SEMI-AUTOMATIC ALIGNMENT OF OLD ROMANIAN WORDS USING LEXICONS

MARIA MORUZ[1], ADRIAN IFTENE[2], ALEX MORUZ[2, 3], DAN CRISTEA[2, 3]

[1] *Centre of Biblical-Philological Studies "Al. I. Cuza" University, Iasi*
[2] *Faculty of Computer Science, "Al. I. Cuza" University, Iasi*
[3] *Institute for Computer Science, Romanian Academy, Iasi Branch*

mhusarciuc@gmail.com
{adiftene,mmoruz,dcristea}@info.uaic.ro

Abstract

This paper discusses an approach for the semi-automatic alignment of old Romanian words taken from three 17th century translations of the Bible. The alignment is first carried out at the verse level, then by means of lexical matching using Levenshtein distance, and then further refined by using a series of heuristics for matching the remaining words; to compensate for synonymy and high lexical variance, we have employed a lexicon. The biblical variants used are the 1688 Bible (in Romanian *Biblia de la București*), Manuscript 45 and Manuscript 4389, and the modern translation of the Bible given in the *Monumenta Linguae Dacoromanorum* series.

1. Introduction

Until recently, natural language processing has been primarily concerned with the modeling and analysis of the modern version of languages, in a synchronic manner. However, there has been a recent increase in the interest for digitizing and analyzing older versions of languages, as shown in (Roselli del Turco, 2010).

In the case of the old Romanian language, this interest has recently been sparked by the availability of digitized old Romanian texts, which were made available in the electronic version of the Romanian Thesaurus Dictionary (Cristea et al. 2009) and within the "Monument Linguae Dacoromanorum" project (Haja and Munteanu, 2010). As a result, a number of applications that make use of these resources have been proposed. One such application, for example, envisages the creation of a diachronical morphology for the Romanian language of 17th century by exploiting a large number of citations from eDTLR (Simionescu et al. 2012).

This paper is concerned with the creation of a lexical similarity equivalence database for the old Romanian language used in three 17th century translations of the Bible. The database is mainly concerned with semantic similarity at the word level, but we also intend to extend this similarity to expressions and idioms. The database is obtained by means of lexical alignment of parallel versions of texts, given that the three translations available are already aligned at the chapter and verse level.

The structure of the paper is the following: section 2 describes the source for the digitized version of the Bible translations, the "Monumenta Linguae Dacoromanorum" project, and the format of the digital version; section 3 presents the methods we have

employed for aligning the translation versions; section 4 describes the results obtained during the alignment process and section 5 presents conclusions and discusses future work.

2. The "Monumenta Linguae Dacoromanorum" project

The "Monumenta Linguae Dacoromanurum – 1688 Bible" project (Haja and Munteanu, 2010) is an international philological project started in 1988 by professor Paul Miron of the Albert Lüdwigs University of Freiburg im Breisgau, Germany, in partnership with the "Al. I. Cuza" University of Iasi, Romania. The purpose of this project is the creation of a philological edition, together with studies and linguistic commentaries, facsimiles and indices for words and variants, of the three contemporary and complete Romanian translations of the Bible from the 17th century: the 1688 Bible (also known as the Șerban Cantacuzino Bible or the Bucharest Bible) printed in Cyrillic characters, Manuscript 45 from the Romanian Academy library in Cluj, which contains the "revised Milescu version" from the 17th century, and the Romanian Academy library Manuscript 4389, which contains the "Daniil Panoneanul" version, also from the 17th century.

Every volume in the series has the following components:

1. The texts of the translation variants, arranged on 5 columns as follows:
    a. The facsimile of the original printed in 1688 (first column)
    b. The phonetic and interpretative transcript of the 1688 Bible (second column)
    c. The phonetic and interpretative transcript of Manuscript 45 (third column)
    d. The phonetic and interpretative transcript of Manuscript 4389 (fourth column)
    e. The modern translation of *Septuagint*, used as an auxiliary tool for the understanding of the old translation variants (fifth column)
2. Philological notes
3. Biblical-philological and linguistic commentaries
4. An exhaustive index of words and variants which contains all of the lexical occurrences in the 1688 Bible
5. Facsimiles of the manuscripts.

