Using Cognates in a French - Romanian Lexical Alignment System: A Comparative Study

Mirabela Navlea
Linguistique, Langues, Parole (LiLPa)
Université de Strasbourg
22, rue René Descartes
BP, 80010, 67084 Strasbourg cedex
navlea@unistra.fr

Amalia Todiraşcu
Linguistique, Langues, Parole (LiLPa)
Université de Strasbourg
22, rue René Descartes
BP, 80010, 67084 Strasbourg cedex
todiras@unistra.fr

Abstract

This paper describes a hybrid French - Romanian cognate identification module. This module is used by a lexical alignment system. Our cognate identification method uses lemmatized, tagged and sentence-aligned parallel corpora. This method combines statistical techniques, linguistic information (lemmas, POS tags) and orthographic adjustments. We evaluate our cognate identification module and we compare it to other methods using pure statistical techniques. Thus, we study the impact of the used linguistic information and the orthographic adjustments on the results of the cognate identification module and on cognate alignment. Our method obtains the best results in comparison with the other implemented statistical methods.

1 Introduction

We present a new French - Romanian cognate identification module, integrated into a lexical alignment system using French - Romanian parallel law corpora.

We define cognates as translation equivalents having an identical form or sharing orthographic or phonetic similarities (common etymology, borrowings). Cognates are very frequent between close languages such as French and Romanian, two Latin languages with a rich morphology. So, they represent important lexical cues in a French - Romanian lexical alignment system.

Few linguistic resources and tools for Romanian (dictionaries, parallel corpora, MT systems) are currently available. Some lexically aligned corpora or lexical alignment tools (Tufiş *et al.*, 2005) are available for Romanian - English or Romanian - German (Vertan and Gavrilă, 2010). Most of the cognate identification modules used by these systems are purely statistical. As far as we know, no cognate identification method is available for French and Romanian.

Cognate identification is a difficult task due to the high orthographic similarities between bilingual pairs of words having different meanings. Inkpen *et al.* (2005) develop classifiers for French and English cognates based on several dictionaries and manually built lists of cognates. Inkpen *et al.* (2005) distinguish between:

- cognates (*liste* (FR) - *list* (EN));
- false friends (*blesser* ('to injure') (FR) - *bless* (EN));
- partial cognates (*facteur* (FR) - *factor* or *mailman* (EN));
- genetic cognates (*chef* (FR) - *head* (EN));
- unrelated pairs of words (*glace* (FR) - *ice* (EN) and *glace* (FR) - *chair* (EN)).

Our cognate detection method identifies cognates, partial and genetic cognates. This method is used especially to improve a French - Romanian lexical alignment system. So, we aim to obtain a high precision of our cognate identification method. Thus, we eliminate false friends and unrelated pairs of words combining statistical techniques and linguistic information (lemmas, POS tags). We use a lemmatized, tagged and sentence-aligned parallel corpus. Unlike Inkpen *et al.* (2005), we do not use other external resources (dictionaries, lists of cognates).

To detect cognates from parallel corpora, several approaches exploit the orthographic similarity between two words of a bilingual pair. An efficient method is the 4-gram method (Simard *et al.*, 1992). This method considers two words as cognates if their length is greater than or equal to 4 and at least their first 4 characters are common. Other methods exploit Dice's coefficient (Adam-

Proceedings of Recent Advances in Natural Language Processing, pages 247–253,
Hissar, Bulgaria, 12-14 September 2011.

son and Boreham, 1974) or a variant of this coefficient (Brew and McKelvie, 1996). This measure computes the ratio between the number of common character bigrams of the two words and the total number of two word bigrams. Also, some methods use the Longest Common Subsequence Ratio (LCSR) (Melamed, 1999; Kraif, 1999). LCSR is computed as the ratio between the length of the longest common substring of ordered (and not necessarily contiguous) characters and the length of the longest word. Thus, two words are considered as cognates if LCSR value is greater than or equal to a given threshold. Similarly, other methods compute the distance between two words, which represents the minimum number of substitutions, insertions and deletions used to transform one word into another (Wagner and Fischer, 1974). These methods use exclusevly statistical techniques and they are language independent.

On the other hand, other methods use the phonetic distance between two words belonging to a bilingual pair (Oakes, 2000). Kondrak (2009) identifies three characteristics of cognates: recurrent sound correspondences, phonetic similarity and semantic affinity.

