ALIGNMENT OF BILINGUAL NAMED ENTITIES IN FRENCH -ARABIC PARALLEL CORPORA

Authoul Abdulhay, Olivier Kraif

ALIGNMENT OF BILINGUAL NAMED ENTITIES IN FRENCH – ARABIC PARALLEL CORPORA

AUTHOUL ABDULHAY
LIDILEM-Université Stendhal Grenoble3
Grenoble, France, 38000
Authoul.Abdulhay@u-grenoble3.fr

OLIVIER KRAIF
LIDILEM- Université Stendhal Grenoble3
Grenoble, France, 38000
olivier.kraif@u-grenoble3.fr

ABSTRACT

Researches in the field of Named Entity recognition and alignment are of strong interest for various applications of natural language processing, such as Cross Lingual Information Retrieval, document management, question-answering systems, data mining etc. But in the processing of Arabic language, the task is particularly difficult and few resources are available to cope with these difficulties.
In this paper, we present a simple method of character transcoding - a kind of transliteration that we call character reduction - *which could improve an aligning system for Named Entities such as anthroponyms and toponyms. This system has been applied and evaluated on a French-Arabic parallel corpus that has been used during the Arcade 2 evaluation campaign.*
The purpose of this method is to bring the graphic forms of both languages close together as much as possible, in order to increase aligning precision. An outcome of such aligning is the ability to project on the target language (Arabic) annotations that has been done on the source language, for which more tools and resources are available (French, English, etc.).

Keywords: *Bilingual aligning, transliteration, anthroponyms, toponyms, Named Entities.*

1. INTRODUCTION

Bilingual aligning consists in identifying matches between units at various levels of granularity: paragraphs, sentences or lexemes.
In the scope of multilingual resource extraction, this study addresses the issue of aligning bilingual Named Entity pairs: more precisely personal names (anthroponyms) and place names (toponyms). These units have the advantage of being relatively stable during translation, and constitute first choice indications in analysing and searching documents from a referential point of view. Thus, Named Entities (NE's) and their translations, when they are comparable, i.e. written with the same alphabet, give interesting clues for matching equivalent sentences [9].

The automatic Named Entity recognition, using monolingual techniques, requires appropriate language resources (dictionaries, POS-tagging and lemmatization, syntactic pattern extraction, etc.). So far, few works have been devoted to the detection of Arabic NE's. For this language, few superficial indications are available, and detection techniques are complex and costly to implement. As other languages – like English – are better equipped in terms of tools and resources, we believe that it would be interesting, before developing a system for Arabic, to recover what is already available for other languages. Bilingual aligning can be a solution to this principle of recovery: if we are able to detect NE's in an aligned corpus with Arabic, then by matching NE source words with their equivalents in Arabic, we can build an annotated corpus which may be useful for the construction of a future system. We call it the bilingual projecting method.
Therefore we focus more specifically on the problem of aligning between languages with different alphabet, such as English-Arabic or French-Arabic. To find equivalents between NE's in bilingual parallel texts, we may rely on the fact that some units share phonetic similarities: it may be a result of a transliteration made by the translator, or the consequence of a common origin in the case of cognate words. As [16] we distinguish between two different relations between name pairs: *transliteration* and *translation*.

e.g. Transliteration case: Milosevic→ميلوسيفيتش [mylwsyfyt$].
e.g. Translation case: Côte d'ivoire →ساحل العاج [saHl alEaj].

To take advantage of transliteration cases, we propose to develop a specific transcoding scheme designed to improve the aligning task.
This paper is organised as follows: Section 2 presents the previous work. Section 3 explains our methodology to develop a transcoding system using a bilingual corpus. Section 4 presents the experiments that has been conducted to evaluate the system and finally Section 5 concludes the paper and indicates advantages, limitations, and possible applications of our method.

2. RELATED WORK

In 1991, some researchers as Brown et al. [3], and Gale & Church [7] developed relatively simple and linguistically poor techniques for bilingual aligning, achieving a good alignment quality at sentence level. Then Débili & Sammouda (1992) [6] show how to implement a "virtuous circle", by extracting from sentence alignment a series of lexical correspondences, from which it is possible to consolidate sentence alignment in return.
Various methods of correspondences extraction were used, as Melamed (1995) [10] implementation of "competitive linking algorithm". An association score is computed for possible correspondences, which compete with each other to find the best pairs. Kraif & Chen (2004) [8] used a null hypothesis which calculates the probability of two units to be not equivalents, basing on word co-occurrence, word distributions, graphic resemblance, word positions, and word parts-of-speech.