Until the time of writing, the "Al. I. Cuza" University Publishing House has published 9 of the projected 25 volumes: I. Genesis (1988), II. Exodus (1991), III. Leviticus (1993), IV. Numeri (1994), V. Deuteronomium (1997), XI. Liber Psalmorum (2003), VI. Iosue. Iudicum. Ruth (2005), VII. Regum I, Regum II (2008), IX. Paralipomenon I, Paralipomenon II (2011).

Starting with *Regum I, Regum II* (Andriescu et al., 2008), the volumes are available in electronic format (Haja et al., 2008), (Patras et al., 2008), together with an exhaustive index for the printed text (the 1688 Bible), ordered by lemma; attached to each lexical

occurrence are the morphologic analysis, the translations into German and French, and the first attested use of the term.

3. Aligning old Romanian words

The alignment of the parallel translations of the Bible was partly inspired by the notion of alignment between parallel texts in different languages (Moore, 2002). This idea was successfully used in the creation of a lexical similarity equivalence database consisting of French and Romanian multi-word expressions aligned according to semantic similarity (Husarciuc, 2008). In order to create a similar taxonomy for old Romanian texts, we first need the world level alignment from which to extract the expression alignment.

In the case of parallel translations in the same language, the problem of aligning parallel texts is simpler, as it is to be expected that at least some of the words are common, even accounting for differences induced by differences in time and region. Because of this we have adopted a bootstrapping based approach that makes use of Levenshtein distances between words and a set of heuristics and resources to improve the alignment result.

3.1. Alignment algorithm

The input for the alignment system consists of pairs of pre-aligned verses extracted from the sources described in section 2. A general outline of the alignment process is given below:

1. Starting from the pre-aligned verse pair, we attempt to match words on the basis of lexical similarity, extracted by means of exact match and Levenshtein distance. Since there is the possibility that a verse contains the same word more than once, the alignment score is weighed by the distance between the positions of the two candidates in the verses. Usually, the words extracted at this stage are proper names and words that have little temporal or regional variance (e.g. “bătrînețe”, “Solomón” etc.), and can be counted upon to be semantically equivalent in a large majority of cases;
2. Given the previously extracted alignments, we consider the aligned words pivots and attempt to align further words on this basis. To this extent we consider that unaligned words that are found next to already aligned words and are lexically similar have a high probability of being aligned, and so their matching score is boosted;
3. In order to account for the semantic similarity of words that are not lexically similar, we have used an automatically generated lexical similarity equivalence database which has been validated by hand. It contains the word pairs that have been aligned by means of heuristics (such as those in step 2), and is manually validated in order to avoid incorrect matches. The score assigned to the alignment obtained by using the taxonomy is computed on the basis of match frequency in the aligned corpus, as a specific word can have multiple semantic equivalents (e.g. “domn - boiarin” and “căpetenie - boiarin”). The taxonomy is described in greater detail in subsection 3.2;

4. The process is repeated from step 2 until no new alignments can be carried out.

3.2. A lexical database for semantic similarity

Given the fact that no suitable lexicon for the old Romanian language exists, the most accessible manner for solving the issue of semantic similarity is the creation of a taxonomy that describes this relation. While solving this issue, such a database is an accomplishment in itself, as this is a valuable resource for the study of the Romanian language of the 17th century.

