professional titles *diputatua* 'Deputy', as well as many orographic or geographical sites *arana* 'Valley', are often written in lowercase, while in the Spanish original documents these are normally written in uppercase (see Table 2). These nouns belong to the type described as 'trigger' words by (Wakao et al. 96), in the sense that they permit the identification of the tokens surrounding them as proper nouns. Then, it has been required to resort to contextual information. The results of the resolution of these singularities are shown in Table

Descriptive levels	Tagset
General encoding	<p>, <s>, <num>, <date> <abbr>, <q>
Document especific	<div>, <classCode> <keywords>, <dateline>, <list><seg>
Proper nouns	<rs>

Table 1: Tagset used for sentence alignment

Proper Noun Classes	Spanish	Basque
Person	*Ana Fernández Gutierrez-Crespo*	*Ana Fernández Gutierrez-Crespo*
Place	*Valle de Arratia*	*Arratiko arana*
Organisation	*Departamento de Presidencia*	*Lehendakaritza Saileko*
Law	*Real Decreto Legislativo*	*Legegintzazko Erret Dekretuko*
Title	*Diputado Foral de Urbanismo*	*Hirigintza foru diputatua*
Publication	*Boletín Oficial de Bizkaia*	*Bizkaiko Aldizkari Ofizialean*
Uncategorised	*Anexo*	*eraskin*

Table 2: Examples of proper nouns

3.

3 Using tags as cognates for sentence alignment

Algorithms for sentence alignment abound and range from the initial pioneering proposals of (Brown et al. 91), (Gale & Church 91a), (Church 93), or (Kay & Roscheisen 93), to the more recent ones of (Chang & Chen 97), or (Tillmann et al. 97). The techniques employed include statistical machine translation, cognates identification, pattern recognition, and digital signal and image processing. Our algorithm, as (Simard et al. 92), and (Melamed 97) employs cognates to align sentences; and similar to (Brown et al. 91), it also uses mark-up for that purpose. Its singularity does not lie on the use of mark-up as delimiter of text regions (Brown et al. 91) in combination with other techniques, but on the fact that it is the sole foundation for sentence alignment. We call it the 'tags as cognates' algorithm, *TasC*. This algorithm is not disrupted by word order differences or small asymmetries in non-literal translation, and, unlike other reported algorithms (Melamed 97), it possesses the additional advantage of being portable to any pair of languages without the need to resort to any language-specific heuristics. Provided an adequate and consistent bitext mark-up, sentence alignment becomes a simple and accurate process also in the case of typologically disparate or orthographically distinct language pairs for which techniques based on lexical cognates may be problematic. One of the best consequences of this approach is that the burden of language dependent processing is dispatched to the monolingual tagging and segmentation phase.

3.1 Similarity calculus between bitexts

The alignment algorithm establishes similarity metrics between candidate sentences which are delimited by corresponding mark-up. Dice's coefficient is used to calculate these similarity metrics (Dice 45). The coefficient returns a real numeric value in the range 0 to 1. Two sentences which are totally dissimilar in the content of their internal mark-up will return a Dice score of 0, while two identical contents will return a Dice score of 1.

For two text segments, P and Q, one in each language, the formula for Dice's similarity coefficient will be:

$$Dice(P,Q) = \frac{2F_{PQ}}{F_P + F_Q}$$

where F_{PQ} is the number of identical tags that P and Q have in common, and F_P and F_Q are the number of tags contained by each text segment P and Q.

Since the alignment algorithm determines the best matching on the basis of tag similarity, not only tag names used to categorise different cognate classes (number, date, abbreviation, proper noun, etc.), but also attributes contained by these tags may help identify the cognate itself: <num num=57>57</num>. Furthermore, attributes

	Spanish			Basque		
Proper Noun Classes	Precision	Recall	% Spanish PN	Precision	Recall	% Basque PN
Person	100%	100%	4.48%	100%	100%	4.76%
Place	100%	100%	6.38%	100%	100%	6.95%
Organisation	99.2%	97.8%	23.96%	100%	100%	24.17%
Law	99.2%	99.2%	47.93%	100%	100%	46.15%
Title	100%	100%	6.55%	97.2%	97.2%	6.59%
Publication	100%	100%	2.58%	100%	100%	2.74%
Uncategorised	100%	100%	8.10%	100%	100%	8.60
Total	99.4%	99.1%	100%	99.8%	99.8%	100%

Table 3: Results of proper noun identification

may serve also to subcategorise proper noun tags: <rs type=place>Bilbao</rs>.