In order to extract EN pairs, Arbabi et al. (1994) [1] have developed a system which produces multiple English spellings for Arabic anthroponyms. At first, the system inserts the appropriate lacking vowels in Arabic name ("*vowelisation* phase") and then converts the name into a phonetic representation in order to build the most likely Latin spellings. Only the Arabic names that comply with strict Arabic morphological rules are processed – and the other names are ignored.
The system combines a knowledge-based system with neural networks to achieve an error rate below 3.1%, but rejects 55% of invalid names.
Their system requires language resources to determine the pronunciation of Arabic words, and the *vowelisation* phase needs to implement a lemmatizer and a morphosyntactic analyser. In addition, many personal names don't comply with the morphological rules, particularly borrowed words and foreign names [13].

To bring strings of two different languages close together, in order to retrieve matches, it is possible to work simultaneously on both languages. Meng et al. (2001) [11] introduced an algorithm for transliteration of OOV names (the "Out-Of-Vocabulary" new names which appear almost daily, and constitute unregistered vocabulary in the lexicon), from English to Chinese in the context of Cross-Language Spoken Document Retrieval. They proposed a method of English-Chinese transliteration, based on: English grapheme-phoneme conversion and interlingual phonological rules (i.e. rules showing the links between English phonemes and Chinese phonemes and between Chinese character-pinyin and basic Chinese syllables). They showed that bigram models are most effective (among the various n-grams) for the recovery. A lattice of Chinese phonemes is built by the most likely Chinese syllables found. These rules have been obtained by aligned parallel data using Transformation-based Error driven Learning (TEL). However, the manually listed rules are unable to balance all the contradicting solutions and yield a high error rate (around 50%).

In this work, we have not implemented a complete transliteration system, because as mentioned previously, getting the pronunciation of an Arabic word involves identification of short vowels, which is a complex issue requiring advanced specialised tools and processing (lemmatizing, morphosyntactic analysis, etc.) – and we do not yet have such tools which give satisfactory results.
Darwish et al. (2001) [5] have used a simpler transliteration scheme, using a 1-n mapping of English and Arabic characters, in order to match unknown word pairs in a CLIR application. As them, we propose to use a simple transformation scheme, but we think that it is possible to take a better advantage of the graphic similarities, by working on both languages, applying a method that we call *reduction*.

Our matching method, described in Kraif & Chen (2004) [8], mainly relies on distributional information for lexical correspondence extraction. But, in the latter paper, we showed that for a French-English corpus, the use of cognate comparison could improve results in a significant way: for instance, results were 6% higher than Melamed (1998) method B. Here we want to show that the graphic comparison may bring useful additional information even for the French-Arabic pair, in order to improve the results.

3. METHOD

3.1. TRANSLITERATION MODEL

The possibility of comparing words at character level may yield interesting indications that can improve the quality of bilingual aligning, either at sentence or at lexical level. When source and target languages don't share the same alphabet, comparisons have to be done through a transliteration process.

One of the more widespread transliteration schemes in Natural Language Processing is the *Buckwalter* transliteration, shown in Figure 1. It is especially designed for automatic processing, because it uses standard ASCII characters, encoded on 7 bits, and it is reversible (with a one-to-one mapping between Arabic and ASCII characters). As it uses just individual characters (monograms) to convert those of the source word, it does not generate ambiguities in the transliterated text that were not in the source spelling (such as other transliteration schemes such as QALAM). Indeed

Buckwalter does not ignore the silent letters, because the transliteration must stick to the source spelling, and these letters can be used to distinguish words [2].

As it has not been designed to give a phonetic representation, Buckwalter does not meet to our requirements. Even though, for some NE's, the transliterated spelling shows graphics similarities, these similarities are difficult to process automatically in a reliable way.

e.g. Buckwalter: *Ignacio* → *AnyAsyw*
Buckwalter: *Ramonet* → *rAmwnh*

To cope with these limitations, we propose another transcoding scheme which is partially based on phonetic properties.

