PROCEEDINGS
OF THE 9TH INTERNATIONAL CONFERENCE "LINGUISTIC RESOURCES AND TOOLS FOR PROCESSING THE ROMANIAN LANGUAGE"
16-17 MAY 2013

Editors

Elena Mitocariu

Mihai Alex Moruz

Dan Cristea

Dan Tufiș

Marius Clim

Organisers

Faculty of Computer Science
"Alexandru Ioan Cuza" University of Iași

Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy, Bucharest

Institute for Computer Science
Romanian Academy, Iași

The publication of this volume was supported by
the Faculty for Computer Science,
"Alexandru Ioan Cuza" University of Iași

ISSN 1843-911X

PROGRAM COMMITTEE

Corneliu Burileanu, Faculty of Electronics', Telecommunications and Information Technology, University of Bucharest and Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Constantin Ciubotaru, Institute of Mathematics and Computer Science, Academy of Science, Chișinău
Mihaela Colhon, Informatics Department, Faculty of Exact Science, University of Craiova
Dan Cristea, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi and Institute for Computer Science, Romanian Academy, Iași branch
Nicolae Curteanu, Institute for Computer Science, Romanian Academy, Iași branch
Cristina Florescu, Institute of Romanian Philology "Al. Philippide", Romanian Academy, Iași branch
Corina Forăscu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași and Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Gabriela Haja, Institute of Romanian Philology "Al. Philippide" of Iași, Romanian Academy
Adrian Iftene, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Diana Zaiu Inkpen, School of Information Technology Engineering, University of Ottawa
Radu Ion, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Cătălina Mărănduc, Institute of Linguistics "Iorgu Iordan – Al. Rosetti", Romanian Academy, Bucharest
Rada Mihalcea, Computer Science and Engineering, University of North Texas
Vivi Năstase, School of Information Technology Engineering, University of Ottawa
Ionuț Pistol, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Dan Ștefănescu, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Elena Isabelle Tamba, Institute of Romanian Philology "Al. Philippide", Romanian Academy, Iași branch
Horia-Nicolai Teodorescu, Institute for Computer Science, Romanian Academy, Iași branch and "Gheorghe Asachi" Technical University of Iași
Amalia Todirașcu, Department d'informatique, Universite de Strasbourg
Diana Trandabăț, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași and Institute for Computer Science, Romanian Academy, Iași branch
Dan Tufiș, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Cristina Vertan, Research Group "Computerphilology" (UHH), University Hamburg
Adriana Vlad, Faculty of Electronics, Telecommunications and Information Technology, University of Bucharest and Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Marius Zbancioc, Institute for Computer Science, Romanian Academy, Iași branch

ORGANIZING COMMITTEE

Dan Cristea, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași and Institute for Computer Science, Romanian Academy, Iași branch
Sabina Deliu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Corina Forăscu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași and Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Lucian Gâdioi, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Daniela Gîfu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Gabriela Haja, Institute of Romanian Philology "Al. Philippide", Romanian Academy, Iași branch
Elena Mitocariu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Alex Moruz, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași and Institute for Computer Science, Romanian Academy, Iași branch
Mădălin Ionel Pătrașcu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași and Institute of Romanian Philology "Al. Philippide", Romanian Academy, Iași branch
Ionuț Pistol, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Liviu Andrei Scutelnicu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași and Institute for Computer Science, Romanian Academy, Iași branch
Radu Simionescu, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Dan Tufiș, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Elena Isabelle Tamba, Institute of Romanian Philology "Al. Philippide", Romanian Academy, Iași branch

TABLE OF CONTENTS

FOREWORD

The series of events organised by the Consortium of Informatisation for the Romanian Language (ConsILR) has reached this year its 9th edition. With a history that goes back to 2001, the ConsILR series of events evolved in these 12 years of existence, by attracting more and more interest from linguists and computational linguists, but also from researchers of the humanities, PhD students and master students in Computational Linguistics, all with a major interest in the study of the Romanian language from a computational perspective. The series of events started in the format of a workshop and was transformed in 2010 into a conference, in order to reach an international visibility, being addressed to researchers working on Romanian language also from outside Romania. This year event was organised in Miclăușeni, in the old and romantic, recently rehabilitated Sturdza castle, which, we believe, has all prerogatives to create a perfect atmosphere for concentration and brainstorming, inviting for dialogues and debates.

The organisers of the Conference *Linguistic Resources And Tools For Processing The Romanian Language*, as in previous years, have been the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iași and two institutes of the Romanian Academy: the Research Institute for Artificial Intelligence "Mihai Drăgănescu", Bucharest, and the Institute for Computer Science of the Iași branch. The organisers were pleased to accept also a satellite workshop organised by the Institute for Romanian Philology "Alexandru Philippide", belonging to the Iași branch of the Academy. The workshop *Romanian Academic Lexicography. Challenges of Going Computational* is meant to show the progresses made in the computational approaches to lexicography in the research institutes of the Academy that developed the Thesaurus Dictionary of the Romanian Language.

In the period from the previous Conference, an event of major importance for the field of Language Technology has happened in Europe: in September last year, the series of White Papers *Languages in the Digital Age* was published as bilingual editions (in English and each out of 30 European languages) by the META-NET consortium. Each bilingual edition[1] includes comprehensive descriptions of the linguistic features of one of the European languages and tables showing comparative positioning of languages from the point of view of the existing resources and the technological development. In these tables Romanian is placed in a rather privileged position (compared against the majority of European languages) with respect to automatic translation, then near the majority of the other languages as regards the text analysis and the acquisition of resources for text and speech, but still on a tail position in the domain of speech processing. This comparative study shows with clarity not only that there is still a lot to be done in all domains of Romanian Language Technology but that we have only made the first steps towards the big science and for shaking the hands with the big industry dedicated to this domain. Another very important document published by the META Technology Council, with the contributions from more than 200 experts worldwide, is

[1] The English-Romanian edition: Diana Trandabăţ, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, Dan Tufiş (2012) The Romanian Language in the Digital Age / Limba română în era digitală, in White Paper Series, Eds. Georg Rehm and Hans Uszkoreit, Berlin, Springer, ISBN 978-3-642-30702-7, 87 p. can be accessed online at <http://www.meta-net.eu/whitepapers/e-book/romanian.pdf>.

"The Strategic Research Agenda for Multilingual Europe 2020"[2] . This comprehensive analysis outlines the major action directions and research priorities towards attaining the goal of a Knowledge Society without language barriers.

A serious warning, stressing that if serious research efforts will not be made 21 of the 30 European languages face "digital extinction", was addressed in a synchronised manner all over Europe (by press releases, TV and radio interviews) on September 26th 2012, a date celebrated as *The European Day of Languages*. By marking this day, the Council of Europe recognises the importance of fostering and developing the rich linguistic and cultural heritage of our continent.

Unfortunately, the alarm seems not to be perceived with the same intensity by the stakeholders deciding the finances of Europe: a likely reduction of the proposed European budget allocated to research for Horizon 2020 (the period 2014-2020, which includes also the Eight Framework Programme) has been announced and created major concerns within the research community of Europe. If this will be the case, the Language Technology domain will certainly not thrive in the next 7 years and the threat for digital extinction of some of the languages in Europe could become a sad reality. Romanian is certainly one of them, digital extinction meaning being less and less used in the internet (because of the lack of stable technologies able to translate it, to summarise it, or to support its automatic interpretation, like text mining, parsing, crawling, etc.). As the internet has become the principal means of communication nowadays, the danger could indeed be very serious, and an explosion of foreign influence from the better "informatized" languages could be manifested in the spoken Romanian as well, sooner than we all expect. The young speakers are the most intensive users of internet, they having a big influence in the trendy evolution of the colloquial language. It is also very well known that text and speech technologies cannot exist without computational resources, since the resources are the ground from which the linguistic technological development takes in its sap. Resources and technology go hand in hand and this is the very reason why this Conference exists.

Observing this axiom, the volume, including 16 papers, is structured in 3 parts: Language Resources, Text Processing and Speech Processing. We decided to structure the content of the volume according to the three mentioned topics and not by the event they were submitted to (the Conference or the Workshop). Each paper has been reviewed by at least two members of the Programme Committee and, in accordance with international practice, the accepted papers were transferred again to authors for final corrections and answers to the reviewers' comments. The volume does not include the presentations for which we received only abstracts, although we have recommended some of them for direct presentation (as part of the Workshop), and in one or two cases we accepted essay-like formats of the papers.

As in other editions, the complete program of the Conference and audio-video recordings of the talks can be consulted online (at http://consilr.info.uaic.ro/2013/), thanks to MEDIAEC – the Multimedia Laboratory of the "Alexandru Ioan Cuza University".

Iași, București, May 2013 The editors

[2] Georg Rehm, Hans Uszkoreit (eds): Strategic Reseach Agenda of Multilingual Europe 2020, Springer, December 2012.

CHAPTER 1
LANGUAGE RESOURCES

EXTRACTING LEXICAL DICTIONARIES FROM COMPARABLE CORPORA IN WHAT CONDITIONS DOES IT WORTH?

IRIMIA ELENA

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest, Romania

elena@racai.ro

Abstract

In previous papers we described DEACC, a tool that extracts lexical dictionaries from comparable corpora (CC) based on an algorithm introduced by Reinhardt Rapp in 1999 and extended by us through various amendments and heuristics. While anterior experiments limited the evaluations at the level of accuracy of the results (the percentage of source words that received correct target translations), we consider that an analysis of the amount of new translation information that such a method can provide is very necessary. Accordingly, we want to look for answers for the following questions: How many new (not seen in the seed lexicon) words were extracted? How many of the new words have accurate translations? How big a seed lexicon should be so that the newly acquired words justify the extraction work?

Keywords: comparable corpora, parallel corpora, parallel data extraction, phrase alignment, machine translation

1. Introduction

Using parallel data to extract translation knowledge is the established practice in the machine translation technology. But to have sufficient parallel corpora available for one's translation needs is a privileged position, in which only the most spoken languages are. Usually, to acquire parallel corpora (PC) involves paying intellectual property rights for a text in the source language and its human translation in the target language. To be used in automatic translation, this corpus must be consequently aligned at document/paragraph/sentence level, morphologically and/or syntactically annotated, etc.; this is expensive both in terms of money and time. For less economically and culturally visible languages (in which category one can consider, between many others, the Eastern-European/Balkan languages), the digitalized amount of texts, and implicitly the amount of parallel data, is significantly reduced.

One of the forthcoming ideas for improving the situation of the under-resourced languages in Machine Translation (MT) was to collect a less pretentious type of corpora: comparable corpora. The EAGLES – Expert Advisory Group on Language Engineering Standards Guidelines (1996)[3] defines a comparable corpus as "one which selects *similar* texts in more than one language or variety." The condition for the data parallelism is replaced by the weaker condition of *similarity*, which makes the corpora much easier to procure. News articles about the same subjects or Wikipedia entries

[3] http://www.ilc.pi.cnr.it/EAGLES96/browse.html

describing the same entities or concepts can be downloaded and collected to compile a comparable corpus.

Of course, to use CC as a basis for MT one needs techniques for extracting translation information from them. Parallel information can be identified at document, paragraph, sentence or inter-sentential level. Intuitively, the translation information that can be extracted from CC is less reliable than the translation tables extracted from PC; this is why the first approaches in the field imagined a base-line machine translation system constructed on as much parallel data as can be acquired and improved by adding data extracted from CC.

2. Methodologies for extracting translation data from comparable corpora

As we already mentioned in the **Introduction**, under-resourced languages must rely on parallel information scattered on the web to compensate the gap to the privileged languages' technologies. Methods for collecting data and developing CC have been created and documented, but we will not pause upon them. For more inside on these matters, we recommend the ACCURAT project report (Paramita et al, 2011).

Parallel data may be found in CC at any textual level: document, sentence, phrase, word. Different algorithms have been developed that focus on a specific level of granularity.

For CC extracted from structured platforms like Wikipedia, which organizes its entries according to some interlingual identifiers, the **document alignment** is an easy task. In other situations, where there are no obvious links between documents in different languages, a variety of alignment techniques have been developed.

Tao and Zhai (2002) employ Pearson's correlation coefficient variant *r* to compute similarities between words in the documents corresponding to the two languages. Using the word similarity measure *r(x,y)*, they construct a document similarity function:

$$s(d_1, d_2) = \sum_{x \in d_1, y \in d_2} r(x,y) p(x|d_1)\, p(y|d_2),$$

where *x* and *y* are words from documents d_1 and d_2 and $p(x|d)$ is the probability for the occurrence of *x* in *d*. The alignment precision of this algorithm doesn't rise above 86%.

But Vu et al. (2009) improve it by adding a Date-Window filter to reduce the search space (assuming that documents on the same subject are created around the same date). Furthermore, a second filter called Title-n-Content favours alignment candidates which have at least one title-word of the source document translated in the target document content. They also add a linguistic feature which concerns terms (multi-word expressions acting like single units) and replace Pearson's correlation coefficient with Discrete Fourier Transform to compute the similarity score of two frequency distributions. (Vu et al., 2009) reports an increase in the alignment precision of 4% for En-Chinese and 8% for En-Malay compared with (Tao and Zhai, 2002).

(Munteanu & Marcu, 2002) and (Munteanu, 2006) use a Cross-Lingual Information Retrieval Technique (CLIR): they translate source words from a document using a bilingual lexicon and use the translations to construct a query which is run against the collection of target documents. The top k documents returned by the IR engine are the most probable pairings for the query document. This approach is designed to ensure a high recall rather than a high precision of the alignment.

Another approach is to translate the target documents with MT systems and compare the translated document *D'* with the candidates D_i. A classical technique is to identify the similar documents D_i through a vector-based clustering algorithm. Montalvo et al. (2006) use named entities and their cognates to perform cross-lingual clustering and obtain 90% accuracy. Abdul-Rauf and Schwenk (2009) measure the closeness of the D_i documents with TER, a standard MT evaluation metric and select the document with the smallest distance:

$$D^* = argmin(TER(D_i, D')).$$

Ion et al. (2011) designed an Expectation Maximization algorithm to align different type of textual units, including documents. They imagined an analogy with the IBM-1 model for word alignment, where the translation probability is computed through an EM algorithm and the hidden variable *a* models an *assignment* (1:1 word alignments). Similarly, an assignment between two sets of documents (a 1:1 sequence of document correspondences) can be modeled by a hidden variable {*true/false*} and is determined by word translations between pairs of documents. The hypothesis is that there are pairs of translation equivalents which are better indicators of a correct document correspondence.

Parallel sentence extraction techniques are based on the assumption that comparable documents/corpora may contain some sentences which are reciprocal translations. Most of the approaches described in the document alignment section have been adapted and used for sentence alignment. Before web-crawling for pages with similar URLs, Resnik and Smith (2003) use a lexicon based on parallel data to compute alignment scores between documents or sentences. Similarly, Zhao and Vogel (2002) find (nearly) parallel sentences in comparable documents through dynamic programming. For each n:m possible alignment between the sentences, they compute an alignment score based on a word alignment model, use special insertion and deletion models and find a path which maximizes the total alignment probability. Abdul-Rauf and Schwenk (2009) use IR techniques (WER, TER) and simple filters like the sentence length rate to identify the most similar sentence in the target language.

More recent approaches have been developed (inside the European project ACCURAT) at RACAI and have been described in (Ştefănescu et al., 2012). LEXACC requires aligned document pairs for a better precision of the sentence alignment. The algorithm interpolates five features functions: 1) a translation overlap score for content words using GIZA++ format dictionaries, 2) a translation overlap score for functional words, 3) the alignment obliqueness score, 4) a punctuation score and 5) a score indicating whether strong content word translations are found at the beginning and the end of each sentence in the given pair.

Phrasal alignment approaches are following similar steps. A standard phrase alignment algorithm relying on the Viterbi path of the word alignment, a binary classifier algorithm and a lexical features based algorithm are the three techniques used by Hewavitharana and Vogel (2011); the best performance in terms of precision, recall and F-measure is reported for the last technique.

PEXACC, developed together with LEXACC at RACAI for the ACCURAT project's purposes, linearly combines a set of feature functions f_i (which output translation similarity scores between 0 and 1) to obtain the final score of parallelism *P* for two phrases *e* (in the source language) and *f* (in the target language)

$$P(e,f) = \sum_i w_i f_i(e,f), \quad \sum_i w_i = 1, \quad 0 \leq f_i(e,f) \leq 1, \forall\, e, f, i.$$

The most popular method to extract **lexical dictionaries** from CC, on which we based the construction of our tool, is described and used by Rapp (1999). It relies on external seed dictionaries and is based on the hypothesis that *word target1 is a candidate translation of word source1 if the words with which target1 co-occur within a particular window in the target corpus are translations of the words with which source1 co-occurs within the same window in the source corpus.*

The translation correspondences between the words in the window are extracted from *seed dictionaries*. A *co-occurrence matrix* is computed both for the source and for the target corpus: each of its rows corresponds to a type word in the corpus and each column corresponds to a type word in the base lexicon. Finally, similarity scores are computed between all the source vectors and all the target vectors computed in the previous step, thus setting translation correspondences between the most similar source and target vectors. Different similarity scores were used in variants of this approach; see (Gamallo, 2008) for a discussion about the efficiency of several similarity metrics combined with two weighting schemes: simple occurrences and log likelihood.

3. DEACC: adapting and extending the original Rapp's approach

3.1. What is new in DEACC

Initially, the co-occurrence matrix is constructed based on the co-occurrence frequencies in the corpus. In a subsequent step, the frequencies are replaced by log-likelihood scores which are able to eliminate word-frequency effects and favour significant word pairs. In our approach, this is followed by a step of LL filtering, in which all the words that occur with an LL smaller than a threshold are eliminated. The filtering was motivated by the need to reduce the space and time computational costs and is also justified by the intuition that not all the words that occur at a specific moment together with another word are significant in the general context of our approach (the LL score is a good measure of this significance).

