Cognate Identification for a French - Romanian Lexical Alignment System: Empirical 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 cognate identification method, used by a lexical alignment system for French and Romanian. We combine statistical techniques and linguistic information to extract cognates from lemmatized, tagged and sentence-aligned parallel corpora. We evaluate the cognate identification model and we compare it to other methods using pure statistical techniques. We show that the use of linguistic information in the cognate identification system improves significantly the results.

1 Introduction

We present a new cognate identification module required for a French - Romanian lexical alignment system. This system is used for French - Romanian law corpora. Cognates are translation equivalents presenting orthographic or phonetic similarities (common etymology, borrowings, and calques). They represent very important elements in a lexical alignment system for legal texts for two reasons:
- French and Romanian are two close Romance languages with a rich morphology;
- Romanian language borrowed and calqued legal terminology from French. So, cognates are very useful to identify bilingual legal terminology from parallel corpora, while we do not use any external terminological resources for these languages.

Cognate identification is one of the main steps applied for lexical alignment for MT systems. If we have several efficient tools for several European languages, few lexically aligned corpora or lexical alignment tools (Tuflă et al., 2005) are available for Romanian - English or Romanian - German (Vertan and Gavrilă, 2010). In general, few linguistic resources and tools for Romanian (dictionaries, parallel corpora, terminological data bases, MT systems) are currently available. Some MT systems use resources for the English - Romanian language pair (Marcu and Munteanu, 2005; Irimia, 2008; Ceauşa, 2009). Other MT systems develop resources for German - Romanian (Gavrilă, 2009; Vertan and Gavrilă, 2010) or for French - Romanian (Navlea and Todiraşcu, 2010). Most of the cognate identification modules used by these systems were purely statistical. No cognate identification method is available for the studied languages.

Cognate identification is a difficult problem, especially to detect false friends. Inkpen et al. (2005) classify bilingual words pairs in several categories such as:
- cognates (reconnaissance (FR) - reconnaissance (EN));
- false friends (blessre (‘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)).

In our method, we rather identify cognates and partial cognates to improve lexical alignment. Thus, we aim to obtain a high precision of our method and to eliminate some false friends using statistical techniques and linguistic information.

To identify cognates from parallel corpora, several approaches exploit the orthographic similarity between two words of a bilingual pair. A simple method is the 4-grams method (Simard et al., 1992). This method considers that two words are cognates if they contain at least 4 characters and...
at least their first 4 characters are identical. Other
methods exploit association scores as Dice’s
coefficient (Adamson 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 words
bigrams. Also, two words are considered as cog-
nates if the ratio between the length of the max-
imum common substring of ordered (and not nec-
essarily contiguous) characters and the length of
the longest word is greater than or equal to a cer-
tain empirically determined threshold (Melamed,
1999; Kraif, 1999). Similarly, other methods
compute the distance between two words, that
represent the minimum number of substitutions,
insertions and deletions used to transform one
word into another (Wagner and Fischer, 1974).
On the other hand, other methods compute the
phonetic distance between two words belonging
to a bilingual pair (Oakes, 2000). Kondrak
(2009) proposes methods identifying three char-
acteristics of cognates: recurrent sound corre-
pondences, phonetic similarity and semantic af-
finity.
We present a French - Romanian cognate identi-
fication module. We combine statistical tech-
niques and linguistic information (lemmas, POS
tags) to improve the results of the cognate identi-
fication method. We compare it with other
methods using exclusively statistical techniques.
The cognate identification system is integrated
into a lexical alignment system.
In the next section, we present our lexical align-
ment method. We present our parallel corpora
and the tools used to preprocess our parallel cor-
pora, in section 3. In section 4, we describe our
cognate identification method. We present the
results’ evaluation in section 5. Our conclusions
and further works figure in section 6.