3.2. REDUCTION OPERATION

This reduction consists of a dual system of transcoding, regarding on one hand the *α-Latin* spelling (we note *α-Latin* for "Latin-based alphabet"), and on the other hand, Buckwalter spelling, as shown in Figure 1.

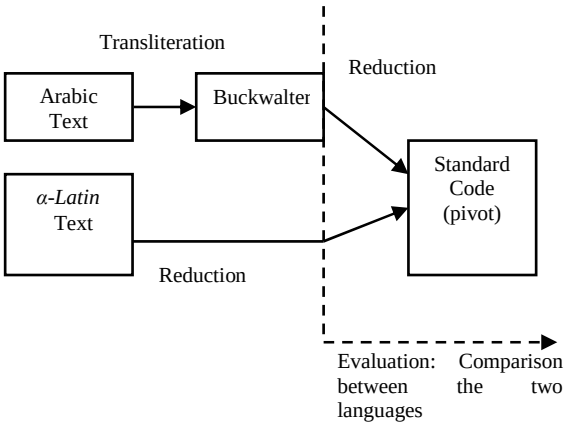

Figure 1: Transcoding system

Reduction rules aims at reducing as far as possible formal differences between both character sets: for instance, on *α-Latin* texts, it converts accented characters in non-accented. These rules are intended to increase similarity, insofar as "Buckwalter" uses no diacritical mark as accents, cedilla, etc.

To measure the graphic similarity, we use an algorithm that extracts the maximum substring (MS) shared by compared forms. If this substring represents 2 / 3 of the longest word [9], we consider the surface forms as graphically similar.
In the example below, the length of the longest common substrings is 1 and 2:
e.g. MS (Ig**n**acio, A**n**yAsyw) = **n**
MS (Ra**mo****n**et, rA**m**w**n**h) = **m-n**

To improve this score, one should not take the case (lower/upper case differences) into account:

e.g. MS (Ig**na**cio A**n**y**A**syw) = **n-a**
MS (**Ram****on**et, **rAm**w**n**h) = **r-a-m-n**

Furthermore, let's suppose that we identify 'w' to 'o', 's' to 'c' and 'y' to 'i'. We now have:

e.g. MS (Ig**nacio**, A**n**y**Asyw**) = **n-a-c-i-o**
MS (**Ramon**et, **rAmwn**h) = **r-a-m-o-n**

This time, the resemblance seems to be relevant enough to be used it in a probabilistic framework.
Such a reduction operation has already been used for a multilingual system, in the work of Pouliquen et al. (2005) [12], where units in every language are converted into an "Internal Standard Representation" to facilitate the recognition of proper names in multilingual news articles (α-Latin and other alphabets). But this Standard representation is made to be shared by many languages, and does not maximize the similarity for a given language pair, as we do.

The reduction operation resulted in the elimination of many distinctions. With this simplification, we lose reversibility: therefore this transcoding is not applied on the corpus itself, which would lead to an impoverishment of the text, but is only implemented at the stage of units' comparison. These rules may affect monogram as well as Polygram, as shown in figure 2.

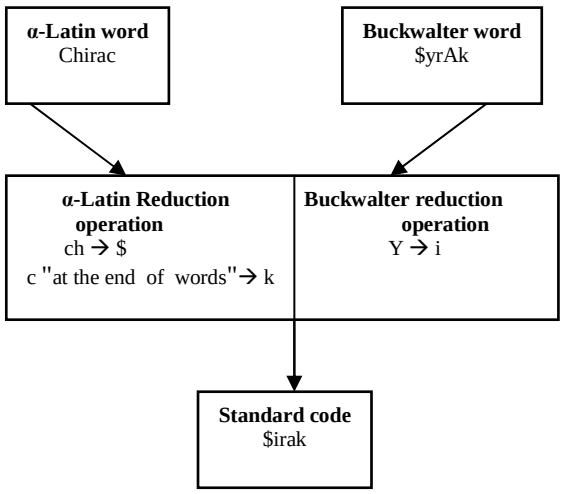

Figure 2: Reduction operation

3.3. LEARNING PHASE

To optimize our transcoding system, we developed our rules (transliteration and reduction) through a learning stage, based on a corpus established manually from UN texts, with 199 anthroponyms pairs and 214 toponyms pairs.