Such subcategorisations are of great value to calculate the similarity metrics. If mark-up is consistent, the correlation between tags in the candidate text segments will be high and Dice's coefficient will come close to 1. For a randomly created bitext sample of source sentences, Figure 1 illustrates how correct candidate alignments have achieved the highest Dice's coefficients (represented by ’*’s), while next higher coefficients (represented by ’o’s) have achieved significant lower values. It must be noted that the latter do not correspond to correct values.

The difference mean between Dice's coefficients corresponding to correct alignments and next higher values is:

$$M = \frac{\sum_{i=1}^{n}(DCc_i - DCw_i)}{n} = 0.45$$

Where for a given source sentence i, DCc_i represents Dice's coefficient corresponding to its correct alignment and DCw_i represents the next higher value of Dice's coefficients for the same source sentence i. In all the cases, this difference is greater than 0.2.

For consistently marked-up bitexts, these results show that sentence alignment founded on the similarity between annotations can be robust criterion.

Figure 2 illustrates how the Dice's coefficient is calculated between candidate sentences to alignment.

3.2 The strategy of the *TasC* algorithm

The alignment of text segments can be formalised by the matching problem in bipartite graphs. Let $G = (V, E, U)$ be a bipartite graph, such that V and U are two disjoint sets of vertices, and E is a set of edges connecting vertices from V to vertices in U. Each edge in E has associated a cost. Costs are represented by a cost matrix. The problem is to find a perfect matching of G with minimum cost. The minimisation version of this problem is well known in the literature as the *assignment problem.*

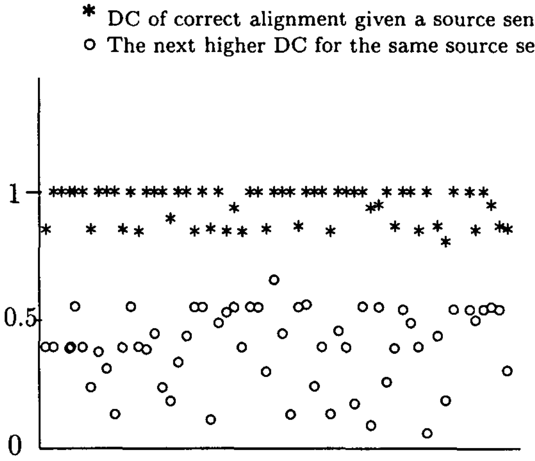

Figure 1: Values of Dice's coefficient between corresponding sentences

Applying the general definition of the problem to the particular case of sentence alignment: V and U represent two disjoint sets of vertices corresponding to the Spanish and Basque sentences that we wish to align. In this case, each edge has not a cost but a similarity metric quantified by Dice's coefficient. The fact that vertices are materialised by sentences detracts gen-

Spanish Sentence:
<s id=sESdoc5-4>Habiéndose detectado en el anuncio publicado en el número<num num=79> 79 </num> de fecha <date date=27/04>27 de abril</date> de este <rs type=publication>Boletín</rs>, la omisión del primer párrafo de la <rs type=law>Orden Foral</rs> de referencia, se procede a su íntegra publicación.</s>

Basque Sentence:
<s id=sEUdoc5-5>Agerkaria honetako <date date=27/04>apirilaren 27ko</date> <num num=79>79k.an </num> argitaratutako iragarkian aipameneko <rs type=law>Foru Aginduaren</rs> lehen lerroaldea ez dela geri detektatu ondoren beraren argitarapen osoa egitera jo da.</s>

The common tags are: <date date=27/04>, <num num=79>, <rs type=law>
The Dice's similarity coefficient will be: Dice(P,Q)= 2x3 / 4+3 = 0.857

Figure 2: Similarity calculus between candidate sentences

erality to the assignment problem and makes it possible to add constraints to the solutions reported in the literature. These constraints take into account the order in which sentences in both the source and target texts have been written, and capture the prevailing fact that translators maintain the order of the original text in their translations, which is even a stronger property of specialised texts.

By default, a whole document delimits the space in which sentence alignment will take place, although this space can be customised in the algorithm. The average number of sentences per document is approximately 18. Two types of alignment can take place:

- 1 to 1 alignment: when one sentence in the source document corresponds to one sentence in the target document (94.39% of the cases).
- N to M alignment: when N sentences in the source document correspond to M sentences in the target document (only 5.61% of the cases). It includes cases of 1-2, 1-3 and 0-1 alignments.

Both alignment types are handled by the algorithm.