The database is populated by adding those words that have lexical similarity scores below an empirically determined threshold, but have been aligned at step 2 in the algorithm given above. These alignments are then manually validated by human annotators, and then a bootstrapping approach is applied, in order to add further pairs to the database. Since the texts on which alignment was carried out are not available in lemmatized form for the manuscripts and the modern translation, the words in the database are given in inflected forms, which greatly reduce the number of cases where a given relation can be inferred. A relation consists of the semantically similar words and an attached confidence score, which is assigned on the basis of frequency in the corpus (we only take into account the corpus represented by the validated alignments). Examples of such relations are given in Fig. 1 below:

fiu [is] fecior [score] 1
jîrtăvnicul [is] altariul [score] 1
ţiitoarea [is] posadnica [score] 1
astruca [is] îngropa [score] 1
boiêri [is] domni [score] 0.6
boiêri [is] căpetenii [score] 0.4
den sîmbătă în sîmbătă [is] în toate sîmbetele [score] 0.7
împărăţi [is] stătu împărat [score] 0.9

Figure 1: Entries in the Semantic Similarity Taxonomy

As can be seen in the examples above, a similarity relation usually holds between two words (one-to-one relations), but we have also allowed for the possibility of one-to-many and many-to-many relations, in order to model similarity for multi-word expressions.

4. Results and discussions

The algorithm described in section 3 was tested on the texts available in (Andriescu et al., 2008), which contains the books “Regum I” and “Regum II”, and the obtained results are described in this section. The reason for our using this particular volume was twofold: it is the first volume in the series to have an electronic index of semantically disambiguated words (Haja et al., 2008), (Patras et al., 2008) and it was available in a structured electronic format that allowed quick access to the aligned verses.

Also, these particular books are less difficult to align, since verses usually contain large numbers of proper nouns, which are easily matched; large numbers of high confidence

matches give high confidence in heuristic matches, and thus the lexical similarity equivalence database is more quickly populated. Once the database is already established and contains large numbers of relations, the alignment of verses that are not lexically similar becomes easier.

Table 1 below gives the results of the application of our algorithm on these books. The alignment was carried out on pairs of translation versions: the 1688 Bible represents version 1, manuscript 45 is version 2, manuscript 4389 is version 3 and version 4 is the modern translation (in order to determine the similarity of the modern translation to a 17^{th} century one, we have decided to also align manuscript 4389 to the modern translation). The numbers in the table represent the number of word alignments that have been determined.

Table 1: Alignment results for "Regum I" and "Regum II"

Match type	**1 – 2**	**2 - 3**	**3 – 4**
Lexical identity	40191	25389	14798
Levenshtein distance	5946	6913	4973
Lexical database	569	443	453
Unmatched words	6600	19055	28866

As can be seen in Table 1, the 1688 Bible and manuscript 45 are very similar, which is to be expected given the fact that one is based on the other. Manuscript 45 and manuscript 4389 are significantly different at the lexical level, as shown by the lower number of lexical matches, while manuscript 4389 and the modern translation have very few lexical similarities, which is to be expected due to the evolution of language. The low number of semantic taxonomy matches is due to the fact that, at the time of testing, the taxonomy contained approximately 100 pairs; given the fact that these pairs contain inflected words, their scope is limited, thus resulting in a low number of matches.

Examples of alignment cases which support the steps of the algorithm in section 3 are given below.

Regum I, 1, 10

B1688*: Și ea, amărîtă la suflet, și s-au rugat către Domnul și plîngînd au plîns*

Ms. 45*: Și ea, amărîtă la suflet, și s-au rugat către Domnul și plîngîndu au plînsu.*

Example 1: Lexical similarity based alignment

Example 1, given above, is a case of near perfect lexical match. This similarity is mainly due to the fact that B1688 is largely based on Ms. 45. For this particular case, step 1 of the algorithm in section 3 solves all of the alignments.

Regum I, 15, 5

B1688 *: Și veni Saul pînă la cetățile lui Amalic și* ***strejui*** *la pîrîu.*

Ms. 45*: Și veni Saul pănă la cetățile lui Amalic și* ***să aleșuiră*** *în părău.*

Example 2: Enriching the semantic similarity taxonomy using heuristics

In the case of example 2, all of the words in the verses are aligned at step 1, with the exception of those words which are highlighted. According to step 2 of the algorithm, since the highlighted words are surrounded by already aligned words, and since there

are no other words that are not aligned, the similarity score of the word pair is boosted, and alignment is found. The pair is inserted in the database candidate pool, awaiting manual validation.