Thus, our method exploits orthographic and phonetic similarities between French - Romanian cognates. We combine n-grams methods with linguistic information (lemmas, POS tags) and several input data disambiguation strategies (computing cognates' frequencies, iterative extraction of the most reliable cognates and their deletion from the input data). Our method needs no external resources (bilingual dictionaries), so it could easily be extended to other Romance languages. We aim to obtain a high accuracy of our method to be integrated in a lexical alignment system. We evaluate our method and we compare it with pure statistical methods to study the influence of used linguistic information on the final results and on cognate alignment.

In the next section, we present the parallel corpora used for our experiments. In section 3, we present the lexical alignment method. We also describe our cognate identification module in section 4. We present the evaluation of our method and a comparison with other methods in section 5. Our conclusions and further work figure in section 6.

2 The Parallel Corpus

In our experiments, we use a legal parallel corpus (*DGT-TM*[1]*)* based on the *Acquis Communautaire* corpus. This multilingual corpus is available in 22 official languages of EU member states. It is composed of laws adopted by EU member states since 1950. *DGT-TM* contains 9,953,360 tokens in French and 9,142,291 tokens in Romanian.

We use a test corpus of 1,000 1:1 aligned complete sentences (starting with a capital letter and finishing with a punctuation sign). The length of each sentence has at most 80 words. This test corpus contains 33,036 tokens in French and 28,645 in Romanian.

We use the *TTL*[2] tagger available for Romanian (Ion, 2007) and for French (Todiraşcu *et al.*, 2011) (as Web service[3]). Thus, the parallel corpus is tokenized, lemmatized, tagged and annotated at chunk level.

The tagger uses the set of morpho-syntactic descriptors (MSD) proposed by the Multext Project[4] for French (Ide and Véronis, 1994) and for Romanian (Tufiş and Barbu, 1997). In the Figure 1, we present an example of TTL's output: *lemma* attribute represents the lemmas of lexical units, *ana* attribute provides morpho-syntactic information and *chunk* attribute marks nominal and prepositional phrases.

```
<seg lang="FR"><s id="ttlfr.3">
<w lemma="voir" ana="Vmps-s">vu</w>
<w lemma="le" ana="Da-fs"
chunk="Np#1">la</w>
<w lemma="proposition" ana="Ncfs"
chunk="Np#1">proposition</w>
<w lemma="de" ana="Spd"
chunk="Pp#1">de</w>
<w lemma="le" ana="Da-fs"
chunk="Pp#1,Np#2">la</w>
<w lemma="commission" ana="Ncfs"
chunk="Pp#1,Np#2">Commission
</w>
<c>;</c>
</s></seg>
```

Figure 1 TTL's output for French (in XCES format)

[1] http://langtech.jrc.it/DGT-TM.html

[2] Tokenizing, Tagging and Lemmatizing free running texts

[3] https://weblicht.sfs.uni-tuebingen.de/

[4] http://aune.lpl.univ-aix.FR/projects/multext/

3 Lexical Alignment Method

The cognate identification module is integrated in a French - Romanian lexical alignment system (see Figure 2).

In our lexical alignment method, we first use GIZA++ (Och and Ney, 2003) implementing IBM models (Brown *et al*., 1993). These models build word-based alignments from aligned sentences. Indeed, each source word has zero, one or more translation equivalents in the target language. As these models do not provide many-to-many alignments, we also use some heuristics (Koehn *et al*., 2003; Tufiş *et al*., 2005) to detect phrase-based alignments such as chunks: nominal, adjectival, verbal, adverbial or prepositional phrases.

In our experiments, we use the lemmatized, tagged and annotated parallel corpus described in section 2. Thus, we use lemmas and morpho-syntactic properties to improve the lexical alignment. Lemmas are followed by the two first characters of morpho-syntactic tag. This operation morphologically disambiguates the lemmas (Tufiş *et al*., 2005). For example, the same French lemma *change* (=*exchange, modify*) can be a common noun or a verb: *change_Nc* vs. *change_Vm*. This disambiguation procedure improves the GIZA++ system's performance.

We realize bidirectional alignments (FR - RO and RO - FR) with GIZA++, and we intersect them (Koehn *et al*., 2003) to select common alignments.

To improve the word alignment results, we add an external list of cognates to the list of the translation equivalents extracted by GIZA++. This list of cognates is built from parallel corpora by our own method (described in the next section).