The seed lexicons we used in our experiments are translation tables, automatically extracted using GIZA++ from parallel corpora. In such a table, a source word can have multiple translations and each pair (*source*, *target*$_i$) is associated with a translation probability. This introduces polysemy in our seed lexicons, situation which is avoided and not discussed in the standard approach. Other approaches either keep

for reference only the first translation candidate in the dictionary or give different weights to the possible translations according to their frequencies in the target corpus. We think one need to take advantage of all possible translations, as the semantic content of a linguistic construction is rarely expressed in another language through an identical syntactic or lexical structure. Our solution was to *distribute the log-likelihood of a word pair* (w_1, w_2) in the source language to all the possible translations of w_2 in the target language as follows:

$$LL(w_1,w_2) = \sum_i LL(w_1,w_2) * p(w_2,t_i)$$

where $p(w_2,t_i)$ is the probability of a word w_2 to be translated with t_i and $\sum_i p(w_2,t_i) = 1$.

As the purpose of DEACC (and of all the other tools in the ACCURAT project) was to extract – from CC – data that would enrich the information already available from parallel corpora, it seemed reasonable to focus (just like Rapp (1999) did) on the open class (versus closed class) words. Because in many languages, the auxiliary and modal verbs can also be main verbs and most often the POS-taggers don't discriminate correctly between the two roles, we decided to eliminate their main verb occurrences as well. For this purpose, the user is asked to provide a list of all these types with all their forms in the languages of interest.

Being based on word counting, the method is sensitive to the frequency of the words: the higher the frequency, the better the performance. In previous works, the evaluation protocol was conducted on frequent words, usually on those with the frequency above 100, an option that ensures very accurate translation candidates. However even if the operation causes loss of precision, the frequency threshold must be lowered when we are interested in extracting *more data*; in our tool, this parameter can be set by the user, according to his/her needs.

Following the conclusions of Gamallo's (2008) experiments, we used as a vector similarity measure the DiceMin function. In computing the similarity scores, we did not allowed the cross-POS translation (a noun can be translated only by a noun, etc.); the user can decide if he/she allows the application to cross the boundaries between the parts of speech, through a parameter modifiable in the configuration file. Each choice has its rationales, as we know that a word is not always expressed through the same part of speech when translated in another language. On the other hand, putting all the words in the same bag increases the number of computations and the risk of error. For the proper nouns, which are more probably to be translated into a similar graphic form from a language to another, we introduced a cognate score (based on Levenshtein Distance), which is used in the computing of the similarity metric to boost the cognate candidates.

If the user's machine has multiple processors, the application can call a function that splits the time consuming problem of computing the vector similarities and runs it in parallel. The tool is implemented in the programming language C#, under the .NET Framework 2, and is language independent, providing that the corpus is POS-tagged according to the MULTEXT-East tag set[4] and that the user is introducing manually in the configuration file the list of source and target verbal forms to be ignored by the algorithm.

[4] http://nl.ijs.si/ME/V3/msd/html/msd.html

3.2. Initial Experiments

The seed lexicon we used is a word-to-word sub-part of a translation table, extracted with GIZA++ from corpora in different registers. Only the content words were kept. The translation table can be loaded as two different dictionaries EN-RO (64,613 polysemic entries) and RO-EN (66,378 polysemic entries).

Tests have been conducted on two different CC of different sizes types/registers:

1. A corpus of articles extracted at RACAI from Wikipedia: 743,194 words for Romanian, 809,137 words for English; strongly comparable one, with little noise (due to the fairly similar structure of the wiki pages, which facilitated the elimination of the boilerplates).

2. A corpora compiled by USFD in the ACCURAT project: journalistic corpora downloaded from Google News through a heuristic based on a list of English paper titles, translated into Romanian. For more details, see (Paramita et al, 2011).

The pre-processing (tokenization, insertion of diacritics, lemmatization, POS-tagging) of the comparable corpora has been described in (Irimia, 2012) and we will not detail it here. Initially, we manually compiled a gold standard lexicon of around 1,500 words (common nouns, proper nouns, verbs and adjectives) from the Wikipedia corpus. In the conditions described by the default parameters in the configuration file, the *precision-1* (the number of times a correct translation candidate of the test word is ranked first, divided by the number of test words) and *precision-10* (the number of correct candidates appearing in the top 10, divided by the number of test words) scores were computed:

Table 1: P-1 and P-10 for the 1,500 test words from Wikipedia corpus

POS	**Precision-1**	**Precision-10**
common nouns	0.5739	0.7381
proper nouns	0.6956	0.7336
adjectives	0.4943	0.6292
verbs	0.6620	0.8275

Because the initial experiments with the USFD corpus were very disappointing, we acknowledged the need for correcting some POS annotations and also for introducing two different frequency thresholds for the two corpora (English: 7,280,609 words; Romanian: 2,170,425 words), to compensate for the difference in size. We also used the Levenshtein Distance for all the analyzed POS, to boost those scores that correspond to graphically similar translations. This boost is done after all the similarity scores between a certain source word and all the target words are computed. The threshold to which the words were considered cognates was LD<0.3 and the boost meant a multiplication with 10 of the similarity score. All the scores that resulted above 1 were reduced to 0.99. After all these heuristics, the results became more reasonable, but still not rising to the performances obtained on the Wikipedia corpus. We see this as a consequence of the serious difference in the degree of comparability between the two corpora.

We constructed gold-standard dictionaries with 100 entries for common nouns, verbs and adjectives and Precision-1 and Precision-10 scores were computed:

Table 2: P-1 and P-10 for the 300 test words from USFD corpora

POS	**Precision-1**	**Precision-10**
common nouns	0.2909	0.5454
adjectives	0.3663	0.5049
verbs	0.24	0.48

The effect of introducing the cognate test for all the POS was important for many of the good results, producing more forms of the same lemma as possible translations, which is consistent with the rich morphology of Romanian and is very useful in a dictionary. This phenomenon occurred for around 46% of the correct translated nouns, 39% of the correct translate adjectives and 29% of the correct translated verbs.

3.3. Using DEACC results to improve SMT systems

There are two basic directions in making use of the translation data extracted from CC to increase the performance of the SMT systems: adding parallel data extracted from parallel corpora to the training PC or constructing mixture translation and interpolated language models from PC and CC. But before thinking about how to integrate our lexical dictionaries to an SMT, we need to evaluate how reliable is the data we obtained. Using a seed lexicon extracted from a diverse and big corpus, as seen in the previous experiments, conducted to good P-1 and P-10 scores. This approach is fitted for domain-adaptation techniques, in situations when the available parallel corpus is general and the comparable corpus is from a specific domain. But for settings where the parallel corpus (and, implicitly, the seed lexicon) is a small one, the method might produce less accurate results. We experimented with a lexicon extracted from "1984" English-Romanian corpus[5]; the one-to-one, content word only version of this small seed lexicon has only 2870 entries, as opposed to the seed lexicon used in our first experiments, who had around 265,000 entries. We also used the opportunity to vary some other parameters of the application: the frequency, the-co-occurrence window and the LD yes/no option.

As can be seen in Table 3, we did not experiment with all the possible parameters combinations, but guided our decisions according to the results in a previous experimental step. The first set of experiments was composed by four settings:

General/F50-10/LDyes/w
General/F30-10/LDyes/w5
1984/F50-10/LDyes/w5
1984/F50-10/LDyes/w5

from which we learned that there is no significant influence coming from two different frequency ratios. We continued by keeping F30-10 and varying the window *w* to 10 and we noticed a good improvement (the maximum, P1 for adjectives: from 0.3 to 0.48).

[5] http://nl.ijs.si/ME/Vault/CD/docs/mte-d21f/node7.html

Table 3: general vs. "1984" dictionaries

		General Dictionary								**"1984" Dictionary**							
		F50-10				**F30-10**				F50-10				**F30-10**			
		LDyes		LDno		**LDyes**		LDno		LDyes		LDno		**LDyes**		LDno	
		w5	w10	w5	w10	w5	**w10**	w5	w10	w5	w10	w5	w10	w5	**w10**	w5	w10
N	P1	0.18	-	-	-	0.17	**0.25**	-	*0.09*	0.12	-	-	-	0.13	**0.12**	-	-
	P10	0.46	-	-	-	0.46	**0.5**	-	*0.26*	0.36	-	-	-	0.36	**0.39**	-	-
A	P1	0.29	-	-	-	**0.3**	**0.48**		*0.17*	0.23	-	-	-	0.26	**0.26**	-	-
	P10	0.57	-	-	-	0.57	**0.65**		*0.46*	0.51	-	-	-	0.51	**0.53**	-	-
V	P1	0.18	-	-	-	0.19	**0.26**	-	*0.17*	0.12	-	-	-	0.11	**0.1**	-	-
	P10	0.53	-	-	-	0.53	**0.6**	-	*0.48*	0.38	-	-	-	0.38	**0.41**	-	-

Next, we set w to 10 and changed LD to no, obtaining serious decreasing to practically the worst scores for the general seed dictionary. The best parameter combination for the general seed dictionary (F30-10, LDyes,w10 – see the first bold column in the table) was used to test the "1984" seed dictionary and the results improved for P10, but decreased or remained the same for P1 (see the second bold column in the table). As expected, there is a significant decrease in performance when using a small seed dictionary, ranging from 13% to 22% for P1 and from 11% to 19% for P10. But we can still use such a dictionary to extract new information from comparable corpora when the available parallel data are poor.

For the two best settings S1 and S2 (the bold column in the table), we computed the number of new words against their specific seed dictionaries.

Table 4: The number of new words extracted from USFD corpora

	A	**N**	**V**
total number of extracted forms	1887	5530	2945
new forms vs. 1984 dict.	1620 (~86%)	4820 (~77%)	2486 (~84%)
new forms vs. general dict.	604 (~32%)	1572 (~28%)	638 (~21%)

Then we computed the P-1 and P-10 scores for S1 and S2, on gold-standards manually validated for 700 word-forms for each of the noun, adjective and verb categories. Below, one can see that the percentage of correctly translated *new* words when using the general dictionary is insignificant in relation to the extraction effort: around 20 new words in 1000 words are on the first positions in the candidate lists. On the contrary, for the 1984 dictionary, where the computational costs are reduced (~20 minutes for ~12,000 new word forms), there are, on average, 18% correct translations in the first positions of the candidate lists.

Table 5: P-1 and P-10 scores for the new words extracted from USFD corpus

		P1	P10
A	**1984**	0.21	0.32
	general	0.02	0.04
N	**1984**	0.20	0.38
	general	0.02	0.05
V	**1984**	0.15	0.33
	general	0.01	0.02

4. Conclusions

In terms of new information added to the available parallel data, the whole process of extraction using a big seed dictionary was costly and almost futile: too less information for too much work. We already had a lot of parallel corpora from different domains and we wanted to extract new information from a comparable corpora which was quite general (the News domain). However, our experiments showed that when a small seed dictionary is the only available and the comparable corpora is a lot out of the dictionary's scope (news versus prose dated from 1949), the procedure is recommended, either for domain-adaptation or for under-resourced pair of languages, as explained in the introduction.

References

Abdul-Rauf, S. Schwenk, H. (2009). Exploiting Comparable Corpora with TER and TERp. *In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP*, 46–54

Gamallo P. (2008) Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. *Proceedings of LREC Workshop on Comparable Corpora*, Marrakech, Marroco, pp. 19-26. ISBN: 2-9517408-4-0.

Hewavitharana S. and Vogel S. (2011). Extracting parallel phrases from comparable data. *ACL: Proceedings of the Fourth Workshop on Building and Using Comparable Corpora*, Portland, Oregon, USA, 61-68.

Ion, R. (2012). PEXACC: A Parallel Sentence Mining Algorithm from Comparable Corpora. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2181—2188, Istanbul, Turkey

Ion, R. Al. Ceauşu and E. Irimia.(2011). An Expectation Maximization Algorithm for Textual Unit Alignment. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora (BUCC 2011) held at the 49th Annual Meeting of the Association for Computational Linguistics* Portland, Oregon, USA

Irimia, E. (2012). Experimenting with extracting lexical dictionaries from comparable corpora for English-Romanian language pair. *In Proceedings of The 5th Workshop on Building and Using Comparable Corpora: "Language Resources for Machine Translation in Less-Resourced Languages and Domains"*, Istanbul, Turkey, 49-55.

Montalvo, S., Martinez, R., Casillas, A., and Fresno, V. (2006). Multilingual Document Clustering: a Heuristic Approach Based on Cognate Named Entities. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, Sydney, 1145–1152.

Munteanu, D. S., and Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, USA, 289–29.

Munteanu, D. S. (2006). Exploiting Comparable Corpora. PhD Thesis, University of Southern California.

Paramita, M., Aker, A., Gaizauskas, R., Clough, P., Barker, E., Mastropavlos, N., Tufis, D. D3.4 Report on methods for collection of comparable corpora", http://www.accurat-project.eu/index.php?p=deliverables

Rapp, R. (1999) Automatic identification of word translations from unrelated English and German corpora. *ACL-1999: 37th Annual Meeting of the Association for Computational Linguistics. Proceedings of the conference*, Maryland, USA, 519-526.

Resnik, P. and Smith, N.A. (2003). The Web As a Parallel Corpus. *In: Computational Linguistics Journal*, 29:3.

Ştefănescu, D., Ion R. and Hunsicker. S. (2012). Hybrid parallel sentence mining from comparable corpora. *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012)*, Trento, Italy.

Tao, T., and Zhai, CX. (2002). Mining Comparable Bilingual Text Corpora for Cross Language Information Integration. *In Proceedings of KDD'05*, Chicago, Illinois, USA.

Vu, T., Aw, A.T., and Zhang, M. (2009). Feature-based Method for Document Alignment in Comparable News Corpora. *In Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens, Greece, 843–851.

Zhao B. and Vogel S. (2002). Full-text story alignment models for Chinese-English bilingual news corpora. *Interspeech 2002: 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, 517-520.

DATABASE ON THE MEDICO-PHARMACEUTICAL TERMINOLOGY [MPHT] IN VARIOUS DISCURSIVE SPACES: ELABORATION-RELATED ISSUES[*]

STELIAN DUMISTRĂCEL, DOINA HREAPCĂ,
LUMINIŢA BOTOŞINEANU

The Romanian Academy, "A. Philippide" Institute of Romanian Philology,

Iaşi Branch – Romania

steliand@uaic.ro, {doina.hrepca, lumi.botosineanu}@gmail.com

Abstract

The elaboration of a database with systematic information on the two levels of use concerning the MPhT (endogenous discourse – of the specialists, and exogenous discourse – of their communication with the public that "consumes" the products of the research and of the industry in the field) is analysed starting from the following concepts and challenges:

a) discourse spaces, terminological levels, and linguistic barriers from the perspective of pragmatics;

b) general characteristics of the MPhT in terms of diachrony and synchrony, as terminology with high socio-cultural impact;

c) the expression of the metalinguistic function of language in the texts of direct contact with the public and in specialized lexicographic works that have as objective instructing the consumer in the field of medicine and pharmacy;

d) effects of globalization: the opportunities of an effective linguistic contact between the specialist and the public in the MPhT field;

e) the necessity and the possibility of elaborating a database of the type "Medico-pharmaceutical protection of the consumer's rights more effective."

Keywords: medico-pharmaceutical terminology, discursive spaces, terminology levels, linguistic barriers, socio-cultural impact

1. Introduction

1.1 A working project

The elaboration of a database comprising systematic information that makes the distinction between the two levels of use [MPhT] (endogenous discourse – of the specialists, and exogenous discourse – of their communication with the public that "consumes" the products of their research and of the industry in the field) is analysed starting with the following concepts and challenges:

[1] pragmatic–discursive spaces; terminology levels and linguistic barriers from the perspective of pragmatics; communication contract;

[*] Main consultants: Pharm. Irina DUMISTRĂCEL, PhD, Technical Manager, Athlone Laboratories Ltd, Roscommon, Ireland; Dan–Doru PLETEA, MD, College of Doctors Iaşi.

[2] general characteristics of the [MPhT] in terms of diachrony as terminology within a field with high socio-cultural impact;

[3] the expression of the metalinguistic function of the language in texts regarding the direct contact with the public and in specialized lexicographic works that have as objective the training of the consumer in the field of medicine and pharmacy;

[4] effects of globalization: the opportunities of an effective linguistic contact between the specialist and the public in the MPhT field;

[5] the necessity and possibility of elaborating a database of "Medico-pharmaceutical Security Glossary" [MPhSG], in order to make the specialized communication with the public more effective and to protect the consumer's rights.

These are, in fact, the matters we shall treat in the present paper.

1.2. Previous research

We mention that the starting points of this approach are represented by the monographs published by Stelian Dumistrăcel and by articles written in collaboration, published in the past two years, as well as several papers with the same status, presented during various scientific events (currently published). We shall emphasize them briefly in the following lines; they are thematic approaches, regarding pragmatic problems in the relationship between the specialist and the consumer in the medico-pharmaceutical field, which represents the diastratic variation, meaning the differences on the level of professional language and particularized discourses. Firstly, this depends on the degree of education and, secondly, on the general scientific education of the speakers (socio-cultural differences), on the diaphasic variation, which involves the different ways of expression in terms of communication performance, including the *expressive* modalities per se (disease and treatment suppose various psycho-linguistic implications). We have paid attention to text analyses of selected publications from various epochs, which underline the application of the programme that answer to the demands of competence expression regarding the diastratic variation. As for the markers concerning the competency in diaphasic variation, we have paid due attention to the "paratext" structures (prefaces, various types of explanatory notes, etc.).

As personal and strictly specialized bibliography, we make reference to the following titles: (Dumistrăcel, 2000); (Dumistrăcel, 2006a); (Dumistrăcel et al., 2011a); (Dumistrăcel et al., 2011b); (Dumistrăcel et al., 2012).

2. Concepts and terminology

2.1. Pragmatic–discursive space

We use a syntagm, *pragmatic–discursive space* ([PDS] hereinafter), starting from certain concepts coined by Dominique Maingueneau, who distinguishes the *discursive space* (which is an element of the triad also comprising *the discursive field* and *the discursive universe*) as the ideological positioning (identity) of the enunciator. The cited author agrees with the relation between the concepts and the concept of «champ scientifique» coined by Pierre Bourdieu, and developed in the study with the same name (Bourdieu, 1976). Generally, our new approach is determined by the fact that the cited starting point refers especially to ideology such as philosophical schools or political currents (cf. Charaudeau – Maingueneau 2002: 97; 453-454). On the other hand, the PDS concept aims at being distinguished from the general concept of «(discursive)

field», which is confrontation-oriented in various areas of spirituality and social field.