3.3 The algorithm

The *TasC* algorithm works in two steps:

1. It obtains the similarity matrix S from Dice's coefficients corresponding to candidate alignment options. Each row in S represents the alignment options of a source sentence classified in decreasing order of similarity. In this manner, each column represents a preference position (1 the best alignment option, 2 the second best and so on). Therefore, each $S_{i,j}$ is the identification of one or more target sentences which match the source sentence i in the preference position j. In order to obtain the similarity matrix, it is not necessary to consider all possible alignment options. Constraints regarding sentence ordering and grouping greatly reduce the number of cases to be evaluated by the algorithm. In the algorithm each source sentence x_i is compared with candidate target sentences y_j as follows: (x_i, y_i); $(x_i, y_j y_{j+1}$ $\ldots$, where $y_j y_{j+1}$ represents the concatenation of y_j with y_{j+1}. The algorithm module that deals with candidate alignment options can be easily customised to cope with different bitext configurations (since bitexts may range from a very simple one-paragraph text to more complex structures). In the current version of the algorithm seven alignment options are taken into account.
2. The *TasC* algorithm solves an assignment problem with several constraints. It aligns sentences by assigning to each *ith* source sentence the $S_{i,j}$ target option with minimum j value, that is, the option with more similarity. Furthermore, the algorithm solves the possible conflicts when a sentence matches with other sentences already aligned. The average cost of the algorithm, experimentally contrasted , is linear in the size of the input, although in the worst case the cost is bigger.

The result of sentence alignment is reflected in the bitext by the incorporation of the attribute 'corresp to sentence tags, as can be seen

Cases	%Corpus	% Accuracy
1 - 1	94.39%	100%
N - M	5.61%	99.68%

Table 4: TasC Algorithm results

in Figure 3. This attribute points to the corresponding sentence identification code in the other language.

4 Evaluation

The current version of the algorithm has been tested against a subcorpus of 500,000 words in each language consisting of 5,988 sentences and has rendered the results shown in Table 4.

The accuracy of the 1 to 1 alignment is 100%. In the N to M case only 1 error occurred out of 314 sentences, which reaches 99.68% accuracy. The algorithm to sentence alignment has been designed in such a modular way that it can easily change the tagset used for alignment and the weight of each tag to adapt it to different bitext annotations. The current version of the algorithm uses the tagset shown in Table 1 without weights.

5 Future work

Once sentences have been aligned, the next step is the alignment of sentence-internal segments. The sentence will delimit the search space for this alignment, and hence, by reducing the search space, the alignment complexity is also reduced.

5.1 Proper noun alignment

Proper nouns are a key factor for the efficient management of the corpus, since they are the basis for the indexation and retrieval of documents in the two versions. For this reason, at present we are concerned with proper noun alignment, something which is not usually done in the mapping of bitexts. The alignment is achieved by resorting to:

- The identification of cognate nouns, aided by a set of phonological rules that apply when Spanish terms are taken to produce loan words in Basque.
- The restriction of cognate search space to previously aligned sentences, and
- The application of the *TasC* algorithm adapted to proper noun alignment.

5.2 Alignment of collocation

The next step is the recognition and alignment of other multi-word lexical units and collocations. Due to the still unstable translation choices of much administrative terminology in Basque, on top of the considerable typological and structural differences between Basque and Spanish, many of the techniques reported in the literature (Smadja et al. 96), (Kupiec 93) and (Eijk 93) cannot be effectively applied. POS tagging combined with recurrent bilingual glossary lookup is the approach we are currently experimenting with.

6 Conclusions

We have presented a sentence alignment approach that, by taking advantage of previously introduced mark-up, obtains accuracy rates close to 100%. This approach is not disrupted by word order differences and is portable to any pair of languages without the need to resort to any language specific heuristics. Provided and adequate and consistent bitext mark-up, sentence alignment becomes an accurate and robust process also in the case of typologically distinct language pairs for which other known techniques may be problematic. The *TasC* algorithm has been designed in such a modular way that it can be easily adapted to different bitext configurations as well as other specific tagsets.

7 Acknowledgements

This research is being partially supported by the Spanish Research Agency, project ITEM, TIC-96-1243-C03-01.

References

Brown, P., Lai, J.C., Mercer, R. (1991). Aligning Sentences in Parallel Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* 169-176, Berkeley, 1991.

Brown, P., Della Pietra, V., Della Pietra, S., Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19(2):263-301 1993.

Catizone, R., Russell, G., Warwick, S. (1993). Deriving Translation Data from Bilingual Texts. *Proccedings of the First International Lexical Acquisition Workshop,* Detroit, MI, 1993.

Chang, J. S., Chen, M. H. (1997). An Alignment Method for Noisy Parallel Corpora based on Image Processing Techniques. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 297-304, 1997.