The set of transcoding rules has been built from our corpus by successive refinements, in an incremental way:

- For a given set of rules, we evaluate its efficiency by computing *R*, the number of pairs that are considered as similar, using the MS comparison described above.
- At each step, whenever a rule allows improvement of *R*, we add it to the set of rules, and we repeat this procedure until stability of all the rules. To avoid interference and confusion, we have eliminated competing rules (rules that are applied on the same characters).

3.4. SIMILARITY DETECTION RESULTS

To learn the rules, and evaluate the results, we have used two corpora of NE pairs:

- *Arcade 2* [4] list, including 383 NE pairs.
- *UN list* that consists in 413 UN pairs, extracted from United Nation texts (http://www.un.org/french/documents/instruments/subj_fr.asp)

In order to evaluate the generality of the resulting sets of rules, we used alternatively UN list, and Arcade 2 list as learning corpus. Then, for each set (which were quite similar), we computed *R* on both corpus.
Results are displayed in Table 1:

Corpus	learning corpus	R (number of similar pairs)	Total Number of couples	Recall
Arcade 2 list	Arcade 2 (*reduction* rules-1)	304	383	79,3%
UN list	Arcade 2 (reduction rules-1)	333	413	80,6%
Arcade 2 list	UN (reduction rules- 2)	306	383	79,8%
UN list	UN (reduction rules-2)	350	413	84,7%

Table 1 : Results of our transcoding system for similarity detection

These results show that a given set of rules may apply to a new corpus without an important loss of efficiency: there is no significant difference between corpora with *rules-1* and only 5% for *rules-2* on UN list - but it yields the same results on Arcade 2 list than *rules-1*. Anyway, *rules-2* seem to be more efficient in any case, may be because the learning corpus was a little bigger.

There is several possible reasons for the non-recognition of certain similar pairs:

- The use of abbreviations in *α-Latin* languages:

  e.g. George **W.** Bush → جورج دبليو بوش [jwrj **dpliw** pw$].

- Translated names:

  e.g. Henri **VII** →السابع هنري [hnry **AlsAbE**].

- Some transformation rules are too rare, and they cannot be integrated into the transcoding system without risk of increasing noise.

  E.g. Karen **Kwiatkowksi**→كفايتوفسكي كارن [kArn **kfaytwfsky**]

- In some case, competing rules we ignored:

  e.g. Mi**ch**ael → مايكل [mAy**k**l]
  e.g. Pino**ch**et → بينوشيه [bynw**$**yh]

4. ALIGNING EXPERIMENTS

The next stage of experiments concerns the integration of transcoded units into the aligning process.

4.1. CORPORA PRE-TREATMENT

First, we have adapted some transcoding rules to be compatible with the XML input of our aligning system. We followed the recommendations of Buckwalter itself [15]. We replaced:

< By I (for hamza-under-alif (إ))
> By O (for hamza-on-alif (أ))
& by W (for hamza-on-waw (ؤ))

We have used the Arcade 2 corpus [3], keeping the NE's annotations on the French part of the corpus, in order to determine the EN that has to be aligned with Arabic.

Then we added the reduced forms for both French and Arabic, as additional mark-up attributes for each token in the input texts.
Here is a sample of the aligning input extracted from the French corpus:

```
<pers>
        <tok id="t3"pivot="ignasio">Ignacio</tok>
        <tok id="t4"pivot="ramont">Ramonet</tok>
</pers>
```

4.2. SENTENCE ALIGNING WITH ALINÉA

Our aligning system, Alinéa, can extract automatic alignment at two levels: sentence and word. To complete this task, Alinéa proceeds in two stages: first, it extracts aligned sentences, making successive sentences groupings that match between the source and the target (based on identical chains, cognate pairs, and lengths of sentences). Then, inside the aligned sentences, it extracts matches between equivalent tokens (simple word or multiword units, depending on how the input corpus has been tokenized).

To reduce noise in sentence alignment, we carried out a manual filtering, by eliminating areas that seemed somehow problematic. The size of the filtered corpus is displayed on Table 2. From this corpus, we manually built a reference alignment, in order to evaluate the results of Alinéa.