> Other works by Pierre Bourdieu that reflect this area of preoccupations are entitled *Le champ politique* or *Le champ religieux dans le champ de manipulation symbolique*; see also *Le champ journalistique et la télévision* [the title of a series of TV shows, 1996]. Joseph Jurt argues, starting from Bourdieu, the symbolic concept of «champ littéraire», in a study referring to the theory of literary field and to the "internationalization" of literature (cf. Jurt 2001: passim).

In general, when we study the PDS concept we focus on the aspects of pragmatics, as mentioned bellow. More precisely, we underline the discourse adjustment by assessing the characteristics specific to the communication setting and to the personal data of the interlocutors. We have developed our concept based on information regarding communication spaces and registers (Dumistrăcel, 2006b) and on the concepts defined by Eugeniu Coşeriu, regarding «linguistic competence» and «linguistic variation» (Dumistrăcel et al. 2012).

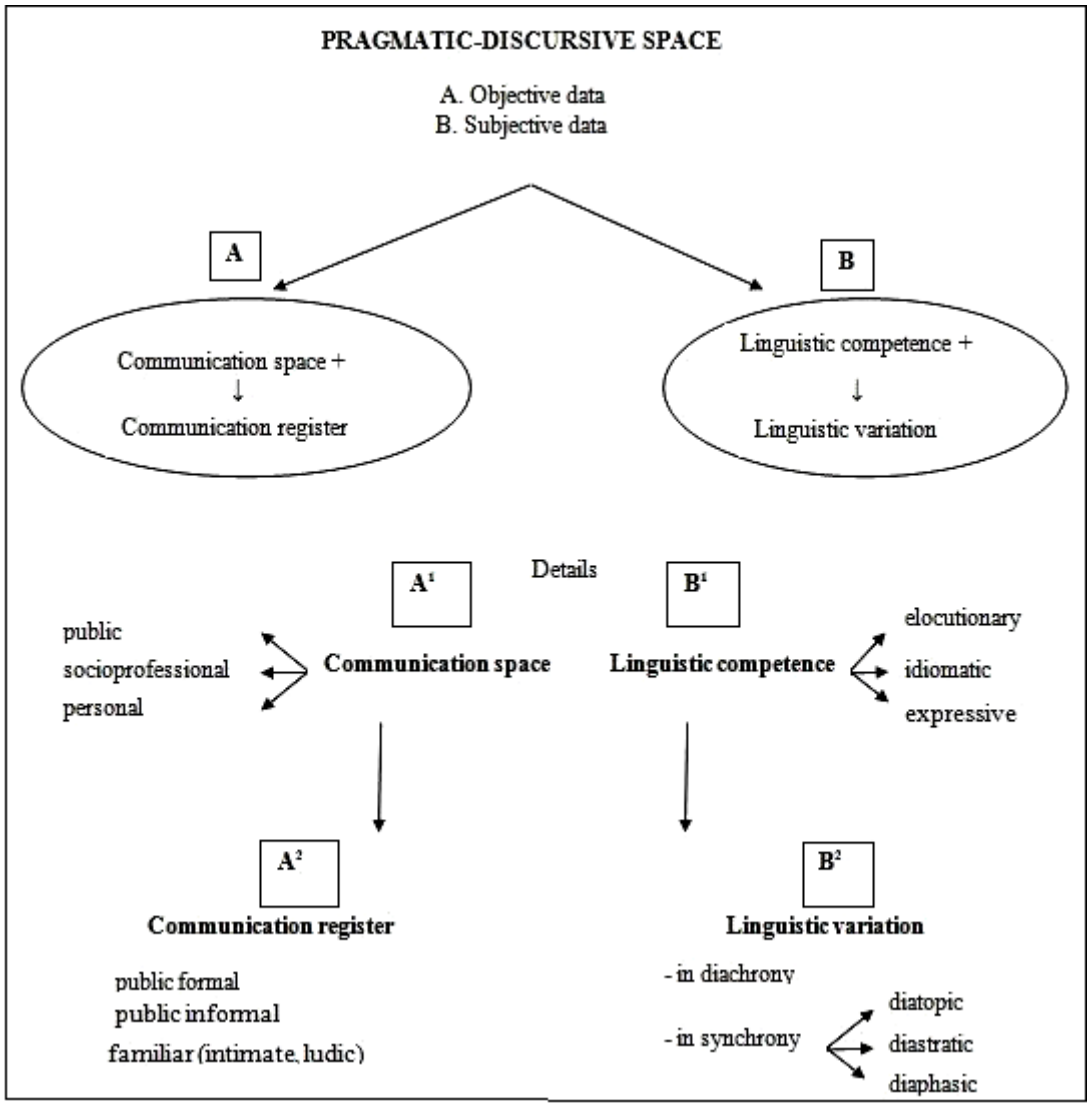


Figure 1. Pragmatic-discursive space

For a brief analysis of Fig. 1, we mention the following coordinates and components: Firstly, two communication coordinates are illustrated: the objective data (A) and the

subjective data (B). The category of *objective data* comprises the communication situation within a given space – the public, socio-professional and personal space. The verbal, paraverbal, and nonverbal correspondents are related (adequate registers): formal public register, informal public register, and familiar register (the last one has also evolved towards intimate or even ludic register). Regarding to *the subjective data* of communication, related to the personal aptitudes of the interlocutors, the first reference is to their *linguistic competence*, mainly to the idiomatic competence (how well the individual knows a language). The second reference is to the expressive competence, which represents the performance "in given situations and concerning certain things, with certain interlocutors" (Coşeriu, 1992-1993), hence, the adequation to the communication situation, to the theme of speech, and to the interlocutor. In other words, this means the capacity/performance of the emitter of placing himself on the same level of idiomatic and expressive competence as the receptor.

2.2. Terminological levels

We have launched the syntagm *terminological levels* (Dumistrăcel, 2000: passim; cf. Also Dumistrăcel et al. 2011a: I, § 1-2), with reference to the distinct level of the discourse specific to the communication in a specialized field, such as the medico-pharmaceutical one, in terms of *endogenous* and *exogenous* discourse (see, above, § 0.1). The starting point in this matter was the analysis of the inventory of terms within the *Vademecum* published by Gheorghe Dănilă (1999). Of the 4,500 entries, only approximately 1% represent terms that can be accepted generally as known and used by a trained public (for the results of the detailed analysis, see also Dumistrăcel et al. 2011b: § 2). In terms of pragmatics, it is worth making a general distinction between terminologies of exegesis *per se*, in fields accessible only to specialists (astronomy, mathematics, chemistry, linguistics, etc.), and terminologies in fields with high socio-cultural impact, with an interest in both exegesis and the public. In this category – where two terminological levels function permanently – one can find, for instance, the legal and administrative terminology, the terminology of religious cult and, the most significantly illustrated, the medico-pharmaceutical terminology. Obviously, our country will be able to add the banking and Internet terminology and maybe other fields.

2.3. The issue of linguistic barriers

Considering the above mentioned aspects, the existence of the specialized level in the area of the terminologies mentioned in the second category – among which the one used for the communication between the specialist (as a doctor and pharmacist), on one side, and the patient, on the other – leads to the creation of true *linguistic barriers*. They result from the social, cultural, status, role, strategic, emotional, etc. differences in the society, seen as barriers.

2.4. Communication contract

The presented elements point out the communication setting and the factors that govern this action, for whose assessment as a whole one can start (with good results) from considering the concept of «communication contract» as a development of the «reading contract» concept, imagined by Eliséo Véron as "communication relation" (Véron 1997: passim). We present certain brief data on the subject: the communication process involves three basic factors: [a] the image of the one who sends a message, the place that he

ascribes to himself concerning what "he says"; [b] the image of the one to whom the message is destined, the place ascribed to him; [c] the relation created based on these images, between enunciator and addressee. In fact, the contractual definition of the speech, which means the existence of two subjects in an intersubjectivity relation, has known various formulations, for convergent visions: "intersubjectivity" (Benveniste), "dialogism" (Bakhtin), "collective intention" (Searle), "joint intentionality" (F. Jacques), "negotiation" (Kerbrat–Orecchioni; for an overall presentation, cf. Dumistrăcel 2006a).

3. General characteristics of [MPhT] from a diachronic perspective

3.1. MPhT characteristics in various stages

At the beginnings of its constitution within the realities of the national culture, the [MPhT] characteristics were analyzed by N.A. Ursu, in a monograph dedicated to the formation of the Romanian scientific terminology. As essential starting point, we outline the translation from French and German, in the second half of the eighteenth century and the first half of the nineteenth century, of various specialized texts. We refer here to short treatises on general medicine, on balneology, as brochures with instructions for epidemics (smallpox, plague) and regarding the cure of diseases, or general norms of hygiene (Ursu, 1962). Within this framework, books of the type "house doctor", of sanitation, hygiene education and treatment occupy a special place, considering the lack of specialized practitioners.

In the phase 1760–1860, but also later, the accessibility in terms of communication was involuntarily ensured for readers of the publications (excepting specialists) by the constant presence of loan translations (that represent "transparent" lexemes and syntagms) and mostly of folk medical terms, besides the neological loans per se. For the current analysis, we have identified a seemingly paradoxical aspect: the linguists who studied the field from the *perspective of the history of the culture language* show (secondarily, of course) certain dissatisfaction for that mixture. The main reason was that the process of creation took more time than what would become the terminology of the modern Romanian literary language, which, by aspiring toward the level "of the endogenous discourse", was proven to have eliminated to a great extent the loan translations and to have decisively got rid of the folk terms.
Another aspect that draws attention – from the same perspective of the study – is the classification (with little differentiation) of all translations done at the end of the eighteenth century and the beginning of the nineteenth century as "books meant to popularize scientific knowledge". In fact, besides such publications and other manuals, publications in the medicine and pharmacy fields are actually books for medical and sanitary *education* or *instruction* per se, thus involving an exogenous discourse (see especially the printings of Ştefan Episcupescul; (Dumistrăcel et al., 2012: I, § 5.3.1 and II, passim).

3.2. Communication perspective

There are not many studies on the specifics of medical terminology in terms of communication; the best known study in Romania is the chapter written by Christian Baylon and Xavier Mignot, entitled *Limbaj şi comunicare medicală* (*Medical language*

and communication), part of the monograph *Comunicarea* (Paris, 1994, translated into Romanian in 2000). Besides general issues, such as "the medical language", "the medical information", the chapter deals especially with aspects such as "the doctor–patient communication" (hence, on the level we call "exogenous discourse"), and "the written communication among doctors" ("the endogenous discourse"). It is worth mentioning the study of the so-called "cryptic function" of the medical language (not of the respective terminology, which has been encrypted anyway since the beginnings of "the division of labour", besides empirical cures, through magical and occult practices). This function refers to both the use of the *jargon* as professional language in medical practice and of the *argot* as communication strategy, aiming at making the patient not understand certain (at least) unpleasant aspects (a cryptic function in action), and to the "technical character" of language, as "potential cryptic function", which gives a "potential power" to the user. Hence, the two authors also take into account the level *of language*, as well as that of the medical *discourse*, depending on the information transmission (Baylon & Mignot, 2000).

4. The expression of the metalinguistic function of language

4.1. Stages in the study of diastratic variation

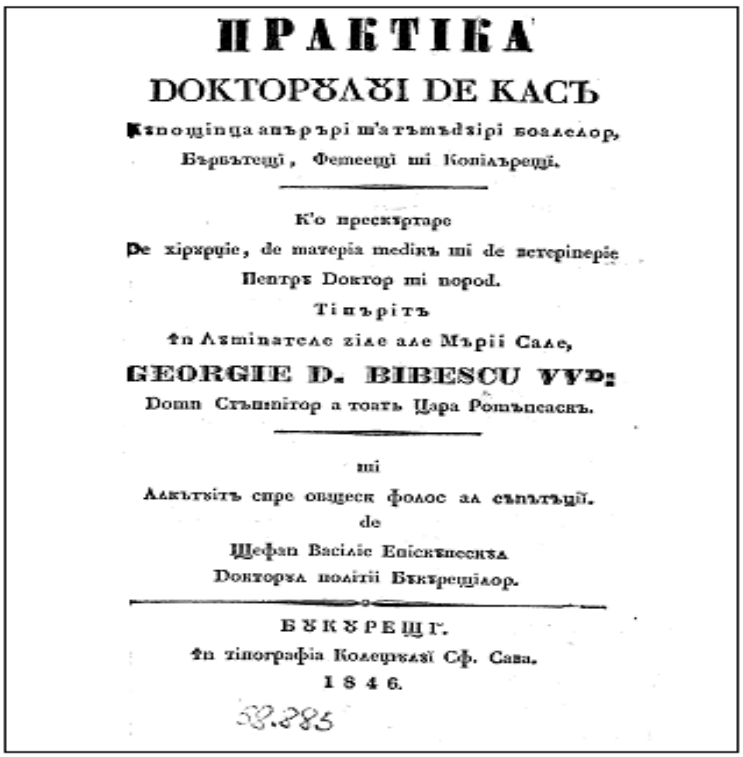
Firstly, we shall outline aspects regarding the diastratic variation in exogenous discourses, in three phases: a) the first half of the nineteenth century; b) the first half of the twentieth century; c) regarding a recent particular specialized work. Secondly, we will refer to the interest for communication with the public in current highly specialized lexicographic works (representing the endogenous discourse, with a minimum opening towards the public).

Books for medical and sanitary instruction and "house doctor" type of dictionaries
The following texts shall comment the idea of "hygiene–medical–sanitary" instruction starting from a Western model from the end of the eighteenth century, which illustrates the application of the Illuminist orientation of mass dissemination of scientific knowledge. We are talking about a "house doctor" in German, printed in Leipzig. The – very instructive – title reads as follows:

> Immanuel Stange, *Der Hausarzt* oder *Anzeige der bewährtesten Hausmittel, und Anweisung sie zur Verhütung oder Heilung der Krankheiten gehörig zu gebrauchen. Ein Handbuch für Landgeistliche, Hausväter und andere Personen, die an Orten leben, wo kein Arzt ist* (Leipzig, Liebeskind, 1797),
>
> which means:
>
> "House doctor or advice on the best known remedies and indications on preventing and curing diseases. A handbook for country priests, heads of families, and for other persons who live in places where there is no doctor".

We find truly remarkable the intuition regarding the necessity of using a language adjusted to the various categories of readers in the works – translations or original works – of a pioneer of the Romanian medical publications, the Walachian doctor Ștefan Episcupescul. He is the author of unsigned translations from Greek, which were attributed to him, and not less than five printed books, which appeared between 1829 and 1846 (for a detailed analysis see Dumistrăcel et al. 2012: part II). As we cannot

discuss here in details the various means through which this author's manifestation of the competence on the diaphasic variation, we shall only cite the facts that illustrate the concern for the differentiation of the text depending on the instruction of the reader categories (diastratic variation). Hence, we shall present certain facts of Episcupescul's book entitled *Practica doctorului de casă. Cunoştinţa apărări ş'a tămăduiri boalelor bărbăteşti, femeeşti şi copilăreşti* [The Practice of the House Doctor. The Knowledge of Preventing and Curing the Diseases of Men, Women, and Children] (1846).



ПРАКТИКА

DOKTORЪLUI DE KASЪ

Кънощінца апърърі ш'а тъмъдзірі боалелор,

Бърбътещі, Фетеещі ші Копілърещі.

К'о прескъртаре

De хірургіе, de матерія медікъ ші de зетерінеріе

Пентрѫ Доктор ші попод.

Тіпъріть

ѫn Луминателе зіле але Мъріі Сале,

GEORGIE D. BIBESCU VVD:

Domn Стъпънітор а тоатъ Цара Ромънеаскъ.

ші

Алкътіть спре общеск фолос ал сънътъцій.

de

Щефан Васіліе Епіскопеснка

Доктораѫ політіі Бъкърещілор.

БЪКЪРЕЩІ.

ѫn тіпографіа Колецілаї Сф. Сава.

1 8 4 6.

38.285

Figure 2. Title page of the book written by Ştefan Vasilie Episcupescul, *Practica doctorului de casă* [House Doctor Practice]

To get an overall picture, firstly we find warnings from contact texts (with a paratext status). After the title page, we find out that the volume is printed "for the Doctor and the people" and that it is "elaborated for the health benefit of the community", an idea also detailed through a distinction within the paratext representing a "dedication" to "Barbu De Ştirbeiu", great *ban* (governor) and "knighted to various orders"; we present the mention below:

> "The book comprises, My lord!, a presentation *suitable for all the categories of our society*: *noble, urban, and rural*, with the simplest indications and the easiest means for any sick person to replace a doctor if necessary" (*op.cit.*, p. VI; our italics).

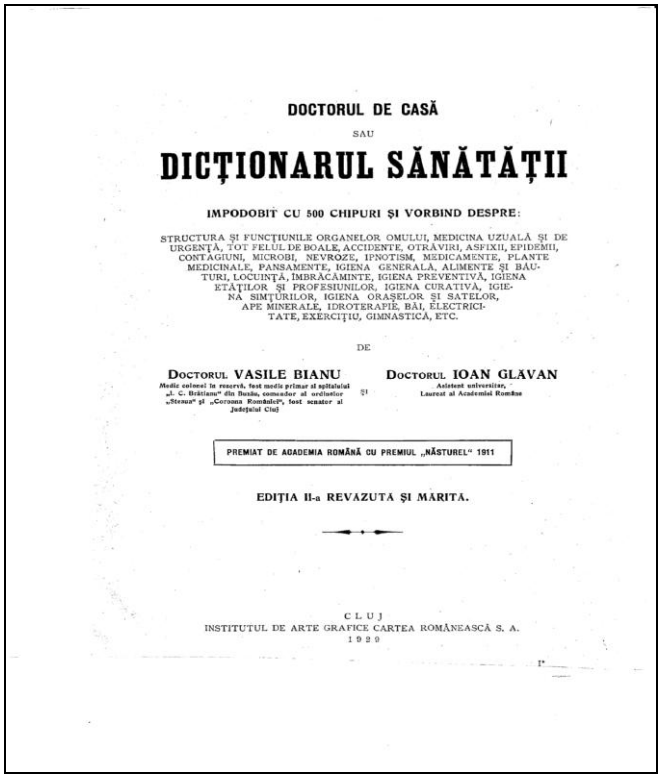
The idea of text adaptation depending of the specialized reader and on the public is illustrated in the general presentation of the book contents (p. XLVII sqq.), "for the people" [a], which represents various practical advice, and "for doctors" [b] "the medicine theory and practice", etc.). We quote the conformation to this communication option through titles from the "Scara cuprinderii cărţi" ["The book contents"] (p. 509; we mention that we have not translated the quotations, as the names of the diseases are transparent as neologisms):

> [a] "for the people": "orânduiala îmbrăcămintei", "~ hrănii", "taina împreunării", "zămislirea pruncului", "îngrijirea lăhuzii", but also "epizootikon, veterineria" (terms indexed by "creşterea şi ţinerea sănătăţii dobitoacelor" – see p. 87-88);
> [b] "for doctors": "Terapia firii: stenia şi astenia sănătăţii"; "Terapia metodică: boala stomahului – morbus gastricus"; "~ răcelii – refrigheraţio"; "~ urechii – otitis"; "~ buboiului şi a sugiului – furunculus panariţium"; "~ pubertăţii – hlorosis şi nostalghia"; "~ întunecimii lintei ochilor – cataracta", etc. (see p. 102 sqq.).