Also, to complete word alignments, we use a French - Romanian dictionary of verbo-nominal collocations (Todiraşcu *et al*., 2008). They represent multiword expressions, composed of words related by lexico-syntactic relations (Todiraşcu *et al*., 2008). The dictionary contains the most frequent verbo-nominal collocations extracted from legal corpora.

To augment the recall of the lexical alignment method, we apply a set of linguistically-motivated heuristic rules (Tufiş *et al*., 2005):

a) we define some POS affinity classes (a noun might be translated by a noun, a verb or an adjective);
b) we align content-words such as nouns, adjectives, verbs, and adverbs, according to the POS affinity classes;
c) we align chunks containing translation equivalents aligned in a previous step;
d) we align elements belonging to chunks by linguistic heuristics. We develop a language dependent module applying 27 morpho-syntactic contextual heuristic rules (Navlea and Todiraşcu, 2010). These rules are defined according to morpho-syntactic differences between French and Romanian.

The architecture of the lexical alignment system is presented in the Figure 2.

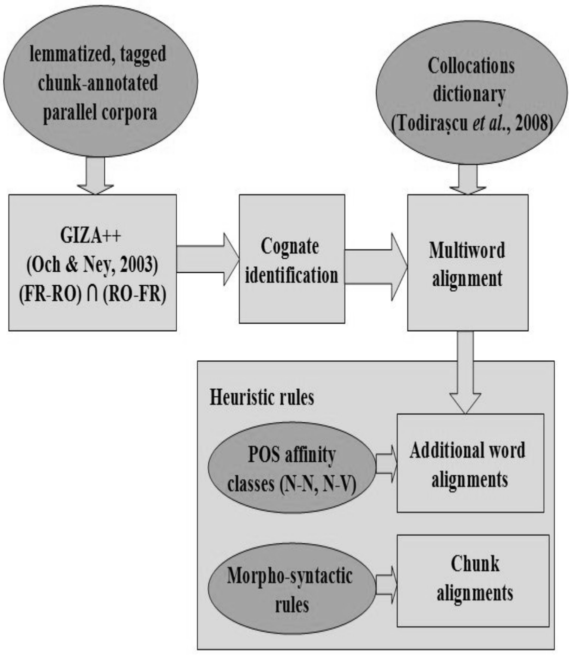
lemmatized, tagged chunk-annotated parallel corpora
Collocations dictionary (Todiraşcu et al., 2008)
GIZA++ (Och & Ney, 2003) (FR-RO) ∩ (RO-FR)
Cognate identification
Multiword alignment
Heuristic rules
POS affinity classes (N-N, N-V)
Additional word alignments
Morpho-syntactic rules
Chunk alignments

Figure 2 Lexical alignment system architecture

4 Cognate Identification Module

In our hybrid cognate identification method, we use the legal parallel corpus described in section 2. This corpus is tokenized, lemmatized, tagged, and sentence-aligned.

Thus, we consider as cognates bilingual word pairs respecting the linguistic conditions below:

1) their lemmas are translation equivalents in two parallel sentences;
2) they have identical lemmas or have orthographic or phonetic similarities between lemmas;
3) they are content-words (nouns, verbs, adverbs, etc.) having the same POS tag or belonging to the same POS affinity class. We filter out short words such as prepositions and conjunctions to limit noisy output. We also detect short cognates such as *il* 'he' vs. *el* (personal pronoun), *cas* 'case' vs. *caz* (nouns). We

avoid ambiguous pairs such as *lui* ‘him’ (personal pronoun) (FR) vs. *lui* ‘s' (possessive determiner) (RO), *ce* 'this' (demonstrative determiner) (FR) vs. *ce* 'that' (relative pronoun) (RO).

To detect orthographic and phonetic similarities between cognates, we look at the beginning of the words and we ignore their endings.

We classify the French - Romanian cognates detected in the studied parallel corpus (at the orthographic or phonetic level), in several categories:

1) cross-lingual invariants (numbers, certain acronyms and abbreviations, punctuation signs);
2) identical cognates (*document* ‘document’ vs. *document*);
3) similar cognates:
   a) 4-grams (Simard *et al*., 1992); The first 4 characters of lemmas are identical. The length of these lemmas is greater than or equal to 4 (*__auto__rité* vs. *__auto__ritate* 'authority').
   b) 3-grams; The first 3 characters of lemmas are identical and the length of the lemmas is greater than or equal to 3 (*__act__e* vs**.** ***act*** 'paper').
   c) 8-bigrams; Lemmas have a common sequence of characters among the first 8 bigrams. At least one character of each bigram is common to both words. This condition allows the jump of a non identical character (*__souscri__r__e__* vs. *__su__b__scrie__* 'submit'). This method applies only to long lemmas (length greater than 7).
   d) 4-bigrams; Lemmas have a common sequence of characters among the 4 first bigrams. This method applies for long lemmas (length greater than 7) (*h__omolo__gué* vs. *__omol__ogat* 'homologated') but also for short lemmas (length less than or equal to 7) (*__group__e* vs. ***grup*** 'group').