Number of sentences	French	Arabic
Before the manual filtering	14 179	14 087
After the manual filtering	9 021	13 076

Table 2 : Size of our corpus Arcade 2

Then, we automatically performed the alignment on two versions of the corpus, with and without transcoding annotations (which is used to identify probable cognate pairs).

The comparison between the reference and the automatic aligning allowed us to compute precision and recall for both corpus version. We got the following values (Table 3), rounded up to 1 / 100:

	Precision	Recall
With transcoding annotation	85,8 %	81,7 %
Without transcoding annotation	74,2%	71,0%

Table 3: Results of sentence alignment, with and without transcoding annotation.

These results show that adding our rules for cognate detection yields a significant improvement in aligning at sentence level, of around 11% for precision and recall. As Kraif (2001) [9] showed, resembling chains may, for some corpus, constitute a valuable indication. In the following section, we try to assess the impact of transcoding rules for word level aligning.

4.2.1. WORD ALIGNEMENT WITH ALINÉA

Aligning at word level will allow us to extract matches between NE's, and evaluate our assumption regarding the possibility of "projecting" French annotations into Arabic. To match lexical equivalents between French and Arabic, we have to create first a parameter file, which records statistics of occurrences and co-occurrences of French and Arabic tokens. Then, we use "stoplists" to take away the more frequent empty words (prepositions, articles, conjunctions…), which do not provide useful information during content unit matching.

For each aligned pair, Alinéa provides a score that measures the association strength between units, calculated on the basis of similarities, relative positions in sentences, and co-occurrence statistics. This score is a relative value; because it gives preference to some matches against others competing associations. The matching pairs are extracted using an iterative one-to-one matching algorithm, similar to the competitive linking algorithm proposed by Melamed [10].

To evaluate the impact of cognate detection using the transcoding system, we used three different parameter files, corresponding to various corpus sizes for occurrence and cooccurrence statistics computing. Usually, the bigger is the corpus, the more relevant are these statistics.

We used the Arcade 2 sub-corpus (for which we have NE annotations), the Arcade 2 full corpus (used to evaluate sentence level aligning), and an extended corpus including Arcade2 and UN parallel texts. Table 4 indicates the words number in each corpus.

	French words	Arabic words
Arcade2 sub-corpus → parameters 0	30 558	28 606
Arcade2 full corpus → parameters 1	206 736	158 699
Arcade2 corpus and UN corpus → parameters 2	383 993	411 063

Table 4: Corpus sizes for 0..2 parameter files

4.3. EVALUATION

To give a complete evaluation of the performance of our transcoding system, we compared our results with two other transliteration systems, SAWS (Scientific Arabic Writing Systems) and QALAM (an electronic morphological transliteration system) [2] which were applied on the same corpora and used during the lexical aligning stage with Alinéa.

QALAM [13] is a morphological system, in the sense that Arabic script words are transliterated according to spelling and diacritics (the marks that represent vowels in Arabic), rather than on phonetic, whereas SAWS [13] is a transliteration system traditionally used for handwriting.

To evaluate the results of NE's projecting based on word alignment, we implemented an evaluation script processing three different files: the word level alignment in cesAlign format, the Arabic corpus including NE's annotations in XML format, and the French annotated corpus in the same format (a sample is shown in Table 4).

The evaluation script implements the following algorithm: for each word in the French corpus, if the word is included in a NE tag we check:

- if there is a corresponding unit in Arabic and if this unit is itself included in a NE tag of

the same type: the match is considered as valid

- if not, the match is not considered as valid.

From these counts, we compute precision and recall.

Table 5, 6 and 7 give the results regarding anthroponyms using "Parameters 0", "Parameters 1" and "Parameters 2" cooccurrence statistics.

Anthroponyms	Transcoding	SAWS	QALAM	Without pivot
Correct couples number	642	602	562	530
Wrong couples number	226	280	321	332
Precision	**73,96%**	68,25%	63,64%	61,48%
Recall	**46,45%**	43,56%	40,66%	38,35%

Table 5: Results obtained using "Parameters 0".

Anthroponyms	Transcoding	SAWS	QALAM	Without pivot
Correct couples number	672	611	590	527
Wrong couples number	240	257	274	287
Precision	**72,31%**	70,39%	68,28%	66,6%
Recall	**45,36**%	44,21%	42,7%	41,4%

Table 6: Results obtained using "Parameters 1".