Even when describing the treatment for various diseases, in the texts "for doctors" there are natural alternations between the technical terms of the profession, which are frequently loans or translations from Greek, and the folk medical terms or common language words with special meanings in the communication regarding the care for the ill. For instance, "*Vărsatul spuzos* – scarlatina"; "*Boala sângerăturii matchii* – menoraghia" [the Greek-based version for *menorrhagia* 'condition of the uterus...'] (Episcupescul 1846: 243; 271).

All these prove the special gift of communicator of the doctor Vasile Episcupescul.

For the first half of the twentieth century, an example of performance regarding the application of the diastratic competence is represented by another "house doctor" book, which employs the syntagm even in the title. We refer to the dictionary published by two doctors, Vasile Bianu and Ioan Glăvan: *Doctorul de casă sau Dicţionarul sănătăţii [House Doctor or the Dictionary of Health]*, a work awarded by the Romanian Academy. The references of the present text concern the second edition of the book, published in 1929 (a massive volume, of 804 p., 25 x 20 format). A true bestseller, the same dictionary got to the fourth edition, in 1942.



DOCTORUL DE CASĂ

SAU

DICŢIONARUL SĂNĂTĂŢII

IMPODOBIT CU 500 CHIPURI ŞI VORBIND DESPRE:

STRUCTURA ŞI FUNCŢIUNILE ORGANELOR OMULUI, MEDICINA UZUALĂ ŞI DE URGENŢĂ, TOT FELUL DE BOALE, ACCIDENTE, OTRĂVIRI, ASFIXII, EPIDEMII, CONTAGIUNI, MICROBI, NEVROZE, IPNOTISM, MEDICAMENTE, PLANTE MEDICINALE, PANSAMENTE, IGIENA GENERALĂ, ALIMENTE ŞI BĂUTURI, LOCUINŢĂ, ÎMBRĂCĂMINTE, IGIENA PREVENTIVĂ, IGIENA ETĂŢILOR ŞI PROFESIUNILOR, IGIENA CURATIVĂ, IGIENA SIMŢURILOR, IGIENA ORAŞELOR ŞI SATELOR, APE MINERALE, IDROTERAPIE, BĂI, ELECTRICITATE, EXERCIŢIU, GIMNASTICĂ, ETC.

DE

DOCTORUL VASILE BIANU
Medic colonel în rezervă, fost medic primar al spitalului „I. C. Brătianu" din Dunăre, comandor al ordinelor „Steaua" şi „Coroana Romîniei", fost senator al Judeţului Cluj

ŞI

DOCTORUL IOAN GLĂVAN
Asistent universitar, Laureat al Academiei Române

PREMIAT DE ACADEMIA ROMÂNĂ CU PREMIUL „NĂSTUREL" 1911

EDIŢIA II-a REVĂZUTĂ ŞI MĂRITĂ.

CLUJ
INSTITUTUL DE ARTE GRAFICE CARTEA ROMÂNEASCĂ S. A.
1929

Figure 3. Title page of the book written by Vasile Bianu and Ioan Glăvan, *Doctorul de casă sau Dicţionarul sănătăţii* [House Doctor or Health Dictionary]

A first necessary mention is that the two authors – famous specialists – also have an important contact with various segments of the public. Bianu was also a military doctor (which helped him know the folk terminology for diseases and treatments in this field) and, concerning the public career, he was also a deputy (he dealt with issues related to sanitation education). In his turn, Glăvan, a professor at the university, he published a significant amount of works in the field. In order to illustrate the concern for an efficient communication with the readers, we shall only present the lexicographic correlation, in regard to the entry words representing *neologism technical terms* and *folk terms* in a synonymy rapport, in the tables [1] and [2], with material taken from Bianu – Glăvan 1929: passim.

Table 1: Synonymic correspondences for *terms belonging to science and therapeutics*

acnee – *coşi*, *funigei*;

cataractă – *perdea*;
cefalalgie, cefalee – see *durere de cap* ("durerea *acută* de cap se cheamă *cefalalgie* şi cea *cronică* se numeşte *cefalee*");
constipaţiune – *încuiere*, *încuietură*;
diabet – *boală de zahăr*;
diaree – *cufureală*, *eşire afară*, *pântecărie*, *pârţuică*, *treapăd*, *urdinare*;
fisuri – *crăpături* sau *plezniturі la şezut* (la anus);
fortifiante, *întăritoare* – look at tonice;
idioţie (see this word) – *nerozie*, imbecilitate – un grad mai mic de *prostie*;
intoxicaţie – look at *otrăvire*;
laringe – *beregată*, *gâtlej*, *răsuflătoare*
ocluziune, ~ intestinală – or *încurcătură de maţe* (look at *intestin*);
placentă – *casa copilului*;
scabie – look at *râie*;
strangulare – *gâtuire*, *sugrumare*;

tuberculoză – *atac*, *tusă seacă*, *ftizie*, *hectică*, *oftică*

Table 2: Synonymic correspondences for *folk medical terms*

abubă – *abces alveolar*;
bale – look at *salivă*;
buline (colloquial) – *capsule*;
căldură – *febră*;
ciumă – *pestă*, *pestă bubonică*, *pestă orientală*;
curăţenie – *purgativ*;

dropică – see *idropizie*;
înmoiere de creieri – *ramolisment cerebral*;
leşin – look at *sincopă*;
maţ – look at *intestin*;
nebunie – look at *alienaţie mentală*;
pogană – look at *afte*;
rac – look at *cancer*;
săpunaş, săpunel – look at *supozitor*;
soare sec, soarele în cap – look at *congestiune* şi *insolaţiune*;
sucitură – look at *entorză*;
vitriol – look at *sulfuric* (*acid*)

However, not only the correspondences illustrated by the tables [T 1] and [T 2] make the object of the concerns for eliminating or, at least, for attenuating the linguistic barriers related to the diastratic competence for the readers of the *Dictionary*. For instance, we find the synonymy between neologisms of various ages (!), such as "*flu* – look at *influenza*", "*tablets* or *plates*" (the second meaning is out of use). On the other hand, the attention paid to the diatopic variation is reflected by the presentation of the synonymic correspondence between words within the folk speech – such as "*săpunel* – look at *odogaci*" or "*măsălariţă* – look at *nebunariţă*" –, as well as the richness of regional synonyms for names of plants of interest for treatments diets.

For instance, for the term *potato* there are no less than 15 equivalents (a serious competition for the inventory of the Academy Dictionary and of a dictionary of synonyms):

> "**Cartofi**, bandraburce, baraboi, barabule, bologeane, cartoafe, crumpene, crumpeni, crumpiri, grumciri, hadeburce, mere de pământ, picioci, piciorcă, poame de pământ, ţermer (Solanum tuberosum, fam. Solaneelor)".

The few elements mentioned above illustrate the vocation of competence of the authors of the *House Doctor* analyzed regarding the diastratic variation in terms of [MPhT] and they explain, of course, the success of their book to the public.

The idea was resumed nowadays, especially regarding alternative treatments, naturist medicine. This way, for instance, *Doctorul de casă* (Bucharest, Rom Direct Impex, 1994) is a translation after J. Frank Hurdle, *A country doctor's common sense health manual* (1975); "The house doctor" or "The doctor of the house" are also titles of blogs (cf. http://healthy13-annelisse.blogspot.ro/ or http://www.gustos.ro/articole/sfaturi-practice/aloe-vera-doctorul-casei.html).

The qualities of the works briefly presented in § 3.1.1 and 3.1.2 are also significantly highlighted through a comparison with a specialized dictionary recommended *from an editorial perspective* as largely accessible work.

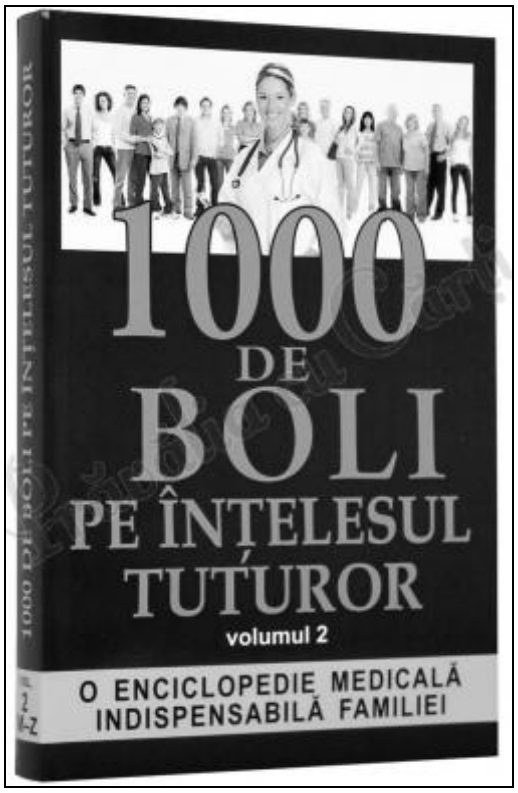


Figure 4. First cover of the book *1000 de boli pe înţelesul tuturor* [1000 diseases in plain language], vol. II

Hence, the translation from French – made by "Dr. Cosmin Pop" – of the work of the two French authors, Ch. Prudhomme and J.-F. D'Ivernois, with a rather easy-to-go original title, *Connaître et comprendre 1000 maladies de A à Z* (Paris, 2009), is presented in "the reading threshold" constituted by the inscription on the first cover. The cover "overestimates" the competence of the addressee in terms of "knowing things" – "in plain language" and "medical encyclopaedia *indispensable to the family*", through seductive formulas for the buyer. If the text on the cover might have been a mere editorial strategy, the declared intention is also present in the title: the very title page "1000 de boli *pe înţelesul tuturor*" ["1000 diseases *in plain language*"]. However, it is not even by far equivalent with "connaître et comprendre" (we shall not discuss here other comparable appealing formulas comprised in the texts that appear on another "reading threshold", the fourth cover of each of the two volumes of the Romanian version of the book).

We have detailed the aspects mentioned above considering that, at first glance, "1000 diseases" could be considered a current counterpart of the Bianu – Glăvan dictionary, which even *mutatis mutandis* is not in conformity with the reality. Most entry words represent scholarly technical terms, such as *choanal atresia*, *eritrasma*, *chronic subdural hematoma*, *polymyositis*, *tularaemia*, etc., while the description of the disease usually belongs to the same register. For instance, Verucile [the *Warts*] [D 7]: „tumori benigne ale pielii provocate de virusuri din familia papilomavirusurilor umane (HPV)..." ["benign skin tumours caused by the human papillomavirus (HPV)..."]. There are also simple correspondences probably representing differences brought by various medical schools: *dracunculoza sau filarioza de Medina* [*dracunculiasis or Medina worm filariasis*], *otita cronică colesteatomatoasă sau colesteatomul* [*chronic otitis cholesteatomatosa or cholesteatoma*]. The cited examples illustrate the status of the targeted readers: specialized doctors or medical students (such situation are also present in Bianu – Glăvan 1929: passim, but they are much less frequent).

However, one may still identify various degrees of accessibility in the case of the diseases representing neologisms that have become part of the common lexicon, such as *angină* [*angina*], *bronşită* [*bronchitis*], *cancer* [*cancer*], *diaree* [*diarrhoea*], *rujeolă* [*measles*], *şancru* [*chancre*], *tuberculoză* [*tuberculosis*], etc. For some of them, there are, sporadically, folk correspondences; for instance, "*Antraxul* sau (vezi) *cărbunele*" ["*Anthrax* or (see) *coal*"], a word (erroneously alphabetised in the entry list) under which the disease is described or, in a more complicated way, „*Lobstein* (boala) sau *boala oaselor de sticlă* sau *osteopsatrioza*" ["*Lobstein* (disease) or *brittle bone disease* or *osteopsathyrosis*"]. There are also other situations of equivalence. For instance, „*bot de iepure* sau (vezi) *palatoschizis*" ["*harelip* or (see) *palatoschisis*"] or „*păduchii* sau *pediculoza*" ["*lice* or *pediculosis*"] – but only in the *Index*; however, the index is not well elaborated, because it does not indicate, technically, which of the elements within the synonymic series is the entry word and which is the variant. For instance, in the above-cited cases, the scholarly term is the entry, and the folk variant is only some sort of commentary (we shall not discuss here other technical dysfunctionalities). Finally, they seemingly belong to the same category, such as in „*boala somnului* sau *tripanosomiaza africană*" ["*sleeping sickness* or *African trypanosomiasis*"], or „*viermele solitar* sau (vezi) *teniaza cu taenia saginata*" ["*tapeworm* or (see) *taeniasis with taenia saginata*"].

It is obvious that such an approach actually ignores the common reader, interested in a diagnostic or/and in a treatment, but the analyzed dictionary has incontestable merits for the reader familiarised with the endogenous discourse. However, we have studied "1000 diseases in plain language" because this encyclopaedia-like lexicon is, to a certain extent, a pale counterpart to works that openly claim the profile of a strictly specialized dictionary, such as the one we study study hereinafter. However, we do warn that Rusu 2012 also treats with interest the issue of folk medical terminology.

4.2. A reliable dictionary

The competence regarding the diastratic variation in the linguistic relation between doctor (pharmacist) and the beneficiary of specialized services is convincingly illustrated by several precepts formulated by the authors of the dictionary Rusu 2012. An example is the following: "From a pragmatic perspective, getting to know the medical terms in circulation [on the level of common and folk speech] can serve to a better communication between doctor and patient". Below, we cite the presentation of medical practice realities:

> "The doctor–patient dialog takes place on *two* distinct *language levels* [this is another formula for what we call «terminological levels»]: the doctor needs precise, mainly anatomical terms, to locate the disease, as well as a clear expression of the symptoms. The patient may indicate the location and characteristics only approximately or in a totally different verbal code".

There is a discussion on the issue of the so-called "exaggerations", which complicate the communication process: the manifestation of shyness, but also of the "vulgarity" (which should be considered with "tolerance"). Under these circumstances, "the use of folk medical terms for both speakers" is required, and

> "Experienced doctors spontaneously adapt the language to the patient's age, profession, reserve, or, on the contrary, the behaviour to the limit of mutual respect, the lack of confidence expressed by the patient" (Rusu 2010: 1433; see also the references to the need to/re/humanise the medical act through the dialog with the patient, in the *Introduction* to the fourth edition, p. 19). See also the criticism to what the author of the preface for the fourth edition, Dr. Gabriel Ungureanu, calls "the aggressive invasion of anglophone terms" in the past few years, as well as to "the irritating filling of the medical language with Americanisms", sometimes "out of pure intellectual snobbism"; *op. cit.*, p. 9).

Thus, in the most pretentious guide of scientific medical terminology in our country – where there are numerous strictly specialized sections that we cannot enumerate here for reasons of space –, as Rusu (2010) represents, in fact, the contemporary higher level of the [MPhT] belonging to the endogenous discourse, given all the possibilities of expression for the potential cryptic function (cf. § 2.2), the folk popular terms are considered very important for performance. By indicating their scientific correspondents or their meaning, these terms are made available to the specialists in a glossary that comprises over one thousand entries. These conclusions of professional common sense, after all appear in the introduction to the *Glossary of folk medical terms*, whose presence is motivated in a highly persuading manner from the perspective of the imperatives of professional communication.

Excursus. A brief assessment allows us to appreciate that a recent publication of the Romanian book, a *Dicţionar medical ilustrat [Illustrated Medical Dictionary]*, representing the translation (planned to be published into 12 volumes) of the Italian original SALUTE. *Dizionario medico* (Milan, RCS Quotidiani, 2006) represents – considering its intention – a welcomed compromise regarding the addressee. In parallel with the endogenous discourse – moderated –, there are (not only within the articles per se) sections which have in view the pragmatic–discursive space of the presumptive patient; certain subdivisions of the articles are even accessible to the readers with a certain level of instruction and of average cultural formation. A simple enumeration of certain types of article substructure highlights this communicative opening; type [1]: a) the generally accessible definition of the entry word and, in parallel, in a special case, "Prophylaxis"; b) treatment; type [2]; a) definition; b) causes; c) symptoms; d) diagnostic; e) treatment (cf. Dicţionar 2013: passim). However, it is obvious that such a dictionary is far from representing a... competition for an information tool of the [MPhSG] type.

5. Effects of the globalization in the [MPhT] field

5.1. Categories of specialized terms

For the study of the linguistic barriers that emerged – mostly in the past decades – through the globalization of industry and of the pharmaceutical market, as a result of international enactments also adopted in Romania and with important linguistic effects, the issue of three categories of names requires further investigations:

– [a] the denomination issue in terms of "inventing" *the drug/medicine* (= M);
– [b] the denomination issue in the phase of prescribing M within the doctor–patient relationship;
– [c] the issue of the denominations used when selling M in pharmacies.

In order to determine the level of communicative performance, of the [a] category of texts, representing the M "invention" and "launching", linguistic analysis assumes the consideration of the effects of terminological regulations, referring to the following criteria:

– [α] *innovative* M,
– [β] *generic* M, for which a study of special interest is the so-called "*umbrella* terms".