Spanish Sentence:
<s id=sESdoc5-4 corresp=sEUdoc5-5>Habiéndose detectado en el anuncio publicado en el número<num num=79> 79 </num> de fecha <date date=27/04>27 de abril</date> de este <rs type=publication>Boletín</rs>, la omisión del primer párrafo de la <rs type=law>Orden Foral</rs> de referencia se procede a su íntegra publicación.</s>

Basque Sentence:
<s id=sEUdoc5-5 corresp=sESdoc5-4>Agerkaria honetako <date date=27/04> apirilaren 27ko</date> <num num=79>79k.an </num> argitaratutako iragarkian aipameneko <rs type=law>Foru Aginduaren</rs> lehen lerroaldea ez dela geri detektatu ondoren beraren argitarapen osoa egitera jo da.</s>

Figure 3: Results of sentence alignment expressed by the corresp attribute

Church, K.W. (1988). A Stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the Second Conference on Applied Natural Language Processing,* 136-143, 1988. Association for Computational Linguistics.

Church, K.W. (1993). Char_Align: A Program for Aligning Parallel Texts at the Character Level. *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics,* Columbus, USA 1993.

Coates-Stephen, S. (1992). The Analysis and Acquisition of Proper Names for Robust Text Understanding, *Ph.D. Department of Computer Science of City University,* London, England, 1992.

Collins, B., Cunningham, P., Veale, T. (1996). An Example Based Approach to Machine Translation. *Expanding MT Horizonts: Proceedings of the Second Conference of the Association for Machine Translation in the Americas:AMTA-96,* 125-134, 1996.

Daille, B., Gaussier, E., Lange, J.M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. *Proceedings of the 15th International Conference on Computational Linguistics,* 515-521, Kyoto, Japan.

Dagan, I., Church, K. (1994). Termigh: Identifying and translating Technical Terminology. *Proceedings Fourth Conference on Applied Natural Language Processing (ANLP-94),* Stuttgart, Germany, 34-40, 1994. Association for Computational Linguistics.

Dice, L.R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology,* 26, 297-302.

Eijk, P. van der. (1993). Automating the acquisition of Bilingual Terminology. *Proceedings Sixth Conference of the European Chapter of the Association for Computational Linguistic,* Utrecht, The Netherlands, 113-119, 1993.

Gale, W., Church, K.W. (1991a). A Program for Aligning Sentences in Bilingual Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* 177-184, Berkeley, 1991a.

Gale, W., Church, K. W. (1991b). Identifying Word Correspondences in Parallel Texts. *Proceedings of the DARPA SNL Workshop,* 1991.

Gale, W., Church, K. W., Yarowsky, D. (1992). Using Bilingual Materials to Develop Word Sense Disambiguation Methods. *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI-92), 101-112, Montreal, Canada 1992.

Harris, B. (1988). Bi-Text, a New Concept in Translation Theory. *Language Monthly #54,* 1988.

Ide,N., Veronis, J. (1994). MULTEXT (Multilingual Text Tools and Corpora.) *Proceedings of the International Workshop on Sharable Natural Language Resources,* 90-96, 1994.

Ide, N., Veronis, J. (1995). The Text Encoding Initiative: Background and Contexts. *Dordrecht: Kluwer Academic Publishers,* 1995.

Kay, M., Roscheisen, M. (1993). Text-Translation Alignment. *Computational Linguistics,* 19:1, 121-142, 1993.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. *Proceedings of the 31st Annual Meeting of the ACL,* Columbus, Ohio, 17-22. Association for Computational Linguistics 1993.

Martínez, R., Casillas, A., Abaitua, J. (1997). Bilingual parallel text segmentation and tagging for specialized documentation. *Proceedings of the International Conference Recent Advances in Natural Language Processing,* RANLP'97, 369-372, 1997.

Melamed, I.D. (1997). A Portable Algorithm for Mapping Bitext Correspondence. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics,* 305-312, 1997.

MUC-6. (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufman.

SIGIR. (1996). *Workshop on Cross-linguistic Multilingual Information Retrieval,* Zurich, 1996.

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

Smadja, F., McKeown, K., Hatzivassiloglou, V.(1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics* Volume 22, No. 1, 1996.

Sumita, E., Iida, H. (1991). Experiments and prospect of example-based machine translation. *Proceedings of the Association for Computational Linguistics.* Berkeley,185-192, 1991.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A. (1997). A DP based Search Using Monotone Alignments in Statistical Translation. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics,* 289-296, 1997.

Wakao, T., Gaizauskas, R., Wilks, Y. (1996). Evaluation of an Algorithm for the Recognition and Classification of Proper Names. *Proceedings of the 16th International Conference on Computational Linguistics (COLING96),*418-423, 1996.

Wolinski, F., Vichot, F., Dillet, B. (1995). Automatic Processing of Proper Names in Texts. *The Computation and Language E-Print Archive, http : //xxx.lanl.gov/list/cmp – lg/*9504001