Anthroponyms	Transcoding	SAWS	QALAM	Without pivot
Correct couples number	630	597	568	545
Wrong couples number	236	258	281	307
Precision	**72,74%**	69,82%	66,90%	63,96%
Recall	**45,58%**	43,2%	41,1%	39,43%

Table 7: Results obtained by using "Parameters 2"

Table 8, 9 and 10 give the results regarding toponyms by using "Parameters 0", "Parameters 1" and "Parameters 2".

Toponyms	Transcoding	SAWS	QALAM	Without pivot
Correct couples number	546	559	545	508
Wrong couples number	652	632	643	692
Precision	45,6%	**46,93%**	45,87%	42,33%
Recall	36,91%	**37,8%**	36,84%	34,34%

Table 8: Results obtained using "Parameters 0".

Toponyms	Transcoding)	SAWS	QALAM	Without pivot
Correct couples number	581	609	605	563
Wrong couples number	622	632	589	639
Precision	48,3%	**50,91%**	50,67**%**	46,83%
Recall	39,28%	**41,17%**	40,90%	38,06%

Table 9: Results obtained using "Parameters 1".

.

Toponyms	pivot (Reduction Operation)	SAWS	QALAM	Without pivot
Correct couples number	588	611	605	586
Wrong couples number	652	628	636	673
Precision	47,41%	**49,31%**	48,75%	45,76%
Recall	39,75%	**41,31%**	40,90%	38,40%

Table10: Results obtained using "Parameters 2".

4.4. DISCUSSION

These results clearly show that the use of any kind of transliteration, for either anthroponym or toponym, improve EN aligning.

The precision and recall for toponyms remain relatively low, whatever parameters are used. They do not depend much on the transliteration system. Indeed, toponyms are likely to be translated rather than transliterated. Furthermore, the use of stoplists and the structure of our algorithm, which is limited by one-to-one matches, could generate certain problems. Here are examples of non recognized toponyms:

Côte d'ivoire →ساحل العاج [saHl alEaj].
Gan → جنوى [jnwy].

The best results for toponyms are obtained with the SAWS transliteration, but it has to be noted that there is no great difference between the various transliterated systems and the not transliterated corpus. Less toponyms appear to bear a surface resemblance, and when it is the case, very few are transliterated by the translators: there are often related words, with some divergences due to etymology. Thus, graphic comparison brings few valuable information, and more noise.

In contrast, for anthroponyms, the results are encouraging: about 72% are correctly recognized using the transcoding method, which performs better than the other transliteration schemes. Most of anthroponym pairs are very similar, and the best way to take advantage of this similarity is to use a transcoding scheme that applies on both language.

When comparing results of the various parameters, it should be noted that parameter 1 yields the best results (if we consider only cooccurrence statistics, without transliteration). We should have got better results with the extended corpus (parameter 3), which is larger and which should give more complete cooccurrence statistics: but the fact that the extended corpus is more heterogeneous, including texts from United Nation that are very different from our test corpus, may explain the loss of precision.

There is a very interesting outcome of these results: when reduction transcoding is used, it does not appear to be necessary to use a bigger corpus for cooccurrence statistic counts. The results are optimal even if the parameters are computed on a relatively small corpus (30 000 words for each language).

It should be taken into account that these results relate to word-to-word alignments, and that we have adopted a tolerant measure, as we considered as correct the fragmentary NE's (e.g. "George" instead of "George Bush").

5. CONCLUSION

We proposed an approach, based on bilingual aligning, which aims to project annotations from French language to Arabic. This type of projection, if it appears to be effective, could allow recycling of language resources through different languages: a list of NE in French could, for example, lead to the creation of a similar list in Arabic through a parallel aligned corpus. In a context of resources scarcity (lexicons, taggers, etc.) for a given language, this would allow to establish inexpensively a capital of linguistic data (annotated corpora), which represents a good start for the development of basic tools.