The study on the performance of the term by the internal criteria of the field imposes the consideration of the text of laws regarding the "visibility tests" and, on the other hand, of the official texts such as *Ghid privind denumirea medicamentelor de uz uman* [*Guideline on the trade name of medicinal products for human use*] (2008), etc. The limitations of encoding the "invented" term may be tracked down based on interpreting the interdictions within the scheme in Fig. 5, referring to the "Decision tree", elaborated according to the instructions of World Health Organization (WHO), which we present below:

Approach on the issues related to international common denominations (ICD) within the proposed invented denominations (ID)

1) The similarity between an invented denomination and an ICD (of the medicine in question or a different ICD) – in the written and/or verbal form – taking into account the medical context and/or the conditions of use and/or administration route of the medicines in question is treated as follows:

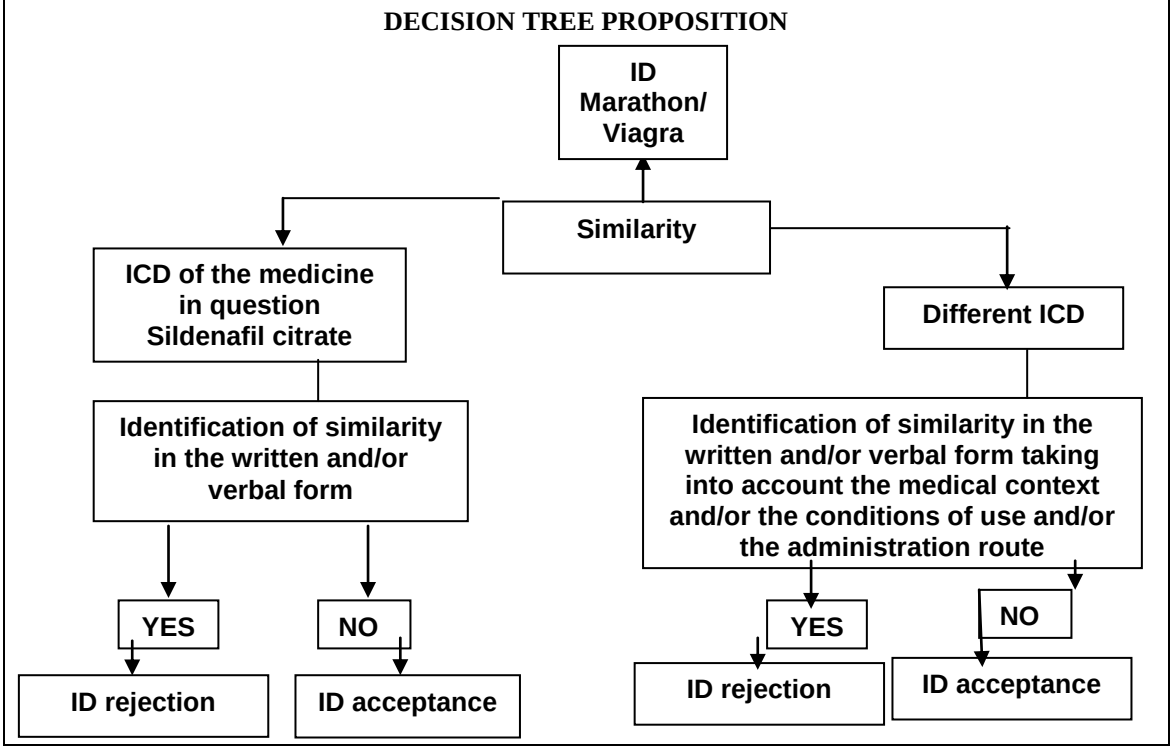


Figure 5. Decision tree proposition for the medicine Marathon/Viagra

The term can be rejected, on one side, if there is a possibility of identification with similar medicines, a similarity in the written and/or verbal form, considering the medical context *and/or the conditions of use and/the administration route*. In addition, on the other side, it can be rejected if – regarding the remedies within the same or a different therapeutic class – the invented term *provides indications on a public health issue*. Furthermore, it can be rejected if there is a *possibility of identifying a similarity* taking into account the *medical context and/or the conditions of use and/the administration route* (we underline the formulations that we have emphasized with italics!). We are dealing here with "branding" and "marketing" issues.

For the [b] phase, the research on the respective terminology considers the existence and practical functioning of the regulations imposed to the family doctor, concerning the concept "Brand *vs* ICD" [=International Common Denomination], comprised within the documents:

– prescriptions following the ICD norms;

– medicine prescriptions according to a "framework agreement" (for instance, that of 2011);

– "pharmaceutical rules of good practice", etc.

The study of legislation referring to the denominations of medicine leads to the conclusion that the respective medicine – though one cannot directly accuse it of exacerbating a "cryptic" function of language – imposes the use of a foreign and even rebarbative terminology for the Romanian consumer. This is why the issue of investigation – in terms of knowledge of the pragmatic–discursive space – proposes the following objectives:

– the concrete manner of social functioning concerning the current legislation on the "international" denomination of the medicine;

– the proceedings of decoding the terms within the endogenous discourse in the functioning of the relation "trade name" *vs* "ICD".

For the [c] phase, the [MPhT] research involves – starting from [α] – the doctor–pharmacist relationship, as endogenous discourse, a theme rarely discussed so far, [β] the study of the relationship between pharmacist and the beneficiary of the treatment, as exogenous discourse. This relationship takes place where the medicines are commercialized, but this issue has been almost absent from the Romanian pragma linguistic research; it is, however, especially interesting for the elaboration of the work we shall refer to in § 5.

> Concerning the [α] component, some of the aspects are studied, for instance, in a paper signed by Dr. Rodica Chirculescu, *Relaţia doctor–farmacist, colaborare şi răspundere asumată* [*Doctor–pharmacist relationship, collaboration and assumed responsibility*] (cf. http://www.pharma-business.ro/oportunitati/relatia-doctor-farmacist-colaborare-si-raspundere-asumata.html). On the same site (www.pharma-business.ro), there is an article on the [β] component, entitled *Principiile comunicării farmacist–client [Principles of the pharmacist–client communication]*, signed by Anda Păcurar (http://www.pharma-business.ro/oportunitati/principiile-comunicarii-farmacist-client.html). There is no doubt regarding the need to study this general issue, as well as to elaborate the planned "Glossary", mostly because there are absurd perspectives, such as the one presented within an interview (with Dr. Cristian Cârstoiu), entitled precisely *Comunicarea farmacist–pacient* [*Pharmacist–patient communication*], published in "Practica farmaceutică" [Pharmaceutical Practice] (vol. V, no. 3-4, 2012, p. 130-132). In this interview, the communication barrier issue – superficially and unprofessionally approached – is simply sent off in a sentence that invokes, by "defending" the pharmacist, besides the lack of time, the difference of education. The sentence reads as follows: "often patients simply do not have the necessary knowledge". Of course, aspects that are even more... human – though general – are considered: "For advice, one should use a language adjusted to the degree of information of a patient, and it should avoid specific terms" (http://heppyportal.projectize.eu/ database/publications/publication_45_ro.aspx).

5.2. Structure of contemporary denomination

We cannot propose to analyze the strictly current specialized nomenclators; a brief presentation in this sense was included in *Vademecum medicamentorum*, published by Gheorghe Dănilă (cf. § 1.2). Compared to the 1% value of the "transparent" names of medicines for the public with medium education, estimated based on the cited source, the percentage is even lower nowadays. We make this statement considering the fact that the Romanian [MPhT] has been radically changing, a characteristic of the globalization in terms of medicine production and commercialisation, with a focus on the existence and functioning of the two above-defined distinct terminological levels regarding the communication between emitter and receptor (cf. § 1.2). The reference is made to *the endogenous discourse*, which became internationalized, with effects on the *exogenous* one.

On the other side, it is also interesting to study the effects of using a so-called pharmaceutical *lingua franca*, based on the scientific terms for the active principles of medicines. Concerning this system, insurmountable linguistic barriers have started emerging in terms of the common communication (we could consider a profile previous to the one illustrated by Hepites 1862; cf. below, § 5.0).

5.3. The issue of elaborating a Medico-pharmaceutical Security Glossary

Our interest – during the planning stage for an overall investigation – for the elaboration of a "Medico-pharmaceutical Security Glossary" also concerns the diversion it's represented for the medicines *per se* ("allopathic", in specialized terms), the real competition, regarding the contact with the users of alternative treatments [AT]. These treatments concern the so-called naturist products, "nutritional supplements" or "food supplements", present in pharmacies. The pharmacies represent a trade setting where an image transfer occurs in favour of the products in the [AT] class.

In the area that we are referring to, transparency is not prohibited, regarding neither the verbal, nor the iconic message; both act directly and effectively through the contact texts represented by packaging, leaflets, and advertising. In this case, we are not talking about prescriptions, which constitute contact texts within the strictly interdisciplinary discursive space, because the prescriptions use only the terms for the active substances, and not the trade names).

Without providing any more details (however, see Dumistrăcel et al. 2011b), we shall present several types of "hyper-transparent" terms on the [AT] level. In pharmacies, there are products whose names – in the contact texts such as packaging and advertising – are transparent, in the sense that they are relatively easily associated, on a certain level of idiomatic competence, with terms for diseases, treatment, or even substances. For instance, *Colonhelp*, *Urinal*, *Acneogel*, *Hepatobil*, or *Calmocard* (with three layers of the reception level through the text on the packaging: „*Calmocard* [1]. *Calmant cardiac* [2]. *Contribuie la buna funcţionare a inimii* [3]" ["*Calmocard* [1]. *Cardiac analgesic* [2]. *Contributes to an optimal heart functioning* [3]"]). The same opening toward immediate acceptance goes for medicines from *Calmoplant*, *Larvalbina*, *Tutunstop*, to "*Hapciu*" („Ceai *Hapciu"* ["*Hapciu Tea"] and „Trusa Hapciu* – un tratament natural contra răcelii şi gripei" *["Hapciu* Kit – a natural treatment against cold and flu"]). Nevertheless, there is a significant distance between the possibility of deciphering the name of a certain medicine (though it may be semitransparent), *Hemorzon*, compared to the name of... a competitor, a "nutritional supplement", *HemoroEasy* (pronounced approximately as *hemoroizi* [haemorrhoids, in Romanian]; the commercial is resounding: "*HemoroEasy* cures you when you have *hemoroizi*"!).

6. Elaboration of a [MPhSG] database

6.1. Antecedents on MPhT decrypting

A *sui-generis* opening toward deciphering the [MPhT] can be tracked down to the emergence of the first Romanian pharmacopoeia, the one published by Constantin C. Hepites in 1862. In the specialized literature, it is defined as a bilingual presentation – in Latin and Romanian –, which itself constitutes a significant step toward "democratization" in terms of communication in this pragmatic–discursive space.

Hence, compared to the traditional scholarly nomenclators, which show the prestige of Latin in the sphere of sciences, in Hepites 1862 the specialized information related to various remedies (constituting the so-called "monographs") is presented on two columns: the first (a smaller text) is in Latin, and the second is in Romanian. However, we are especially interested in the linguistic transparency, which begins in the very title of the monographs, an efficient area of the paratext: the names of the remedies are not provided only in Latin and Romanian, but also in French and German. This way, this nomenclator becomes a multilingual one, actually. We illustrate this view through the title of two remedies based on «deer antler» (the spelling is the original one):

> "***Cornu Cervi***, ***Raspatum*** – *Cornŭ de cerbŭ*, *Răsătura* – Gall. *Corne de cerf râpée*, Germ. *Hirschhorn geraspeltes*" şi "***Cornu cervi ustum*** – *Cornŭ de cerbŭ arsŭ* – Gall. *Corne de cerf brûlée*, Germ. *Weisgebranntes Hirschhorn*" (Hepites 1862: 69).

There is additional information in this sense, discovered after a minute research, regarding the animal pharmaceutical remedies, within Hepites' pharmacopoeia, present in the collection of the Museum of Pharmacy History in Sibiu. In that period, the pharmacists' interest concerned the substances used to treat diseases. For instance, "Castoreum – Castoreŭ", "Cetaceum – Spermacetŭ", "Ossa Sepiae – Ósse de sepii" (Hepites 1862: 48, 51, 126); or: "Cancrorum lapides (ochi de raci), Conchae (scoici), Fel bovinum (fiere de bou), Ichthyocolla (clei de peşte), Sebum ovillum (seu de oaie)" (Toma et al. 2012: passim).

> N.B. "Bila de bou" is still used today, and the substance called *castoreu* was registered as an antispasmodic remedy and as emmenagogue in Bianu – Glăvan 1929: s.v.

6.2. Downsides of leaflets

We have seen the low effectiveness of the presence – on the book market – of term inventories such as "1000 diseases in plain language" (cf. § 3.1.3) and of the (brief) information on the WHO norms of enciphering the "invented" common denominations, meaning of the new products (cf. § 4.1.1). These aspects make it easy to accept, in terms of consumer's protection, the idea of the necessity of elaborating, in the [MPhT field specialized nomenclators and glossaries with explanations accessible to the public. Obviously, the elaboration of such work involves many issues, some of which are rather hard to identify before ordering and applying the facts referring to: [A] elaborating the general necessary database (categories of sources, material transcription, etc.) and [B] selecting the title–words depending on the two corresponding terminological levels, in order to elaborate the list of terms to appear in a planned [MPhSG]. The task becomes even more difficult because of the lack of proper Romanian dictionaries with more or less common or even "folk" (meaning really "in plain language") terms for diseases and treatments.

In regard to the elaboration of a database that reflects the terminology belonging to the current level of the exogenous discourse, the most important aspect is that the medicine treatment for certain diseases has been evolving rapidly; hence, new terms appear all the time. On the other hand, it is difficult to establish terminological correspondences within a certain area of treatments, meant to orient toward a prospecting approach of a synonymic nature. On principle, as a rough guide, a terminological group of remedies

can be outlined starting from the core, represented by [the scientific name of the disease] + [the generic name of the remedy used for its treatment]. Around this core, an onomasiologic group can be outlined; this group comprises, from a linguistic perspective, firstly the word family of the basic term.

In this phase, we do not intent to present samples or lexicographically organized materials; however, we do provide a brief example of this working hypothesis. Around the core formed by the term *spasm* (defined by the Romanian Explicative Dictionary as an "involuntary, strong contraction, with variable duration, of muscle or of a group of muscles") and by the term *antispasmodic* ("a medicine against spasms"), the file [F] may comprise the terms *spasmodic*, *spastic*, *spasmofilie*, *spasmogen*, *spasmolitic*, and *antispastic*. This inventory will be subjected to a selection for the list of words [LW] of the [MPhSG], depending on the results of the surveys with doctors and pharmacists regarding the presence/occurrence of some of these terms in their discussions with various categories of patients and in the paratext structures represented by leaflets.

> In fact, such an approach also includes the issue of the lexicographic presentation of texts. We could consider giving up the alphabetical order of entries, in the favour of the presentation by onomasiologic groups, following the concept of "structural lexicology" formulated by Hallig – Wartburg 1963, by a "rational system of concepts", matter on which we cannot afford to discuss here. In any case, for such a lexicographic formula, an index of words solves the issue of easy orientation.

Obviously, the project and elaboration of [F] should start from basic nomenclators in the field of the description of the medicines present at a certain point in time (such as *Farmacopeea română [The Romanian Pharmacopoeia]*, elaborated under the patronage of the National Medicines Agency). It should also start from lists within the documents emitted periodically by official bodies; these documents are extremely important because they present the advantage of reflecting the product circulation. For instance, in a list comprising the international common denominations (ICD) and the common terms for the medicines available to the persons with medical insurance within the social health insurance system (emitted by the National Health Insurance House), there are around 2,000 names (in electronic format, the document has 44 p. x 47 names per page). For [F], essential criteria must be considered: medicine classes by the ATC (Anatomical–Therapeutic–Clinic) system, by the administration route, maybe even by the presentation forms, etc. Another way of assessing and enriching [F] can also be represented by the extraction of terms from the paragraphs reserved to "treatment" within complex articles such as those in Dicţionar 2013.

In regard to the selection for the [LW], the decisive element is the experience of the potential collaborators to this project, doctors and pharmacists, especially the latter. This occurs because, as easily concluded, among the patients who come to purchase medicines, a great part did not go to a doctor first; this way, often pharmacy is the setting where symptoms are described and medication recommendations are obtained. It is not less true that even the patients who come in with a prescription also choose to discuss with the pharmacist on the selection of medication, starting from common names per se, which correspond to the coded notes of doctors. In fact, the regulations in force do mention the functioning of a consultation room within the space of the pharmacy. Of course, the idea is also to ensure, if necessary, the confidentiality (in that professional setting, the concepts of "pharmaceutical care" and "confidentiality of the information" are quite common; cf. Cristea 2013).

Besides the relationship with the qualified personnel (doctors, pharmacists), the patient also has the opportunity of a direct contact with the [MPhT] that goes beyond reading the prescription (often indecipherable; this has actually become proverbial: when it is said of someone that he "has a doctor's handwriting", the idea is that the handwriting is indecipherable). We refer to the consultation of the main paratext structure: the *Leaflet* of medicines (subtitled "*Information for the user*") and, more rarely, to the consultation of the same type of structures represented by the inscriptions on packaging or even on the bottle (we take into account, as always, the priority of the most used medicines).

An analysis sample of the leaflet for the medicine called in the pragmatic–discursive space of trade SERMION 30mg (a name followed by the scientific "gloss" NICERGOLINE) is meant to clarify the difficulties of the approach. It is also meant to maintain a certain optimism – moderately, of course – regarding the practical possibilities of the project. This way, firstly, all the section titles of the leaflet are formulated in terms that are accessible to the reader with medium education; this status should also be taken into account considering the interest for deciphering the technical terms within the text of several sections. They are as follows:

- section [1] "What is Sermion 30 mg and what it is used for";
- paragraphs within section [2] "Before you take Sermion 30mg";
- section [4] "Possible adverse effects".

Those utterances represent around 25% of the entire leaflet text and they comprise a rather large number of diseases and therapies.

We mention that we do not take into account utterances that belong to the pragmatic–discursive space of an endogenous level and that are of no interest for the common reader, such as "the class of ergot alkaloid derivatives" (in section [1]), and, mostly, aberrations from the perspective of the user's competency, which we shall refer to hereinafter.