Regum II, 24, 18

Ms. 45*: Și veni Gad către David întru dzua acêea și-i dzise lui: "Suie-te și pune Domnului jirtăvnic întru ariia lui Orna, ievuseului!".*

Ms. 4389*: Și veni Gad într-acea zi la David și-i zise: "Suie-te și pune altar lui Dumnezeu în arătura lui Iornei ievuseul".*

Example 3: Semantic similarity based alignment

In those cases where the semantic similarity taxonomy contains a word pair, alignment is carried out directly on the basis of that relation. Such is the case in example 3, where two alignments are carried out by this method. It is worth noting that without the taxonomy, the alignment would be very difficult to predict, given the low lexical similarity of the words in the vicinity. Table 2 shows the improvement brought by the lexical similarity equivalence database to the alignment process (between 1 and 2):

Table 2: Alignment results with and without the lexical similarity equivalence database

Run type	Exact	Ontology	Levenshtein	Not aligned
With database	40191	569	5946	6600
Without database	40191	0	6042	7073

Alignment is made more difficult by a series of translation inconsistencies within the variants. Such an inconsistency is given by the fact that parts of the original text in some verses are missing and are given in endnotes or not given at all. Such is the case in example 4, where part of the text in manuscript 45 is missing and is given in an endnote.

Regum I 20, 30

B1688*: Și să mînie cu urgie Saul pre Ionathan foarte și zise lui: "Fecior de fêtele cêle ce mergu de bunăvoie, au nu știu că părtaș ești tu cu* ***fiiul lui Iesei*** *întru rușinea ta și întru rușinea descoperirii maicii tale?*

Ms 45 *: Și să mînie cu urgie Saul preste Ionathan foarte și-i dzise lui: "Fiu a fêtelor ce mărgu de bunăvoie, au nu știu că părtaș ești tu[10] întru rușinea ta și întru rușinea dăscoperirei maicii tale?*

+note *: Marginal note in another hand: "****fiiului lui Iese****".*

Example 4: Missing text in verses

Another type of inconsistency which greatly hinders automatic word alignment is the use of proper and common nouns for denoting the same concepts (this occurs most commonly in the case of names of peoples). In extreme cases, the word forms in the translation variants are very different, as is the case in example 5 below. The verse from manuscript 45 is also missing some text, which does not exist even in the endnotes.

Regum I 15, 6*:*

B 1688*: Și zise Saul către Chineu: "ferêște-te și te abate den mijlocul* ***Amalichitului****, ca să nu te adaogă împreună cu el; și tu ai făcut milă cu toți fiii lui Israil cînd să suia ei den Eghipet". Și să abătu Chineul den mijlocul lui Amalic.*

***Ms. 45**: Și dzise Saul cătră Chineu: "ferêşte-te // și te abate den mijlocul **amalichitului**, ca să nu te adaog împreună cu el. Și tu ai făcut milă cu toți fiii lui Israil cînd să suia ei den Eghiptu".*

***Ms. 4389**: Și zise Saul lui Chinei: "Pasă și te dă în laturi den mijlocul **ammalitênilor** și nu te apropiia de dînșii să te concenesc depreună cu dînșii, că tu ai făcut milă cu feciorii lui Israil cînd ieșiia den Eghipet". Și se dêde Chinei într-o lature den mijlocul ammalitênilor.*

Example 5: Proper and common nouns for the same concept

5. Conclusions and future work

This paper proposes an algorithm for aligning translation variants of old Romanian texts by means of lexical and semantic similarity. It also proposes a method for extending existing semantic similarity taxonomies by using a set of heuristics. The results obtained prove the potential of our proposed method, as increases of the lexical similarity equivalence database are directly correlated to increases in the number of alignments.