2 The Lexical Alignment Module

The output of the cognate identification module
is exploited by a French - Romanian lexical align-
ment system.
Our lexical alignment system combines statistical
methods and linguistic heuristics. We use GI-
ZA++ (Och and Ney, 2000, 2003) implementing
IBM models (Brown et al., 1993). These models
realize word-based alignments. Indeed, each
source word has zero, one or more translation
equivalents in the target language, computed
from aligned sentences. Due to the fact that these
models do not provide many-to-many align-
ments, we also use some heuristics (Koehn et al.,
2003; Tufiş et al., 2005) in order to detect
phrase-based alignments such as chunks: nomin-
al, adjectival, verbal, adverbial or prepositional
phrases.
We use lemmatized, tagged and annotated at
chunk level parallel corpora. These corpora are
described in details in the next section.
To improve the lexical alignment, we use lem-
mas and morpho-syntactic properties. We pre-
pare the corpus in the input format required by
GIZA++, providing also the lemma and the two
first characters of the morpho-syntactic tag. This
operation morphologically disambiguates the
lemmas (Tufiş et al., 2005). For example, the
same French lemma traité (=treaty, treated) can
be a common noun or a participial adjective:
traité_Nc vs. traité_Af. This disambiguation pro-
cEDURE improves the GIZA++ system’s perfor-
mance.
In order to obtain high accuracy of the lexical
alignment, we realize bidirectional alignments
(FR-RO and RO-FR) with GIZA++, and then we
intersect them (Koehn et al., 2003). This heuristic
only selects sure links, because these align-
ments are detected in the two lexical alignment
process directions.
To obtain sure word alignments, we also use a
set of automatically identified cognates. Indeed,
we filter the list of translation equivalents ob-
tained by alignment intersection, using a list of
cognates. To extract cognates from parallel cor-
pora, we developed a method adapted to the stu-
died languages. This method combines statistical
techniques and linguistic information. The me-
thod is presented in section 4.
We obtain sure word alignments using multiword
expressions such as collocations. They represent
polylexical expressions whose words are related
by lexico-syntactic relations (Todirașcu et al.,
2008). We use a multilingual dictionary of ver-
bo-nominal collocations (Todirașcu et al., 2008)
to align them. This dictionary is available for
French, Romanian and German. The dictionary is
completed by data extracted from legal texts and
it contains the most frequent verbo-nominal col-
llocations from this domain. The external re-
source is used to align this class of collocations
(for legal corpora), but it do not resolve the
alignment problems of other classes (noun +
noun, adverb + adjective, etc.).
Finally, we apply a set of linguistically motivated
heuristic rules (Tufiş et al., 2005) in order to
augment the recall of the lexical alignment me-
thod:
i. we define some POS affinity classes (a noun can be translated by a noun, a verb or an adjective);
ii. we align content-words such as nouns, adjectives, verbs, and adverbs, according to the POS affinity classes;
iii. we align chunks containing translation equivalents aligned in a previous step;
iv. we align elements belonging to chunks by linguistic heuristics; At this level, we developed a supplementary module depending on the two studied languages. This module uses a set of 27 morpho-syntactic contextual heuristics rules. These rules are defined according to morpho-syntactic differences between French and Romanian (Navlea and Todiraşcu, 2010). For example, in Romanian relative clause, the direct object is simultaneously realized by the relative pronoun care ‘that’ (preceded by the pe preposition) and by the personal pronoun il, -l, o, ii, -i, le. In French, it is expressed by que relative pronoun. Thus, we define a morpho-syntactic heuristic rule to align the supplementary elements from the source and from the target language (see Figure 1). The rule aligns que with the sequence pe care (accusative) le.

Figure 1 Example of lexical alignment using morpho-syntactic heuristics rules (the case of relative clause)

| French                      | Romanian                  |
|-----------------------------|---------------------------|
| les                         | problemele                |
| problèmes                   | pe                        |
| que                         | care                      |
| créerait                    | le-                       |
| publication                 | genera                    |
| generation, publication     | publicarea                |

These corpora are based on the Acquis Communautaire multilingual corpus available in 22 official languages of EU. It is composed of laws adopted by EU member states and EU candidates since 1950. For our project, we use a subset of 228,174 pairs of 1-1 aligned sentences from the JRC-Acquis, selected from the common documents available in French and in Romanian. We also use a subset of 490,962 pairs of 1-1 aligned sentences extracted from the DGT-TM.