Often, in pharmacies, when the patient tries to initiate a discussion on issues within the *Leaflet* of a medicine (and for all the right reasons, considering the subtitle "Information for the user"), the pharmacist blocks the potential dialog by replying "The Leaflet is for US!" So who is right? Both interlocutors are. The patient because he is approached through such a text, not to mention the direct information formulas regarding his particular situation (for instance, for the medicine of reference, in section 2, under "The use of Sermion 30 mg during pregnancy or if you intent to become pregnant should be extremely cautious"). He is also right considering the presence of a discourse that claims to be exogenous, as the medical terms used are among the generally known ones (for instance, in the same situation, "Tell your doctor ... if you have kidney diseases...; if you have high/low blood pressure", etc.).

However, only the pharmacist can help the patient, even by simply translating utterances that clearly belong to the endogenous discourse, though, we underline again, the text is addressed to the patient. For instance, in the case of Sermion, the patient (in section 2, paragraph "Interactions") is informed of the following. "Tell your doctor if you use... anticoagulant and plaquetary antiaggregant medicines – as nicergoline inhibits platelet aggregation and reduces blood viscosity, it is necessary to frequently monitor the parameters of blood coagulation in case of the more prone patients". The utterance itself – just like many others – is grammatically incongruent. Actually, the

patients are also preoccupied by the terminology; in texts present in social networks on the Internet there are sarcastic characterizations on the names of the medicines, some of which are considered, in the ludic register, "funny" or "stupid", etc. (cf., for instance, http://e-lari.blogspot.ro/2009/12/denumiri-haioase-de-medicamente.html, as well as http://www.krossfire.ro /un-plic-de-fluimucil/).

6.3. Necessity of computer-based sources

In order to elaborate the database in question, we plan to use mainly computer-base sources, in order to permanently ensure practical operations of correlation/assessment. Furthermore, for the same reasons, we intend to edit the corpus material and the [MPhSG] in electronic format.

7. Conclusions

We believe that the above-discussed facts justify our interest for the elaboration of a database on the medico-pharmaceutical terminology representing the concrete communication possibilities characteristic to the pragmatic–discursive space of the user of medicine treatment.

In any case, the exclude the – chimerical – illusion that the communication profitability could be obtained, in this field, by training the patients to learn the codes of specialists, meaning the scientific terms for diseases and treatments. This illusion was hazardously considered possible (cf. Marin-Omer 2003), in full, but surprisingly unaware of the communication realities within a pragmatic–discursive space with the highest socio-cultural relevance.

Whether our starting point represents the result of a correct assessment and if the issue of elaborating a practical working tool of the type "Medico-pharmaceutical Security Glossary" was considered realistical, we shall find out on this occasion, of the first "declaration" of our intention. We are looking forward to and we welcome the objections and suggestions of any nature, as well as any potential criticism the specialists.

References

A. Sources

Bianu V., Glăvan I. (1929). Doctorul de casă sau Dicţionarul sănătăţii..., Ediţia a II-a revăzută şi mărită, Cluj, Institutul de Arte Grafice Cartea Românească S.A.

Dănilă Gh. (1999). Vademecum medicamentorum, Iaşi, Polirom.

Dicţionar (2013). Dicţionar medical ilustrat, vol. I-X, Bucureşti, Editura Litera Internaţional [traducere după SALUTE. *Dizionario medico*, Milano, RCS Quotidiani, 2006].

Episcupescul Ș.V. (1846). Practica doctorului de casă. Cunoştinţa apărări ş'a tămăduiri boalelor bărbăteşti, femeeşti şi copilăreşti. C-o prescurtare de hirurgie, de materie medicală şi de veterinerie, pentru doctor şi norod. În tipografia Colegiului Sf. Sava, Bucureşti.

Hepites C. (1862). Pharmacopea română, Typographia Jurnalului Naţional, XIV + 790 p., Bucuresci.

Prudhomme Ch., D'Ivernois F. (2012). 1000 de boli pe înţelesul tuturor. O enciclopedie medicală indispensabilă familiei, vol. I-II, Bucureşti, Editura Orizonturi.

Rusu V. (2010). Dicţionar medical, Ediţia a IV-a revizuită şi adăugită, Bucureşti, Editura Medicală.

B. Exegeses

Baylon Ch., Mignot X. (2000). Comunicarea, traducere de Ioana Ocneanu şi Ana Zăstroiu, Editura Universităţii „Alexandru Ioan Cuza", Iaşi.

Bourdieu P. (1976) . Le c*hamp scientifique.* „Actes de la Recherche en Sciences Sociales", vol. 2, 88-104.

Charaudeau P., Maingueneau D. (2002) . Dictionnaire d'analyse du discours, Paris, Éditions du Seuil.

Coşeriu E. (1994). Competenţa lingvistică. *Prelegeri şi conferinţe*, excerpted from The "Anuar de lingvistică şi istorie literară", XXXIII (1992–1993), 27–47.

Cristea A.N. (2013). Consilierea pacientului în farmacia de comunitate. *"Pharma Buisiness"* (www.pharma-business.ro/oportunitati/consilierea-pacientului-in-farmacia-de-comunitate.html).

Dumistrăcel S. (2000) . Paliere terminologice. "*Cronica*", XXV, no. 1, 19.

Dumistrăcel S. (2006a). Limbajul publicistic românesc din perspectiva stilurilor funcţionale, Iaşi, Institutul European.

Dumistracel S. (2006b). Discursul repetat în textul jurnalistic. Tentaţia instituirii comuniunii fatice prin mass-media, Editura Universităţii "Alexandru Ioan Cuza", Iaşi.

Dumistrăcel S., Stoica D., Dumistrăcel I. (2011a). Barrières linguistiques, notamment au niveau terminologique, dans des domaines de communication à grand impact socio-culturel. Approche pragmalinguistique. *Conférence internationale "La formation en terminologie"*.

Dumistrăcel S., Hreapcă D., Botoşineanu L. (2011b). Paliere terminologice din perspectiva barierelor lingvistice: încifrare şi transparenţă în terminologia medico-farmaceutică. *In the volume comprising the Acts of the International Conference "Paradigm of the ideological discourse. Dynamics of terminologies and (re)modelling of the systems of ideas"* (fourth edition; PID 4), Facultatea de Litere, Universitatea "Dunărea de Jos", Galaţi.

Dumistrăcel S., Hreapcă D., Botoşineanu L. (2012). Variaţie diastratică şi variaţie diafazică în comunicarea specializată: paliere terminologice. Spaţiul discursiv al publicaţiilor româneşti de instruire şi educaţie medico-sanitară (I). *In the volume comprising the Acts of the International Conference.*

Hallig R., Wartburg W. (1963). Begriffsystem als Grundlage für die Lexicographie. Versuch eines Ordnungsschemas, second edition, Berlin, Akademie Verlag.

Jurt J. (2001). La théorie du champ littéraire et l'internationalisation de la littérature. /www.freidok.uni-freiburg.de/.

Marin -Omer I. (2003). The Role of Medical Special Code and Slang. *In Communication between Doctor and Patient in Oncology Departments*, in vol. *Limba şi vorbitorii*, Bucureşti (Tatiana Slama-Cazacu ed.), Editura Arvin-Press, 272–287.

Toma E.C., Mesaros A.M., Carată A. (2012). Remedii farmaceutice de origine animală prezente în prima farmacopee română de la 1862 şi în colecţia Muzeului de Istorie a Farmaciei din Sibiu /http://www.srif.eu/fisiere/7978_JxYq_TOMA_Remedii%20origine%20 animala%202012_preg_prezentare.pdf/.

Ursu N.A. (1962). Formarea terminologiei ştiinţifice româneşti, Editura Ştiinţifică, Bucureşti.

Véron E. (1997). Entre l'épistémologie et la communication. *In "Hermes" 21, Sciences et médias*, Paris, Editions CNRS, 23–32.

C. Regulations, legislation

Ghid privind exprimarea concentraţiei în denumirea comercială a medicamentelor de uz uman (2010).

Hotărârea Consiliului Ştiinţific al Asociaţiei Naţionale a Medicamentului şi Dispozitive Medicale, nr. 2/29.02.2008: *Ghidul privind denumirea medicamentelor de uz uman.*

Id., Anexa 1: abordarea problemelor referitoare la *denumirile comune internaţionale* (DCI) în cadrul *denumirilor inventate* propuse (DI).

Ordin nr. 75/2010 pentru aprobarea Regulilor de bună practică farmaceutică.

Ordinul Ministerului Sănătăţii din 1 aprilie 2009 privind prescrierea reţetelor pe DCI.

Reglementări privind modalitatea de gestionare a propunerilor de denumiri comerciale tip „umbrelă" şi alte denumiri comerciale pentru medicamentele de uz uman în raport cu denumiri ale suplimentelor alimentare, ale produselor cosmetice şi ale dispozitivelor medicale (2012).

Regulamentul de prescriere medicamente din Contractul Cadru 2011, anexa 30.

ELECTRONIC LINGUISTIC RESOURCES FOR HISTORIC STANDARD ROMANIAN

ELENA BOIAN, SVETLANA COJOCARU, CONSTANTIN CIUBOTARU, ALEXANDRU COLESNICOV, LUDMILA MALAHOV, MIRCEA PETIC

Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, Chişinău, Republic of Moldova

lena@math.md

Abstract

This article describes digitization of old Romanian texts, problems at their recognition, and motivates the necessity to create specific electronic resources mirroring the history of the standard Romanian language. We analyze printed texts since the 16th century when the Romanian typography begins. We also provide statistics of results of recognition of documents in a Romanian text of the 19th century by modern OCR (optical character recognition) software.

Keywords: digitization, Romanian linguistic resources, text recognition, language technology

1. Introduction

The main directions of the cultural policy into zones where the Romanian language is spoken, refer to study, evaluation and digitization of cultural and historic heritage. Process of heritage digitization requires the solving of many problems that refer to recognition, editing, translation, interpretation, circulation and reception of texts printed in Romanian and other modern languages. These problems became more complicated for Romanian, as we need to consider the historic period when the source was printed, and we have several periods.

This paper presents a short description of periods of the Romanian language evolution, and aspects of the development of the main language components: alphabet, lexicon, and orthography, specific for each period. Taking into account a specific period, we will propose a technology to obtain these components. In particular, we study the problem of digitization of printed Romanian texts using different writing systems starting since the 16th century (Ivănescu, 1980).

The first book printed in the Romanian territory was the Church-Slavonic *Liturgy Book* (1508) edited by Serbian hieromonk Makarie. The first printed book in Romanian appeared in Brashov in 1535 (Panaitescu, 1965). It was *The Romanian Catechism* published by deacon Coresi.

The National Library of the Republic of Moldova possesses approximately 21,000 old and rare books. The collection contains approx. 20 books printed in Romanian in the Romanian Cyrillic and transitional scripts in Bessarabia (Chişinău and Dubăsari).

Public libraries of Sankt-Petersburg keep important quantities of old Romanian books (the 16th–19th centuries). For example, there are 66 titles in The Catalog of Cyrillic editions of Southern Slavonians and Romanians. 45 volumes are of the Southern Slavonian origin, while 21 can be attributed to Romanian lands (Valori, 2008).

In its history, the Romanian language has passed through a long and rich evolution. The existent studies explain appearance of each vowel and consonant at each specific stage of the language evolution that is necessary to determine the alphabet and specific letters (Ivănescu, 1980; Munteanu & Țâra, 1978). This information permits us to construct linguistic resources and to use specific tools for a specific period of the language history.

Our work is a long-term project that is in its beginning now. We implement it using the principle "from now into the depths of time".

In this paper we describe our approach to digitization of Romanian texts from the 20th century and back until the 19th. Three types of texts can be selected:

1. Moldavian Cyrillic script that was used in 1924–1989, and is used now in Transnistria;
2. Latin script with additional letters, different depending on period;
3. Transitional script.

We performed this categorization based on the alphabet. We should note that each of these periods can be subdivided on the basis of the corresponding orthography and lexicon.

The structure if the paper is as follows: state-of-the-art in old text recognition, with orientation to South-Eastern Europe (Sec. 2); a short list of the historic periods of the Romanian language and script evolution (Sec. 3–4); exposition of techniques to digitize and to recognize printed texts (Sec. 5); examples and considerations on recognizing texts from specific periods (Sec. 6).

2. State-of-the-arts in working with historical texts of South-Eastern Europe

The problem of digitization and preservation of historical linguistic heritage is a domain of priority in the digital agenda for Europe. The EU highlights the necessity for coordinated effort in the domain, and manifests vast actions to activate this process. These actions include development of the *Europeana* virtual library supported by a resolution of the European Parliament of May 5, 2010, and by adopting the Work Plan for Culture 2011–2014. Let us mention also the European Commission Recommendation on the digitization and online accessibility of cultural material and digital preservation of October 27, 2011.

For Romanian historical linguistic heritage, the solution of this problem presents specific difficulties: a large number of periods in the language evolution; relatively small number and big dispersion of deposited resources; big variety of used alphabets, in particular, several so-called "transitional" (mixed Latin-Cyrillic) alphabets. The difficulties in digitization and preservation of this heritage lie in correct recognition of

characters and in lack of adequate lexicons corresponding to the periods of the texts printing. One of solutions of the lexicon problem could be aligning of old texts to contemporary linguistic norms (Moruz & al., 2012).

As to OCR of printed and handwritten Cyrillic characters, we can mention a paper (Kornienko & al., 2011) where both standard ABBYY FineReader and AI techniques are used, in particular, artificial neural networks. There exists an application of methods based on knowledge technologies to the digital archive and multimedia library for Bulgarian traditional culture and folklore (Pavlov & al., 2011). Problems of transliteration caused by parallel use of two alphabets, Cyrillic and Latin, which appear at processing of written texts in modern Serbian, were solved applying monolingual and multilingual corpora and various e-dictionaries (Vitas & al., 2003).

3. Periods of evolution of the Romanian language

The history of the Romanian language contains two epochs of its evolution. The first one is that of formation of the Daco-Romanian dialect and continues since the taking of Sarmizegetusa (106 A.D.) until the 15th century (Ivănescu, 1980). The Cyrillic alphabet was used in the end of the epoch because of the Orthodox Church domination.

The second epoch (16th–20th centuries) of the evolution of standard Romanian begins since the appearance of the first texts written in Romanian as the result of a long and complex development (Munteanu & Țâra, 1978). This second epoch can be divided in two big stages.

The first stage begins since the appearance of the first Romanian literature texts, and ends in the beginning of the 18th century. This stage can be subdivided in three periods:

1. 1532–1588, the first steps in language standardization;
2. 1588–1656, consolidation of the main variants of standard Romanian (Muntenian, Moldavian, and South-West-Ardealian);
3. 1656–1715, mutual influence of variants.

In 1688 *Biblia de la Bucureşti* [*the Bible of Bucharest*] appeared. Its publication became a milestone in the linguistic unification that led to the second stage of the second epoch (Gheție, 1978). This second stage covers 1715–1960 and consolidates a unified over-dialectal language. We can subdivide this stage in four periods:

1. 1715–1780, the first unification, approx. at 1750;
2. 1780–1836, linguistic diversification;
3. 1836–1881, stabilization of main norms of the unified standard language;
4. 1881–1960, fixing of norms of the modern standard Romanian language.

The last period signifies also stylistic consolidation of standard Romanian. In 1904, the orthography was changed to be definitively based on the phonetic principle that is kept for standard Romanian till now, with some further refinements.

4. Periods of Romanian scripts development

In the 17th century, a Romanian Cyrillic script had appeared, with up to 47 letters. Most letters were taken from Old Church Slavonic. Several Greek letters were added to convey names exactly. An original Romanian letter  was used as prefix or preposition **în**, **îm** (**in**), or as the modern letter **î** in the beginning of words. *Varlaam's Homiliary* was printed in 1643 with this script (Fig. 1). The first Romanian ABC book was printed in Bălgrad (Alba Iulia) in 1699, and in 1757 D. Evstatievich published a Romanian grammar.

In the 18th century Romanian belle letters appeared.

Since 1830, until the official adoption of Romanian Latin-based alphabet in 1862, the script was not regulated thoroughly, and at least seven modifications of so-called "Transitional alphabets" mixed from Cyrillic and Latin letters were used (Fig. 4, 7). Forexample, **e** - **є** (1830) - **ε** (1846); **к** - **k**; **щ** - **шt**; **s** - **дз** – **dz** - **ḑ** (1846).

Usage of the Latin-based script in Romania had not influenced the typography practice in Bessarabia.

After the ceding of Bessarabia to imperial Russia in 1812, the official language has migrated to Russian. In 1833, Romanian was excluded from all official communications but remained in eparchial administration until 1873. The church typography in Chisinau was closed in 1883, and reopened in 1906. Except church books, we can also mention: ABC books, 1814 and later, 1861, 1863; a booklet on emancipation of serfs, 1861; calendars, etc. Instructive booklets on agriculture and hygiene in Romanian were published and distributed by local authorities. In 1867-1871, the Romanian version of Chisinau Eparchial Gazette was printed in civil Slavonic script with several traditional letters and γ-like **u** (**ȣ**). In several cases, transitional (1859) and even Latin-based script were used (Ciobanu, 1923).

In the 1880–1890, the printing in the Romanian language was ceased in Bessarabia, resuming at the beginning of the 20th century. The religious printing used both church and civil scripts.

It is necessary to distinguish the Romanian Cyrillic alphabet and the Moldavian Cyrillic alphabet (Fig. 8). The former was used for Romanian writing since the 14th–15th centuries until 1862. The latter is, in fact, an adaptation of the Russian Cyrillic alphabet to reproduce the Romanian phonetics by Russian orthographical norms that led to some weird orthographical effects.

This second variant based on the Russian alphabet was used in the Moldavian Autonomous Soviet Socialist Republic (MASSR) in 1930–1932 and 1938–1940, then in the Moldavian Soviet Socialist Republic (MSSR) since its formation in 1940, and until 1989. This alphabet is still used in Transnistria. Between 1932–1938, the Latin-based alphabet was used in the MASSR.

We can therefore exhibit the following periods in the development of the Romanian script since the publication of *Varlaam's Homiliary* (Tab. 1).