The proposed alignment can be used for the extraction of inflection variants for the old Romanian language, which is useful for the creation of an old Romanian grammar; also, the alignment to the modern version of the translation allows for the observation of the evolution of words and expressions.

As future work we intend to apply our algorithm to a new volume in the MLD series, *Paralipomenon I, II*, in order to further test and enhance our algorithm. Also, for future alignments, we will use the B1688 version as a pivot, aligning the other variants only to it, mainly because most of the lexical semantic disambiguation has been carried out on this version. Also, the generic Levenshtein string matching algorithm should be modified in order to accommodate a series of linguistic phenomena, such as ignoring the final "u" in some words (e.g. "plînsu" vs. "plîns") or aligning "dz" to "z" ("zise" vs. "dzise").

Acknowledgements. This work was partly funded by the "Al. I. Cuza" University of Iasi and the Sector Operational Program for Human Resources Development through the project ―Development of the innovation capacity and increasing of the research impact through post-doctoral programs POSDRU/89/1.5/S/49944 and the METANET4U – Enhancing the European Linguistic Infrastructure project.

References

Andriescu, A., Miron, P., Haja, G. (coordinators) (2008). Monumenta linguae Dacoromanorum. Biblia 1688. Pars VI. Regum I, Regum II. *Iași, "Alexandru Ioan Cuza" University Publising House*, 560 p. + DVD. Authors: Tamara Adoamnei, Mădălina Andronic, Mioara Dragomir, Gabriela Haja, Elsa Lüder, Paul Miron, Alexandra Moraru, Mihai Moraru, Adrian Muraru, Veronica Olariu, Elena Tamba Dănilă. Scientific consultant: Eugen Munteanu. Electronic format on DVD by Vlad-Sebastian Patraș.

Cristea, D., Răschip, M., Moruz, A. (2009). Steps in Building the Electronic Version of a Thesaurus Dictionary of the Romanian Language. *Buletinul Institutului Politehnic din Iasi*. Sectia: Matematica. Mecanica Teoretica. Fizica, 1244-7863.

Haja, G., Munteanu, E. (2010). Monumenta linguae Dacoromanorum. 1688 Bible Project. *In Clarin, Newsletter of Clarin Project*, 8, 4-5 http://www.clarin.eu/files/cnl08_web.pdf.

Haja, G., Dănilă, E., Clim, M. R., Patraș, V. (2008). Contribuții la informatizarea cercetării filologice românești: Biblia 1688 și eDTLR. *Simpozionul internațional Distorsionări în comunicarea lingvistică, literară și etnofolclorică românească și contextul european*, Iași, 25-28 septembrie.

Husarciuc, M. (2009). Echivalarea în limba română a unităților frazeologice infinitivale din limba franceză. *În Lucrările Atelierului Resurse lingvistice și instrumente pentru prelucrarea limbii române* (Iași, 19-21 noiembrie 2008), Editura Universității "Alexandru Ioan Cuza" Iași, 115-124.

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. *In "Lecture Notes In Computer Science"*, 2499 - Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, on Machine Translation: From Research to Real Users, Springer-Verlag, London, ISBN: 3-540-44282-0, 135-144.

Patraș, V. S., Pavel, G., Haja, G. (2008). Resurse lingvistice în format electronic. Biblia 1688. Regi I, Regi II – probleme, soluții. *În volumul Resurse lingvistice și instrumente pentru prelucrarea limbii române*, editori Ionuț Cristian Pistol, Dan Cristea, Dan Tufiș, Iași, Editura Universității "Alexandru Ioan Cuza", 51-60.

Roselli del Turco, R. (2010). Filologia digitale: ragioni, problemi, prospettive di una disciplina. *III Incontro di Filologia Digitale*, Verona, 3-5 marzo.

Simionescu, R., Cristea, D., Haja, G., Minuț, A. M. (2012). Inferarea unei morfologii diacronice folosind eDTLR, to appear.