We iteratively extract cognates by identified categories. In addition, we use a set of orthographic adjustments and some input data disambiguation strategies. We compute frequency for ambiguous candidates (the same source lemma occurs with several target candidates) and we keep the most frequent candidate. At each iteration, we delete reliable considered cognates from the input data.

We start by applying a set of empirically established orthographic adjustments between French - Romanian lemmas, such as: diacritic removal, phonetic mappings detection, etc. (see Table 1).

Levels of orthographic adjustments	**French**	**Romanian**	**Examples FR - RO**
diacritics	x	x	d**é**p**ô**t - d**e**p**o**zit
double contiguous letters	x	x	ra**pp**ort **-** ra**p**ort
consonant groups	ph th dh cch ck cq ch ch	f [f] t [t] d [d] c [k] c [k] c [k] ş [ʃ] c [k]	**ph**ase **-** **f**ază mé**th**ode - me**t**odă a**dh**érent - a**d**erent ba**cch**ante **-** ba**c**antă sto**ck**age - sto**c**are gre**cq**ue - gre**c** fi**ch**e - fi**ş**ă **ch**apitre **-** **c**apitol

q	q (final)
	qu(+i) (medial)
	qu(+e) (medial)
	qu(+a)
	que (final)
		c [k]
		c [k]
		c [k]
		c(+a) [k]
		c [k]
			cin**q** - cin**c**i
			é**q**uilibre - e**c**hilibru
			mar**q**uer - mar**c**a
			qualité **-** **c**alitate
			prati**q**ue - practi**c**ă

intervocalic s	v + s + v	v + z + v	pré**s**ent **-** pre**z**ent
w	w	v	**w**agon **-** **v**agon
y	y	i	**y**aourt **-** **i**aurt

Table 1 French - Romanian cognate orthographic adjustments

While French uses an etymological writing and Romanian generally has a phonetic writing, we identify phonetic correspondences between lemmas. Then, we make some orthographic adjustments from French to Romanian. For example, cognates *stockage* 'stock' (FR) vs. *stocare* (RO) become *stocage* (FR) vs. *stocare* (RO). In this example, the French consonant group *ck* [k] become *c* [k] (as in Romanian). We also make adjustments in the ambiguous cases, by replacing with both variants (*ch* ([ʃ] or [k])): ***fiche*** vs. ***fişă*** 'sheet'; ***chapitre*** vs. ***capitol*** 'chapter'.

We aim to improve the precision of our method. Thus, we iteratively extract cognates by identified categories from the surest ones to less sure candidates (see Table 2).

To decrease the noise of the cognate identification method, we apply two supplementary strategies. We filter out ambiguous cognate candidates (*autorité - autoritate|autorizare*), by computing their frequencies in the corpus. In this case, we keep the most frequent candidate pair. This strategy is very effective to augment the precision of the results, but it might decrease the recall in certain cases. Indeed, there are cases where French - Romanian cognates have one form in French, but two various forms in Romanian (*spécification* 'specification' vs. *specificare* or *specificaţie*). We recover these pairs by using regular expressions based on specific lemma endings (*ion* (fr) vs. *re|ţie* (ro)).

Then, we delete the reliable cognate pairs (high precision) from the input data at the end of the extraction step. This step helps us to disambiguate the input data. For example, the identical cognates *transport* vs. *transport* 'transportation', obtained in a previous extraction step and deleted from the input data, eliminate the occurrence of candidate *transport* vs. *tranzit* as 4-grams cognate, in a next extraction step.

We apply the same method for cognates having POS affinity (N-V; N-ADJ). We keep only 4-grams cognates, due to the significant decrease of the precision for the other categories 3 (b, c, d).