Table 1: Development of Romanian script since 1642

Romania	**Bessarabia**
1642 – 1710 (Romanian Cyrillic script)	
1710 – 1830 (modified Romanian Cyrillic script)	1710 – 1814 (modified Romanian Cyrillic script)
1830 – 1862 (mixed Cyrillic-Latin transitional script)	1814 – 1880 (Cyrillic scripts based on Russian civil and Old Church Slavonic scripts; occasionally, transitional and Latin-based script)
1862 – 1904 (Latin-based script)	1880 – 1905 (No Romanian typography) 1905 – 1918 (Cyrillic script based on Russian civil script)
1904 – 1960 (modified Latin-based script)	1919 – 1940, 1941 – 1944 (modified Latin-based script) 1940 – 1941 (Moldavian Cyrillic script) [See above in the text on situation in the MASSR]
1960 – 1993 (modified Latin-based script)	1944 – 1989 (Moldavian Cyrillic script; in 1967 letter ӂ appeared)
1993 – now (modern Romanian Latin-based script)	1989 – now (modern Romanian Latin-based script) [See above in the text on the situation in Transnistria]

There are more factors except of script, which characterize periods of language development. They are also orthography and lexicon.

We show in Fig. 1–8 examples of printed texts at different periods of the language evolution.

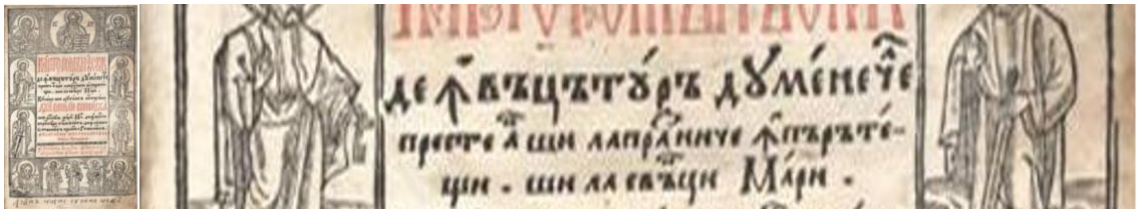


Figure 1: Varlaam's Homiliary, Iași,1643

"Romanian book of learning during the year and at the Christian feasts, and of the Great Saints. Under the order and all costs paid by Vasilie [Lupu], Prince and Ruler of Moldavia, compiled and translated from many sources, from Slavonic into Romanian, by Varlaam the Metropolitan of Moldavia. At Ruler's typography."

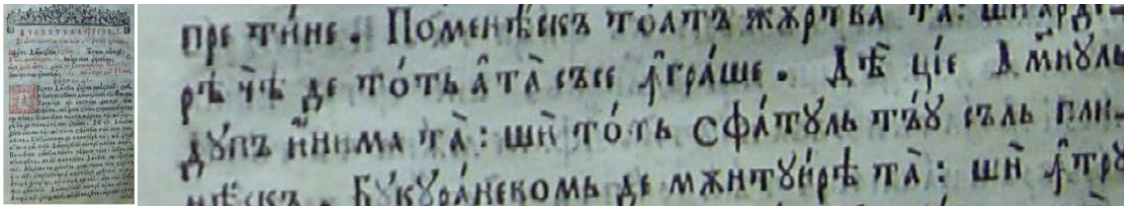


Figure 2: Horologion, 1748

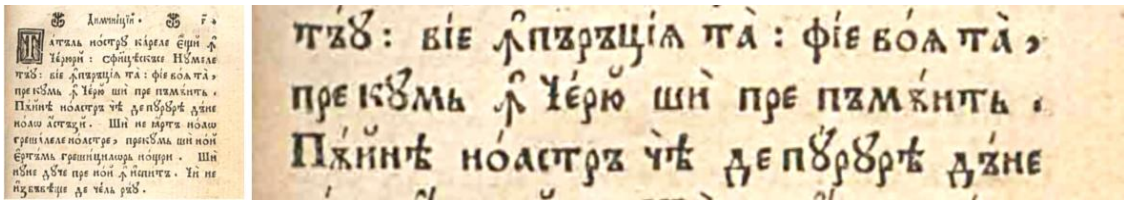
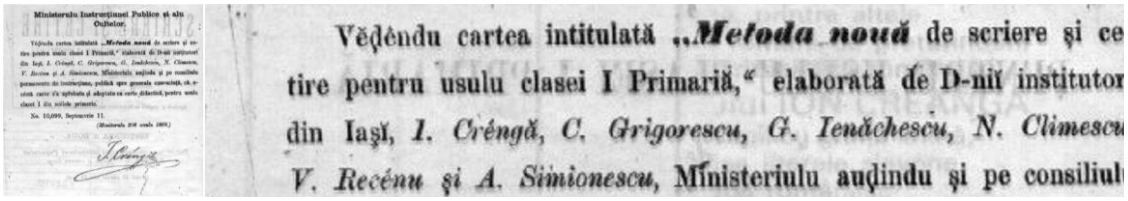


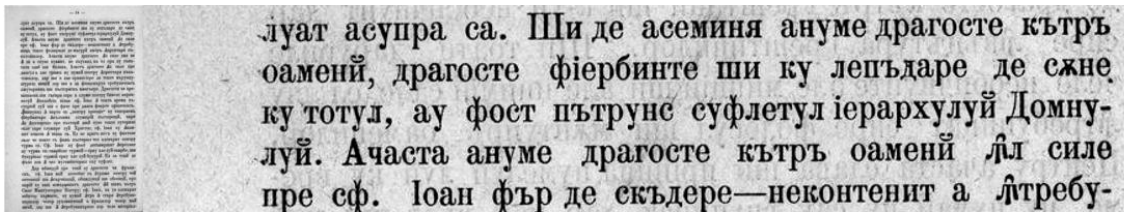
Figure 3: Lord's Prayer. In: Book of Akathists with many selected prayers for humbleness of each Christian, Printed in the third time. Blaj: Typography at the Theological School, 1786

Figure 4: Chronicles of the State of Moldova published for the first time ever by Mihail Kogălniceanu. Volume I. Iaşii. Available in all bookstores. 1852



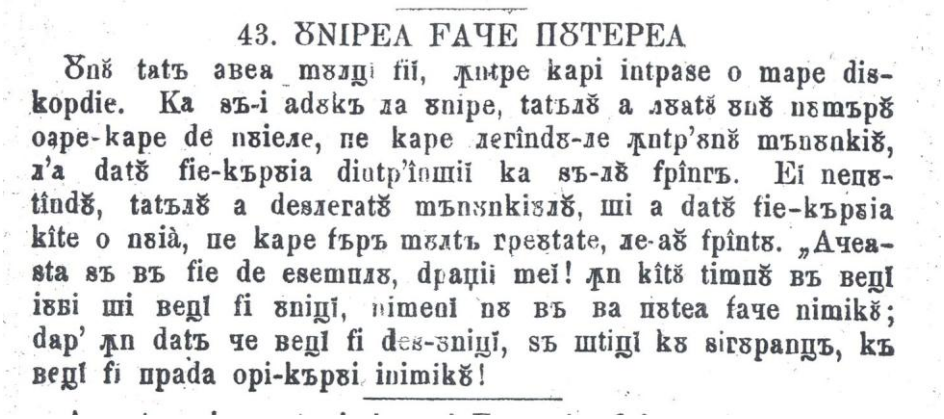
Vĕḍèndu cartea intitulată „*Metoda nouă* de scriere şi ce
tire pentru usulu clasei I Primariă,“ elaborată de D-nii institutor
din Iaşi, *I. Créngă, C. Grigorescu, G. Ienăchescu, N. Climescu*
V. Recénu şi A. Simionescu, Ministeriulu auḍindu şi pe consiliul

Figure 5: A new method of writing and reading: For the 1st year at primary school / I. Creangă , C.Grigorescu, G. Ienăchescu, 2nd ed. – Iassy: H. Goldner's Tipography, 1868. – 71 p



луат асупра са. Ши де асеминя ануме драгосте кътръ
оамений, драгосте фіербинте ши ку лепъдаре де сѫне
ку тотул, ау фост пътрунс суфлетул іерархулуй Домну-
луй. Ачаста ануме драгосте кътръ оамений л̆л силе
пре сф. Іоан фър де скъдере—неконтенит а л̆трeбу-

Figure 6: A page from magazine "Luminătorul" ["Enlightener"], 1908, Nr. 1



43. ѲNIPEA FAЧE ПѲTEPEA

Ѳnѳ tatъ авеа мѳлџі fiї, л̆nрe kapi intpase o mape dis-
kopdie. Ka въ-i adѳkъ ла вnipe, tatьлѳ a лѳatѳ вnѳ пѳmърѳ
oape-kape de пѳieле, пe kape лeгîndѳ-ле л̆ntp'вnѳ mънѳnkiѳ,
л'a datѳ fie-kърѳia diutp'înшii ka въ-лѳ fpînгъ. Ei пeпѳ-
tîndѳ, tatьлѳ a deѳлeratѳ mънѳnkiѳлѳ, ші a datѳ fie-kърѳia
kîte o пѳià, пe kape fъръ мѳлтъ гpeѳtate, ле-аѳ fpîntѳ. „Ачea-
sta въ въ fie de eѳemплѳ, dpaџii meї! л̆n kîtѳ timnѳ въ вeџl
iѳві ші вeџl fi вniџї, nimeol nѳ въ ва пѳtea faче nimikѳ;
dap' л̆n datъ че вeџl fi deѳ-вniџї, съ штиџї kѳ вігѳpaпдъ, kъ
вeџl fi пpada opi-kърві inimikѳ!

Figure 7: One of variants of the transitional alphabet from ABC book by I. Creangă

Ш'ачел реже-ал поезией, вечник тынэр ши фериче,
Че дин фрунзе ыць дойнеште, че ку флуерул ыць зиче…

Figure 8: A text printed in Moldavian Cyrillic alphabet (1967–1989) used till now in Transnistria. From: M. Eminescu, "Epigones"

5. Recognition of characters in printed texts

Manuscript digitisation and recognition is complicated because it requires additional operations, such as adjusting the contrast, cleaning the image, text segmentation. We also need to develop special algorithms of recognition and specialized lexicons. Further, we only take into account Romanian texts printed with Latin letters.

Process of digitisation and recognition consists of the following stages:

- Digitization of the text resulting in its graphical electronic copy.
- Recognition by standard techniques, namely, using OCR (Optical Character Recognition) (OCR) software, possibly, with its training. Without OCR, procedures of conversion using Artificial Intelligence techniques should be applied. Transliteration of the text is performed taking into account specific letters from the initial text.
- Verification of the recognised text is performed using reusable resources specialized for the corresponding period.

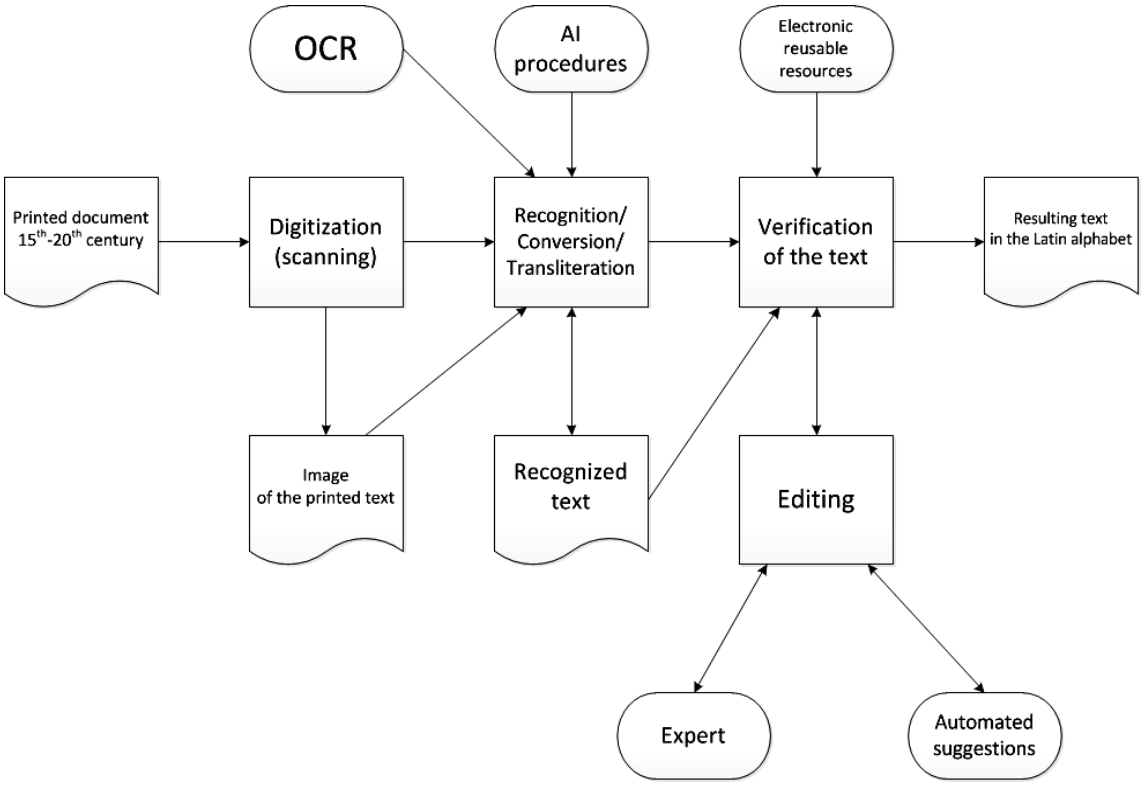


Figure 9: Technological stages of printed text recognition

Digitizing texts is their scanning and obtaining their electronic version as an image. OCR is used to recognize text from its image. Standard OCR systems use different methods to recognize texts.

We tried two systems: IRIS and ABBYY FineReader. Results of experiments in recognition of a printed 19th-century text are exposed in Section 6. We found that IRIS does not offer the possibility to select an arbitrary fragment of image during training. Therefore, we cannot correct the fragmentation proposed by the system. This system does not satisfy our purposes, as it is impossible to train it to recognize old printed Romanian text.

The ABBYY FineReader OCR system allowed us to adapt it for the alphabet of a corresponding period. We trained the system by enlarging the alphabet. It should be noted that OCR systems recognize the actual text if its internal spelling checker uses lexical resources that corresponds to the historical period of text. The OCR systems using standard (modern) lexicons do not always obtain a satisfactory result. To improve the results, we need further processing of the scanned text.

Pattern recognition techniques are used to identify individual characters in the page, including punctuation, spaces and end of lines. The recognized text appears as an editable file.

Transliteration is a strictly individual process that is dependant on the examined period. It uses programs that depend on the initial text and contain information on specific letters in that text. Transliteration supposes creation of bidirectional relations between two systems of writing considering that a specialist could reconstruct the original text from its transliterated variant. Transliteration should be performed only as necessary.

Text verification is performed by a special application that uses specific resources for the historic period of the printed text (Burlaca & al., 2010). Newly obtained words can be entered into the corresponding lexicon.

6. Results of experiments in recognition of printed 19th-century texts

6.1. Processing of texts in the Moldavian Cyrillic script

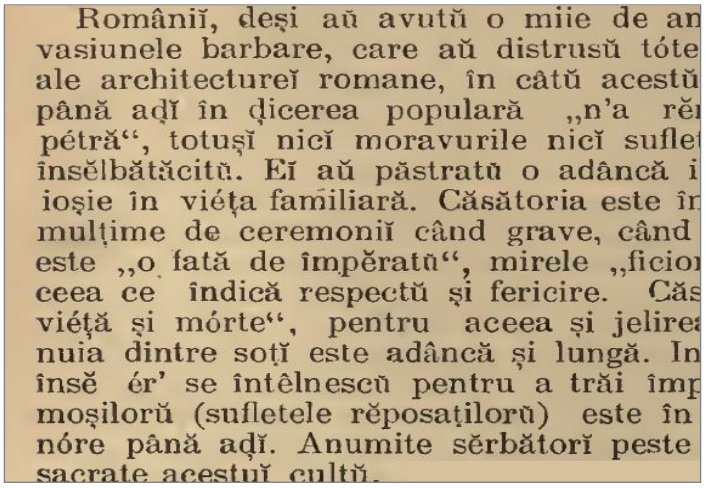
To perform OCR (Fig. 9) of such texts, it is necessary to train the OCR system to recognize an additional letter ѫ (since 1967), and to provide the corresponding lexicon. For the end of the period (1951–1989), we can obtain the dictionary transliterating the modern Romanian dictionary in the Latin script. The transliteration is not simple because of several irregularities in this system of writing, e.g.:

- absence of **и** (**i**) in the Cyrillic equivalent of words like **pâine** (**пыне** – **bread**), **câine** (**кыне** – **dog**); in other words containing diphthongs, this letter is kept: **cârâitor** (**кырыитор** – **croaking**), **târâitură** (**тырыитурэ** –**creeping**);
- replacement of **а** with **я** (instead of **а**) in words like **funcția** (**функция** – **function**); in other words the diphthongs **ia** is transliterated as **я**: **boia** (**боя** – **color**), see the next point;
- representation of **ea** and **ia** as **я**, with a single exception of pronoun **ea** (she); at the same time, the verb **ia** (a derived form of **lua** – **take**) was written as **я**;
- replacement of **i** with three different letters (**и**, **й**, **ь**), etc.

The initial period of Moldavian Cyrillic writing (1924–1951) is associated with an extremely specific lexicon. It is characterized by:

- use of Russian words, for example, **совет**, **указ**, **словарь** (council, decree, dictionary) instead of their Romanian equivalents (**consiliu**, **decret**, **dicționar**);
- deletion of Romanian neologisms that were claimed as "bourgeois";
- fixing of local (Transnistrian) lexicon;
- introduction of self-invented neologisms for abstract notions that cannot be found in the language of Bessarabian countrymen, for example, **амувремник** (**amuvremnic** = **contemporary**), instead of **contemporan**; the word was constructed from dialectal forms of words **now** and **time** with an adjectival suffix;
- fixing of peculiarities of local (Transnistrian) accent, like **ди** (**di)** instead of **de** (preposition), **мержи** (**merji**) instead of **merge** (go), **сунити** (**suniti**) instead of **sunete** (sounds), **кы** (**cî**) instead of **că** (how), etc.

We need to create several specific lexicons for this type of writing, reflecting dictionary and orthography of sub-periods: 1924–1932; 1938–1940; 1945–1951. We need one or two more for 1951–(1967)–1989.