As the JRC-Acquis and the DGT-TM are legal corpora, we built other multilingual corpora for other domains (politics, aviation). Thus, we manually selected French - Romanian available texts from several websites according to several criteria: availability of the bilingual texts, reliability of the sources, translation quality, and domain. The used corpora are described in the Table 1:

| Corpora Source                | Number of words / French | Number of words / Romanian |
|-------------------------------|--------------------------|----------------------------|
| JRC-Acquis                    | 5,828,169                | 5,357,017                  |
| DGT-TM                        | 9,953,360                | 9,142,291                  |
| European Parliament’s website | 137,422                  | 126,366                    |
| European Commission’s website | 200,590                  | 185,476                    |
| Romanian airplane companies’ websites | 33,757             | 29,596                     |

Table 1 French - Romanian parallel corpora

We preprocess our corpora by applying the TTL\(^2\) tagger (Ion, 2007). This tagger is available for French and for Romanian as Web service. Thus, the parallel corpora are tokenized, lemmatized, POS tagged and annotated at chunk level. TTL uses the set of morpho-syntactic descriptors (MSD) proposed by the Multext Project\(^3\) for French (Ide and Véronis, 1994) and for Romanian (Tuşif and Barbu, 1997). TTL’s results are available in XCES format (see Figure 2).

---

1 http://langtech.jrc.it/DGT-TM.html
2 Tokenizing, Tagging and Lemmatizing free running texts
3 http://aune.lpl.univ-aix.fr/projects/multext/
Figure 2 TTL’s output for French

In the example of the Figure 2, lemma attribute represents the lemmas of lexical units, ana attribute provides morpho-syntactic information and chunk attribute marks nominal and prepositional phrases. We exploit these linguistic information in order to adapt lexical alignment algorithm for French and for Romanian. Thus, we study the influence of linguistic information to the quality of the lexical alignment.

4 Cognate Identification

We did our lexical alignment and cognate identification experiments on a legal parallel corpus extracted from the Acquis Communautaire. We automatically selected 1,000 1:1 aligned complete sentences (starting with a capital letter and finishing with a punctuation sign). Each selected sentence has no more than 80 words. This corpus contains 33,036 tokens in French and 28,645 tokens in Romanian. We tokenized, lemmatized and tagged our corpus as mentioned in the previous section. Thus, to extract French - Romanian cognates from lemmatized, tagged and sentence-aligned parallel corpus, we exploit linguistic information: lemmas, POS tags. In addition, we use orthographic and phonetic similarities between cognates. To detect such similarities, we focus rather on the beginning of the words and we ignore their endings. First, we use n-gram methods (Simard et al., 1992), where n=4 or n=3. Second, we compare ordered sequences of bigrams (an ordered pair of characters). Then, we apply some data input disambiguation strategies, such as:

- we iteratively extract sure cognates, such as invariant strings (abbreviations, numbers etc.) or similar strings (3- and 4-grams). At each iteration, we delete them from the input data;
- we use cognate pairs frequencies in the studied corpus.

We consider as cognates the words belonging to a bilingual pair simultaneously respecting the following linguistic conditions:

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 showing POS affinities. So, we filter out short words as prepositions and conjunctions to limit noisy output. Thus, we do not generally restrict lemmas length. We also detect short cognates as il 'he' vs. el (personal pronouns), cas 'case' vs caz (nouns). We avoid ambiguous pairs such as lui 'him' (personal pronoun) (FR) vs. lui (possessive determiner) (RO), ce 'this' (demonstrative determiner) (FR) vs. ce 'that' (relative pronoun) (RO).

We classify French - Romanian cognates (detected in the studied parallel corpus) in several categories:

1) cross-lingual invariants (numbers, certain acronyms and abbreviations). In this category, we also consider punctuation signs;
2) identical cognates (civil vs civil);
3) similar cognates (at the orthographic or phonetic level) :
   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 (autorité 'authority' vs. autoritate).
   b) 3-grams; The first 3 characters of lemmas are identical and the length of the lemmas is greater than or equal to 3 (mars 'March' vs. marie);
   c) 8-bigrams; Lemmas have a common sequence of characters (eventually discontinuous) among the first 8 bigrams. At least one character of each bigram is common to the two words. This condition allows the jump of a non identical character (fonctionne-
ment ‘fonction’ vs. functionare). This applies only to long lemmas, with the length greater than 7.

d) 4-bigrams; Lemmas have a common sequence of characters (eventually discontiguous) among the 4 first bigrams: rembourser ‘refund’ vs. rambursa; objet ‘object’ vs. object. This applies to long lemmas (length > 7) but also to short lemmas (length less than or equal to 7).