Extraction steps by category of cognates	**Content-words / Same POS**	**Frequency**	**Deletion from the input data**	**Precision (%)**
1 : cross lingual invariants			x	100
2 : identical cognates	x		x	100
3 : 4-grams (lemmas' length >= 4) ;	x	x	x	99.05
4 : 3-grams (lemmas' length >=3) ;	x	x	x	93.13
5 : 8-bigrams (long lemmas, lemmas' length >7)	x		x	95.24
6 : 4-bigrams (long lemmas, lemmas' length > 7)	x			75
7 : 4-bigrams (short lemmas, lemmas' length =< 7)	x	x		65.63

Table 2 Precision of cognate extraction steps

5 Evaluation and Methods' Comparison

We evaluated our cognate identification module against a list of cognates initially built from the test corpus, containing 2,034 pairs of cognates.

In addition, we also compared the results of our method with the results provided by pure statistical methods (see Table 3). These methods are the following:

a) thresholding the Longest Common Subsequence Ratio (LCSR) for two words of a bilingual pair; This measure computes the ratio between the longest common subsequence of characters of two words and the length of the longest word. We empirically establish the threshold of 0.68.

$$LCSR(w1, w2) = \frac{length(common\_substring(w1, w2))}{\max(length(w1), length(w2))}$$

b) thresholding DICE's coefficient; We empirically establish the threshold of 0.62.

$$DICE(w1, w2) = \frac{2 * number\_common\_bigrams}{total\_number\_bigrams(w1, w2)}$$

c) 4-grams ; Two words are considered as cognates if they have at least 4 characters and their first 4 characters are identical.

We implemented these methods using orthographically adjusted parallel corpus (see Table 1). Moreover, we evaluate 4-grams method on the initial parallel corpus and on the orthographically adjusted parallel corpus to study the impact of orthographic adjustments step on the quality of the results.

These methods generally apply for words having at least 4 letters in order to decrease the noise of the results. Cognates are searched in aligned parallel sentences. Word characters are almost parallel (***rembourser*** vs. ***rambursare*** 'refund').

Methods	P (%)	R (%)	F (%)
LCSR	44.13	58.95	50.47
DICE	56.47	60.91	58.61
4-grams	91.55	72.42	80.87
Our method	**94.78**	**89.18**	**91.89**

Table 3 Evaluation and methods' comparison; P=Precision; R=Recall; F=F-measure

Our method extracted 1,814 correct cognates from 1,914 provided candidates. The method obtains the best scores (precision=94.78% ; recall=89.18% ; f-measure=91.89%), in comparison with the other implemented methods. The 4-grams method obtains a high precision (90.85%), but a low recall (47.84%). Orthographic adjustments step improves significantly the recall of 4-grams method with 24.58% (see Table 4). This result is due to the specific properties of the law parallel corpus. Indeed, many Romanian terms were borrowed from French and these terms present high orthographic similarities.

Methods	P (%)	R (%)	F (%)
4-grams - Adjustments	90.85	**47.84**	62.68
4-grams + Adjustments	91.55	**72.42**	80.87

Table 4 Evaluation of the 4-grams method before and after orthographic adjustments step

However, our method extracts some ambiguous candidates such as *numéro 'number' - nume 'name'*, *compléter* 'complete' - *compune* 'compose'. Some of these errors were avoided by keeping the most frequent candidate in the studied corpus. So, the remaining errors mainly concern hapax candidates.

Also, some cognates were not extracted: *heure - oră* 'hour', *semaine - săptămână* 'week', *lieu - loc* 'place'. These errors concern cognates sharing very few orthographic similarities.

The lowest scores are obtained by the LCSR method (f-measure=50.47%), followed by the DICE's coefficient (f-measure=58.61%). These general methods provide a high noise due to the important orthographic similarities between the words having different meanings. Their results might be improved by combining statistical techniques with linguistic information such as POS affinity or by combining several association scores.

As we mentioned, the output of the cognate identification module is exploited by a French - Romanian lexical alignment system (based on GIZA++) described in section 3. We compared the set of cognates provided by GIZA++ with our results to study their impact on cognate alignment. GIZA++ extracted 1,532 cognates representing a recall of 75.32% (see Table 5). Our cognate identification module significantly improved the recall with 13.86%.