To illustrate this approach, we propose a very simple method dedicated to Named Entity aligning. This method is based on superficial string processing, and requires no precondition: it is a simple transcoding scheme, easy to develop, and called "graphical reduction", which is applied on both languages, and aims at making the equivalent strings closer together. Experiments show that this method allows improving Named Entity aligning when they correspond to phonetically similar units, as most anthroponyms. On our corpus, graphical reduction yielded better results than other kind of transliteration schemes, as SAWS and Qalam, because it is really designed to take advantage of similarities. The results for toponyms seem much less satisfactory, because cognate words may have diverged for etymological reasons.

Future works may focus on other language pairs using different alphabets, in order to show that graphical reduction is a generic, simple, and language independent method.

REFERENCES

[1] Arbabi, Mansur, S. M. Fischthal, V. C.Cheng, and E. Bar, "Algorithms for Arabic name transliteration,". *IBM Journal of research and Development*, 38(2):183-193, 1994.

[2] Banouni M., A. Lazrek, K. Sami, "Une translittération arabe/roman pour un e-document," i*n 5^e Colloque International sur le Document Électronique, Conférence Fédérative sur le Document*, Hammamet, Tunisi, pp. 123–137, 2002.

[3] Brown P., Lai J., Mercer R., " Aligning sentences in parallel corpora, ",*in Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, p. 169-176, 1991.

[4] Chiao Y.-C., O. Kraif, D. Laurent, T. M. H. Nguyen, N. Semmar, F. Stuck, J. Véronis, W. Zaghouani (2006-forthcoming) "Evaluation of multilingual text alignment systems: the ARCADE II project", *in Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006*, Genova, May 2006.

[5] Darwish K., D. Doermann, R. Jones,D. Oard and M. Rautiainen, "TREC-10 experiments at Maryland: CLIR and video,". *In proceedings of TREC 2001*, Gaithersburg: NIST, 2001.http://trec.nist.gov/pubs/trec10/t10_proceedings.html

[6] Debili F.,Sammoud E. "Appariements de phrases de texts bilingues Français-Anglais et Français-Arabes ".*in Actes de COLING-92*,Nantes,pp.528-524,1992.

[7] Gale W., Church K., " A program for aligning sentences in bilingual corpora, "*in Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkley, California, p. 177-184, 1991.

[8] Kraif O., B.Chen, "Combining clues for lexical level aligning using the Null hypothesis approach," *in Proceedings of Coling 2004*, Geneva, pp. 1261-1264, August 2004.

[9] Kraif O, "Exploitation des cognats dans les systèmes d'alignement bi-textuel: architecture et évaluation," *TAL N°. 42 vol. 3, ATALA*, Paris, pp. 833-867, 2001.

[10] Melamed D., "Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons," *in Proceedings of the 3rd Workshop on Very Large Corpora (WVLC3)*, Boston, MA. pp. 184-198,1995.

[11] Meng H. M., W.-K. Lo, B. Chen and T. Tang. 2001. "Generate Phonetic Cognates to Handle Name Entities in English-Chinese Cross-Language Spoken Document Retrieval, " *In Proceedings of the Automatic Speech Recognition and Understanding Workshop* (ASRU), Italy, pp.311-314.

[12] Pouliquen B., R. Steinberger, C. Ignat, I. Temnikova, A. Widiger, W. Zaghouani, J. Žižka,"Multilingual person name recognition and transliteration," *Journal CORELA - Cognition, Representation, Language*. Numeros speciaux, Le traitement lexicographique des noms propres Poitiers, France, CERLICO. ISSN 1638 -5748, vol. 3/3, no. 2, pp. 115-123, 2005.

[13] Stalls, B. and Knight, K., "Translation Names and Technical Terms in Arabic Text," *in Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, Montreal, Quebéc, 1998.

[14] Site of QALAM: A convention for morphological arabiclatin-Arabic transliteration. (1985).Heddaya,Abdelsalam URL:http://eserver.org/langs/qalam.txt.

[15] Site of the association of: Qamus llc. (2002). http://www.qamus.org/transliteration.htm, visite on 23/09/2006.

[16] Wan S., Cornelia Maria V., " Automati English-Chinese name transliteration for development of multilingual resources," *in Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, Montreal, Quebec, Canada ,p.p 1352–1356. 1998.

[17] William A.G., K. W. Church, "A program for aligning sentences in bilingual corpora," *in Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, CA, pp. 177-184, 1991