Our method mainly follows three stages. In the first place, we apply a set of empirically established orthographic adjustments between French - Romanian lemmas, such as: remove diacritics, detect phonetic mappings, etc. (see Table 2). As French uses an etymological writing and Romanian has a phonetic writing, we identify phonetic correspondences between lemmas. We make some orthographic adjustments from French to Romanian. For example, cognates phase ‘phase’ (FR) vs. fază (RO) become faze (FR) vs. faza (RO)). In this example, we make two adjustments: the French consonant group ph [f] become f (as in Romanian) and the French intervocalic s [z] become z (as in Romanian). We also make adjustments in the ambiguous cases, by replacing with both variants (ch ([ʃ] or [k]): machine vs. mașină ‘car’, chlorure ‘chlorure’ vs. clorură.

| Levels of orthographic adjustments | French | Romanian | Examples |
|-----------------------------------|--------|----------|----------|
| Diacritics                        | x      | x        | dépôt - depozit |
| double contiguous letters         | x      | x        | rapport - rapport |
| consonant groups                  |        |          | ph - f [f] |
|                                  |        |          | th - t [t] |
|                                  |        |          | dh - d [d] |
|                                  |        |          | cch - c [k] |
|                                  |        |          | ck - c [k] |
|                                  |        |          | cq - c [k] |
|                                  |        |          | ch - s [ʃ] |
|                                  |        |          | ch - c [k] |
|                                  |        |          | q (final) - c [k] |
|                                  | qu(+i) (medial) - c [k] |
|                                  | qu(+e) (medial) - c [k] |
|                                  | qu(+a) - c(+a) [k] |
|                                  | que (final) - c [k] |
| intervocalic s                    | v + s + v | v + z + v | présent - prezent |
| w                                 | w      | v        | wagon - wagon |
| y                                 | y      | i        | yaourt - iaur |

Table 2 French - Romanian cognates orthographic adjustments

Secondly, we apply seven cognate extraction steps (see Table 3). To extract cognates from parallel corpora, we aim to improve the precision of our method. Thus, we extract cognates by applying the categories 1 - 3 (a-d) (see Table 3). Moreover, in order to decrease the noise of cognate identification method, we apply two supplementary strategies. We filter out ambiguous cognate candidates (a same source lemma occurs with several target candidates), by computing their frequencies in the corpus. Thus, we keep the most frequent candidate pair. This strategy is very effective to augment the results precision, but it might decrease the recall in certain cases. Indeed, there are cases when French - Romanian cognates have one form in French, but two various forms in Romanian (information ‘information’ vs. informație or informare; manifestație or manifestare). We recover these pairs by using regular expressions based on specific lemma ending (ion (FR) vs. fiere (RO)).

Then, we delete the reliable cognate pairs (high precision) from the input data at the end of the extraction step. Thus, we disambiguate the data input. For example, the identical cognates transport vs. transport, obtained in a previous extraction step and deleted from the input data, eliminate the occurrence of candidate transport vs. transport as 4-grams cognate, in a next extraction step.
These strategies allow us to increase the precision of our method. We give below some examples of correct extracted cognates: *autorité* (FR) - *autoritate* (RO); *disposition* (FR) - *dispoziție* (RO); *directive* (FR) - *directivă* (RO). We also eliminate some false friends: *autorité* (FR) - *autorizare* (RO); *layout* (FR) - *dispozitiv* (RO); *direction* (FR) - *directivă* (RO).

5 Evaluation

We evaluate our method on a parallel corpus of 1,000 sentences described in the previous section. We compare the results with another two methods (see Table 4):

- a) the method exclusively based on 4-grams;
- b) a combination of the 4-gram approach and the orthographic adjustments.

| Extraction steps by category of cognates | F | Deletion from input data | P (%) |
|------------------------------------------|---|--------------------------|-------|
| 1 : cross lingual invariants             |   | x                        | 100   |
| 2 : identical cognates                   |   | x                        | 100   |
| 3 : 4-grams (lemmas’ length >= 4)        |   | x x                      | 99.05 |
| 4 : 3-grams (lemmas’ length >= 3)        |   | x x                      | 93.13 |
| 5 : 8-bigrams (long lemmas, lemmas’ length >7) |   | x                        | 95.24 |
| 6 : 4-bigrams (long lemmas, lemmas’ length > 7) |   | 75                       |
| 7 : 4-bigrams (short lemmas, lemmas’ length <= 7) |   | x                        | 65.63 |

Table 3 Precision of cognates’ extraction steps; F=Frequency; P=Precision

However, our system extracts some false candidates, such as: *numéristere* (FR) - *nume* (RO); *consommation* (FR) - *considerare* (RO); *compléter* (RO) - *compune* (FR). We apply the same method for cognates having POS affinity (N-V; N-ADJ). We keep only 4-gram cognates, due to a significant decrease of the precision for the other categories (3-grams, 8-bigrams and 4-bigrams).