Systems	Number of extracted cognates	Number of total cognates	Recall (%)
GIZA++	1,532	2,034	75.32
Our method	1,814		89.18

Table 5 Improvement of our method's recall

6 Conclusions and Further Work

We present a French - Romanian cognate identification module required by a lexical alignment system. Our method combines statistical techniques and linguistic filters to extract cognates from lemmatized, tagged and sentence-aligned parallel corpus. The use of the linguistic information and the orthographic adjustments significantly improves the results compared with pure statistical methods. However, these results are dependent of the studied languages, of the corpus domain and of the data volume. We need more experiments using other corpora from other domains to be able to generalize. Our system should be improved to detect false friends by using external resources.

Cognate identification module will be integrated in a French - Romanian lexical alignment system. This system is part of a larger project aiming to develop a factored phrase-based statistical machine translation system for French and Romanian.

References

George W. Adamson and Jillian Boreham. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles, *Information Storage and Retrieval*, 10(7-8):253-260.

Chris Brew and David McKelvie. 1996. Word-pair ex-traction for lexicography, in *Proceedings of International Conference on New Methods in Natural Language Processing*, Bilkent, Turkey, 45-55.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, 19(2):263-312.

Nancy Ide and Jean Véronis. 1994. Multext (multilingual tools and corpora), in *Proceedings of the 15th International Conference on Computational Linguistics, CoLing 1994*, Kyoto, pp. 90-96.

Diana Inkpen, Oana Frunză, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English, RANLP-2005, Bulgaria, Sept. 2005, p. 251-257.

Radu Ion. 2007. *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română, Ph.D. Thesis*, Romanian Academy, Bucharest, May 2007, 148 pp.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation, in *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL 2003*, Edmonton, May-June 2003, pp. 48-54.

Grzegorz Kondrak. 2009. Identification of Cognates and Recurrent Sound Correspondences in Word Lists, in *Traitement Automatique des Langues (TAL)*, 50(2) :201-235.

Olivier Kraif. 1999. Identification des cognats et alignement bi-textuel : une étude empirique, dans *Actes de la 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, TALN 99*, Cargèse, 12-17 juillet 1999, 205-214.

Dan I. Melamed. 1999. Bitext Maps and Alignment via Pattern Recognition, in *Computational Linguistics*, 25(1):107-130.

Mirabela Navlea and Amalia Todiraşcu. 2010. Linguistic Resources for Factored Phrase-Based Statistical Machine Translation Systems, in *Proceedings of the Workshop on Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages, 7th International Conference on Language Resources and Evaluation*, Malta, Valletta, May 2010, pp. 41-48.

Michael P. Oakes. 2000. Computer Estimation of Vocabulary in Protolanguage from Word Lists in Four Daughter Languages, in *Journal of Quantitative Linguistics*, 7(3):233-243.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, in *Computational Linguistics*, 29(1):19-51.

Michel Simard, George Foster, and Pierre Isabelle. 1992. Using cognates to align sentences, in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, pp. 67-81.

Amalia Todiraşcu, Ulrich Heid, Dan Ştefănescu, Dan Tufiş, Christopher Gledhill, Marion Weller, and François Rousselot. 2008. Vers un dictionnaire de collocations multilingue, in *Cahiers de Linguistique*, 33(1) :161-186, Louvain, août 2008.

Amalia Todiraşcu, Radu Ion, Mirabela Navlea, and Laurence Longo. 2011. French text preprocessing with TTL, in *Proceedings of the Romanian Academy, Series A*, Volume 12, Number 2/2011, pp. 151-158, Bucharest, Romania, June 2011, Romanian Academy Publishing House. ISSN 1454-9069.

Dan Tufiş and Ana Maria Barbu. 1997. A Reversible and Reusable Morpho-Lexical Description of Romanian, in Dan Tufiş and Poul Andersen (eds.), *Recent Advances in Romanian Language Technology*, pp. 83-93, Editura Academiei Române, Bucureşti, 1997. ISBN 973-27-0626-0.

Dan Tufiş, Radu Ion, Alexandru Ceauşu, and Dan Ştefănescu. 2005. Combined Aligners, in *Proceedings of the Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 107-110, Ann Arbor, USA, Association for Computational Linguistics. ISBN 978-973-703-208-9.

Cristina Vertan and Monica Gavrilă. 2010. Multilingual applications for rich morphology language pairs, a case study on German Romanian, in Dan Tufiş and Corina Forăscu (eds.): *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Romanian Academy Publishing House, Bucharest, pp. 448-460, ISBN 978-973-27-1972-5.

Robert A. Wagner and Michael J. Fischer. 1974. The String-to-String Correction Problem, *Journal of the ACM*, 21(1):168-173.