Finally, we recover initial cognates lemmas for both languages.

6 Conclusions and Further Work

We present here a cognate identification module for two morphologically rich languages such as French and Romanian. Cognates are very important elements used by a lexical alignment system. Thus, we aim to obtain high precision and recall of our cognate identification method by combining statistical techniques and linguistic information. We show that an orthographic adjustment step between French - Romanian lemmas bilingual pairs and linguistic filters improve significantly module’s performance.

The cognate identification method is integrated into a French-Romanian lexical alignment module. The alignment module is part of a larger
project aiming to develop a French - Romanian factored phrase-based statistical machine translation system.

References

Adamson, George W., 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.

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

Brown, Peter F., 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.

Ceaușu, Alexandru. 2009. Tehnici de traducere automată și aplicabilitatea lor limbii române ca limbă sursă, Ph.D. Thesis, Romanian Academy, Bucharest, April 2009, 123 pp.

Gavrilă, Monica. 2009. SMT experiments for Romanian and German using JRC-Acquis, in Proceedings of the Workshop on Multilingual resources, technologies and evaluation for central and Eastern European languages, 7th Recent Advances in Natural Language Processing (RANLP), 17 September 2009, Borovets, Bulgaria, pp. 14-18.

Ide, Nancy, and Jean Véronis. 1994. Multext (multilingual tools and corpora), in Proceedings of the 15th International Conference on Computational Linguistics, CoLing 1994, Kyoto, August 5-9, pp. 90-96.

Inkpen, Diana, Fruzan Oana, and Kondrak Grzegorz. 2005. Automatic Identification of Cognates and False Friends in French and English, in Proceedings of Recent Advances in Natural Language Processing, RANLP-2005, Bulgaria, Sept. 2005, p.251-257.

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

Irimia, Elena. 2008. Experimenter de Traducere Automată Bazată pe Exemple, in Proceedings of Workshop Rezurse Lingvistice Românești și Instrumente pentru Prelucrarea Limbi Române, Iasi, 19-21 November 2008, pp. 131-140.

Koehn, Philipp, Franz Josef 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.

Koehn, Philipp, and Hieu Hoang. 2007. Factored translation models, in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, June 2007, 868-876.

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

Kraif, Olivier. 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.

Marcu, Daniel, and Dragoș Ş. Munteanu. 2005. Statistical Machine Translation: An English-Romanian Experiment, in 7th International Summer School EUROLAN 2005, July 25 - August 6, Cluj-Napoca, Romania.

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

Navlea, Mirabela, 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 (LREC 2010), Malta, Valletta, May 2010, pp. 41-48.

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

Och, Franz Josef, and Hermann Ney. 2000. Improved Statistical Alignment Models, in Proceedings of the 38th Conference of the Association for Computational Linguistics, ACL 2000, Hong Kong, pp. 440-447.

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

Simard, Michel, 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.

Steinberger, Ralph, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiš, and Dăniel Varga. 2006. The JRC-Acquis: A Multilin-
Todirașcu, Amalia, Ulrich Heid, Dan Ștefănescu, Dan Tufiş, Christopher Gledhill, Marion Weller, and François Roussilot. 2008. Vers un dictionnaire de collocations multilingue, in Cahiers de Linguistique, 33(1) :161-186, Louvain, août 2008.

Tufiş, Dan, 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, București, 1997. ISBN 973-27-0626-0.

Tufiş, Dan, Radu Ion, Alexandru Ceașcu, and Dan Ștefănescu. 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.

Vertan, Cristina, and Monica Gavrilă. 2010. Multilingual applications for rich morphology language pairs, a case study on German Romanian language pairs, in Dan Tufiş and Corina Forâscu (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.

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