Românii, deşi aŭ avutŭ o miie de an
vasiunele barbare, care aŭ distrusŭ tóte
ale architectureĭ romane, în câtŭ acestŭ
până aḑĭ în ḑicerea populară „n'a rĕ
pétră", totuşĭ nicĭ moravurile nicĭ suflet
însĕlbătăcitŭ. Eĭ aŭ păstratŭ o adâncă i
ioşie în viéţa familiară. Căsătoria este în
mulţime de ceremoniĭ când grave, când
este „o fată de împĕratŭ", mirele „ficio
ceea ce îndică respectŭ şi fericire. Căs
viéţă şi mórte", pentru aceea şi jelire
nuia dintre soţĭ este adâncă şi lungă. In
însĕ ér' se întêlnescŭ pentru a trăi împ
moşilorŭ (sufletele rĕposaţilorŭ) este în
nóre până aḑĭ. Anumite sĕrbătorĭ peste
sacrate acestuĭ cultŭ.

Figure 10: Digitized text, 1984 (Densușianu, 1984, p. 130)

> Românîĭ, deşi au avută o miie de an! se sufere in vasiunele barbare, care au distrusă tote operele măreţe ale architectureT romane, în cătu acesta faptă a rămasă până adt în dicerea populară "n 'a rămasă petră pe petră", totuşt nici moravuri1e niel sufletuln loru nu s'a tnsălbătăcitu. Ei au păstrate o adâncă intimitate şi doioşie in vieţa familiară. Căsătoria este încungiurată de-o mulţime de ceremonil când grave, când vesele. Miresa este "o, fată de tmpăratn", mirele "ficioru de împărată", ceea ce indică respectă şi fericire. Căsătoria este "pe vieţă şi morte", pentru aceea şi jelirea la m6rtea unuia dintre sot} este adâncă şi lungă. In ceealaltă lume tnsă er' se tntelnescn pentru a trăi împreună. Cultulă moşiloră (sufletele răposatilorn) este în forte mare 0- n6re până adt. Anumite sărbătort peste anu suntu consacrate acestui cultă.

Figure 11: Text recognized with OCR system IRIS

The next step was manual correction of the text from Figure 11 resulting in the text shown on Figure 12. Words in old writing are underlined.

> Româniĭ, deşi aŭ avutŭ o miie de anĭ se sufere invasiunele barbare, care aŭ distrusŭ tóte operele măreţe ale architectureĭ romane, în câtŭ acestŭ faptŭ a rémasŭ până aḑĭ în ḑicerea populară „n'a remasŭ pétră pe pétră", totuşĭ nicĭ moravurile nicĭ sufletulŭ lorŭ nu s'a însélbătăcitŭ. Eĭ aŭ păstratŭ o adâncă intimitate şi doioşie în viéţa familiară. Căsătoria este încungiurată de-o mulţime de ceremoniĭ când grave, când vesele. Mirésa este „o fată de împératŭ", mirele „ficiorŭ de împératŭ", ceea ce îndică respectŭ şi fericire. Căsătoria este „pe viéţă şi mórte", pentru aceea şi jelirea la mórtea unuia dintre soţĭ este adâncă şi lungă. În ceealaltă lume însé ér' se întélnescŭ pentru a trăi împreună. Cultulŭ moşilorŭ (sufletele réposaţilorŭ) este în fórte mare onóre până aḑĭ. Anumite sérbătorĭ peste anŭ suntŭ consacrate acestuĭ cultŭ.

Figure 12: Manual correction of the text

6.2. Processing of texts in the Latin script with additional letters

To illustrate the described technology we will investigate recognition and verification of digitized text from (Densușianu, 1984) that was published in 1894 (Fig. 10).

The text on Fig. 10 was recognized with the OCR system IRIS with Romanian mode that uses modern lexicon.

As we compare the resulting (Fig. 11) and source (Fig. 10) texts we see that unrecognized words are those written in the old orthography with letters specific for the 19th century. For example, we got **tnsălbătăcitu** instead of **însélbătăcitŭ**.

This result cannot be improved, because IRIS in its training mode does not permit arbitrary fragmentation of image fixing its own fragmentation.

The use of modern lexicon lead, for example, in recognizing of **avutŭ** as **avută**, while the right word is **avut** in this context. Words from the 19th-century lexicon were not recognized because we need for their correct recognition dictionaries specific for the corresponding period that, in our case, would contain words like **remasŭ**, **viéţa**, **împêratŭ**, etc.

Underlined word in Figure 11 are those erroneous or written differently comparing with the modern Romanian language.

> *Româniĭ, deşi *aŭ *avutŭ o *miie de *anĭ se sufere *invasiunele barbare, care *aŭ *distrusŭ *tóte operele măreţe ale *architectureĭ romane, în *câtŭ *acestŭ *faptŭ a *rémasŭ până *ad̨ĭ în *dicerea populară „n'a *remasŭ *pétră pe *pétră", *totuşĭ *nicĭ moravurile *nicĭ *sufletulŭ *lorŭ nu s'a *însélbătăcitŭ. *Eĭ *aŭ *păstratŭ o adâncă intimitate şi *doioşie în *viéţa familiară. Căsătoria este *încungiurată de-o mulţime de *ceremoniĭ când grave, când vesele. *Mirésa este „o fată de *împératŭ", mirele „*ficiorŭ de *împératŭ", ceea ce *îndică *respectŭ şi fericire. Căsătoria este „pe *viéţă şi *mórte", pentru aceea şi jelirea la *mórtea unuia dintre *soţĭ este adâncă şi lungă. În *ceealaltă lume *însé *ér' se *întélnescŭ pentru a trăi împreună. *Cultulŭ *moşilorŭ (sufletele *réposaţilorŭ) este în *fórte mare *onóre până *ad̨ĭ. Anumite *sérbătorĭ peste *anŭ *suntŭ consacrate *acestuĭ *cultŭ.

Figure 13: Text checked with RomSP

The corrected text was checked with RomSp spelling checker (Burlaca & al., 2010) with the lexicon of approx. 1 million words of modern Romanian (Fig. 13). An asterisk * marks words not understood by the spelling checker that can be attributed as belonging to the 19th-century lexicon.

The source text in Fig. 10 contains 130 words. 57% of words were found correct but 43% were suspicious. The "correct" words are those whose writing was kept intact since the 19th century, for example: **sufere**, **acesta**, **fericire**. "Suspicious" words are those affected by the changes in orthography, for example: **ceealaltă** (***cealaltă***), **doioşie** (***duioşie***), **miie** (***mie***), **avutŭ** (***avut***), **aḑĭ** (***azi***). It is seen that only part of "old" words contains specific letters.

To recognize the text correctly, we need to train the OCR system to recognize specific letters and to add into the lexicon a set of new words specific for the 19th century, for example: **avutŭ, miie, nicĭ, doioşie, viéţa, ficiorŭ**, etc.

The OCR system ABBYY Fine Reader has more elaborated features of training. We used this system to perform another experience with the same text from Fig. 10. The system recognizes the whole Unicode set of letters in many font faces. The user can select any subset of Unicode as a "user-defined language", adding to it his own lexicon (list of words). In rare cases, a "real" training over images of letters can be necessary but we had not used it in this case. First of all, we instructed the system to include as recognizable specific letters for 19th-century Romanian:

- **ŭ** (a final letter, can be mute or pronounced),
- **é** (is pronounced as diphthong **ea**),
- **ó** (is pronounced as diphthong **oa**),
- **ḑ** (is pronounced as **z** or **dz**),
- **ĭ** (**i** is written now, with special rules of pronunciation),
- **ê** (is used as **â**).

The resulting text is shown in Fig. 14 (accuracy of 63%).

> Româniĭ, deşi aŭ avutŭ o miie de anĭ se sufere in- vasiunele barbare, care aŭ distrusŭ tóte operele măreţe ale architectureĭ romane, în câtŭ acestŭ faptŭ a rémasŭ până adĭ în cjicerea populară „n'a remasŭ pétră pe pétră", totuşĭ nicĭ moravurile nicĭ sufletulŭ lorŭ nu s'a însélbătăcitŭ. Eĭ aŭ păstratŭ o adâncă intimitate şi do- ioşie în viéţa familiară. Căsătoria este încungiurată de-o mulţime de eeremoniĭ când grave, când vesele. Mirésa este „o iată de împératŭ", mirele „ficiorŭ de împératŭ", ceea ce îndică respectŭ şi fericire. Căsăteria este „pe viéţă şi mórte", pentru aceea şi jelirea la mórtea u- nuia dintre soţĭ este adâncă şi lungă. In ceealaltă lume însé ér' se întélnescŭ pentru a trăi împreună. Cultulŭ moşilorŭ (sufletele réposaţilorŭ) este în fórte mare o- nóre până a^ĭ. Anumite sérbătorĭ peste anŭ suntŭ con- sacrate acestuĭ cultŭ.

Figure 14: Text recognized with ABBYY Fine Reader set for the 19th-century alphabet, without spell checking

As the next step, the system was equipped with a dictionary containing words marked in Fig. 12, namely, those that do not exist in the modern lexicon. This lexicon was set as the additional one to the modern Romanian lexicon. This time ABBY Fine Reader recognized the source image (Fig. 10) with the accuracy of 98% correct words and 2% of suspicious words (Fig. 15). It is seen that most of "bad" words were not recognized because of poor image quality (**adĭ**, **cjicerea, a(Jĭ**)). Comparing Fig. 14 and Fig. 15, we see that even hyphenated words were recognized correctly.

> Româniĭ, deşi aŭ avutŭ o miie de anĭ se sufere invasiunele barbare, care aŭ distrusŭ tóte operele măreţe ale architectureĭ romane, în câtŭ acestŭ faptŭ a rémasŭ până **adĭ** în **cjicerea** populară „n'a remasŭ pétră pe pétră" totuşĭ nicĭ moravurile nicĭ sufletulŭ lorŭ nu s'a însélbătăcitŭ. Eĭ aŭ păstratŭ o adâncă intimitate şi doioşie în viéţa familiară. Căsătoria este încungiurată de-o mulţime de ceremoniĭ când grave, când vesele. Mirésa este „o fată de

> împératŭ‟, mirele „ficiorŭ de împératŭ‟, ceea ce îndică respectŭ şi fericire. Căsătoria este „pe viéţă şi mórte‟, pentru aceea şi jelirea la mórtea unuia dintre soţĭ este adâncă şi lungă. în ceealaltă lume însé ér’ se întélnescŭ pentru a trăi împreună. Cultulŭ moşilorŭ (sufletele réposaţilorŭ) este în fórte mare onóre până **a(Jĭ**. Anumite sérbătorĭ peste anŭ suntŭ consacrate acestuĭ cultŭ.

Figure 15: Text recognized with ABBYY Fine Reader set for the 19th-century alphabet, with spell checking and an additional dictionary (accuracy 98%)

Thus equipped, FineReader was used to recognize another five pages from the same source (Densușianu, 1984), and, later, for pages from another book of the same period. We sum the results in Tab. 2. The errors can be attributed to the absence of words in the lexicon, or to the poor image quality.

Table 2: Results of experiments in OCR of 19th–century texts

Mode of recognition	Correct words	Suspicious words
IRIS	57%	43%
ABBYY FR, no training	63%	37%
ABBYY FR, trained, dictionary’s source page	98%	2%
ABBYY FR, trained, more pages, the same book	95%	5%
ABBYY FR, trained, pages from another book	95.4%	4.6%

If we want to obtain better results at the verification of printed text, we need that for the corresponding historic period:

- the scanner (scanning software) would be trained to recognize specific characters,
- a lexicon of words used in the specific period would be composed.

6.3. Processing of texts in the transitional script

There are at least seven versions of transitional (mixed Cyrillic and Latin) script. Most of the letters of this script can be recognized with ABBYY Fine Reader by forming the “language” from the corresponding Unicode glyphs. Only one specific Romanian Cyrillic letter is absent in the Unicode. It is necessary to include in the language its letter equivalent (linguists simply use an arrow ; we may use, for example, Slavonic *yus* **Ѧѧ**), and to train the system over its graphical forms in different font faces. We experimented with the text from Fig. 7, with accuracy of 93.2%. With a small volume of training material and poor scan quality, this is a quite a good result.

7. Conclusions

Digitized resources are specific records that are kept in the database accessible through Internet. To ease the access to these resources for users, it is necessary to develop interfaces and a special technology that allows text recognition.

Our technology is oriented to solve, for each period of the language development, two main problems: 1) development of algorithms to recognize alphabets of a specific period; and 2) development of tools and interfaces needed to create the corresponding linguistic resources (lexicons). This would permit to recognize words and to align texts conforming to contemporary linguistic norms.

As we move from one period to another, we can use previously elaborated tools and resources, thus implementing the principle "from now into the depths of time".

The proposed technology can be used in the formation and completion of specific linguistic resources with new words extracted from digitized materials and certified by language experts. It would allow construction of parallel corpora of different nature. Development of the proposed technology would provide opportunities to transliterate digitized text into modern Romanian, to customize graphics, to offer possibilities for corpora building, to preserve the original texts.

Specific electronic resources can be placed on the Internet for public access contributing to the development of the informational communicative media for the Romanian language. Moreover, these resources constitute an essential support for researchers, and conversions into modern standard text can be used as didactic materials at teaching.

References

Burlaca, O., Ciubotaru, C., Cojocaru, S., Colesnicov, A., Magariu, G., Malahov, L., Petic, M., Verlan, T. (2010). Applications based on reusable linguistic resources. *Multilinguality and interoperability in language processing with emphasis on Romanian*, 461–476.

Cartea Moldovei (sec XVII – înc. sec XX). (1992). Ediţii cu caractere chirilice (sec XVII – înc. sec XX). Catalog general. Chişinău,. [Moldavian Books (XVII–beg.XX cen.). Editions in Cyrillic Script (XVII–beg.XX cen.). General Catalog. Chişinău, 1992. – In Romanian.]

Ciobanu, Ş. (1923). Cultura românească în Basarabia sub stăpânirea rusă. [Ciobanu, Ş. Romanian culture in Bessarabia under Russian rule. Chişinău, 1923. – In Romanian.**]**

http://www.scribd.com/doc/75147025/%C5%9Etefan-Ciobanu-Cultura-romaneasc%C4%83-in-Basarabia-sub-st%C4%83panirea-rus%C4%83-1923

Densuşianu, A. (1894). Istoria limbii şi literaturii române. Iaşi. *[*Densuşianu, A. History of the Romanian language and literature. Iaşi, 1894. – In Romanian.]

http://ru.scribd.com/doc/123035210/Istoria-limbii-si-literaturii-romane

Gheţie, I. (1978). Istoria limbii române literare. Bucureşti. [Gheţie I. History of the standard Romanian language. Bucharest, 1978. – In Romanian.]

Ivănescu, G. (1980). Istoria limbii romane. Iaşi. [Ivănescu, G. History of the Romanian language. Iaşi, 1980. – In Romanian.]

Корниенко С.И., Айдаров Ю.Р., Гагарина Д.А., Черепанов Ф.М., Ясницкий Л.Н. (2011). Программный комплекс для распознавания рукописных и старопечатных текстов. *Информационные ресурсы России*, №1, с. 35–37. [Kornienko S.I. et al. Program tools for recognition of handwritten and old-printed texts. *Informational Resources of Russia*, 2011, nr. 1, p. 35–37. – In Russian.]

Moruz, M., Iftene, A., Moruz, A., Cristea, D. (2012). Semi-automatic alignment of old Romanian words using lexicons. *Proceedings of the 8-th International Conference „Linguistic resources and tools for processing of the Romanian language"*, Iași, Editura Universității „A.I. Cuza", 119-125.

Munteanu, Ş., Ţâra, V. (1978) Istoria limbii române literare. Editura Didactică şi Pedagogică, Bucureşti. [Munteanu, Ş., Ţâra, V. History of the standard Romanian language. Editura Didactică şi Pedagogică, Bucharest, 1978. – In Romanian.]

OCR (Optical Character Recognition) Technology

http://www.unescap.org/stat/pop-it/pop-guide/capture_ch01.pdf

Panaitescu, P. (1965). Începuturile şi biruinţa scrisului în limba română, Bucureşti. [Panaitescu, P. The beginning and the victory of the Romanian writing. Bucureşti, 1965. – In Romanian.**]**

Pavlov, R., Bogdanova, G., Paneva-Marinova, D., Todorov, T., Rangochev, K. (2011). Digital archive and multimedia library for Bulgarian traditional culture and folklore. *International Journal "Information Theories and Applications"*, Vol. 18, Number 3, 276–288.

RRRL: Reusable Resources for the Romanian Language: http://www.math.md/elrr/

Valori Bibliofile-2008. *Gazeta bibliotecarului*, Iunie-Iulie 2008, nr. 6-7, p. 1. [Bibliophile Values-2008. *Librarian's Gazette*, June-July 2008, nr. 6-7, p. 1. – In Romanian.]

http://87.248.191.115/bnrm/publicatii/files/3/93.pdf

Vitas, D., Krstev, C., Obradović, I., Popović, L., Pavlović-Lažetić, G. (2003). An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts.

http://poincare.matf.bg.ac.rs/~cvetana/biblio/Solun03MATF.pdf

CLRE – PARTIAL RESULTS IN THE DEVELOPMENT OF A ROMANIAN LEXICOGRAPHIC CORPUS

MĂDĂLIN IONEL PATRAŞCU, ELENA TAMBA, MARIUS-RADU CLIM, ANA-VERONICA CATANĂ-SPENCHIU

[1]*The Romanian Academy, "A. Philippide" Institute of Romanian Philology,*

Iaşi Branch – Romania
[2] *"Alexandru Ioan Cuza" University of Iaşi, Faculty of IComputer Science,*

Iaşi – Romania

ionel.patrascu@info.uaic.ro, isabelle.tamba@gmail.com, marius.clim@gmail.com, anaspenchiu@gmail.com

Abstract

The aim of this paper is to point out the current status of creating an essential Romanian lexicographic corpus, which contains eDTLR (the digitalized version of the *Romanian Language Thesaurus Dictionary*) and other essential Romanian dictionaries (old and new dictionaries, general or specialized ones), aligned at entry level.

Keywords: Romanian lexicography, CLRE, eDTLR, computerized lexicography, Linguistic resources, computerized lexicographic instruments

1. Introduction

The CLRE project is financed by CNCS – UEFISCIDI, PN II - Human Resources area, with the purpose to encourage the training of young teams of researchers, for a period of three years (August 2010-July 2013), with a team formed of three lexicographers and an IT specialist. The *CLRE* project aims at achieving a corpus which will include 100 dictionaries from the *Romanian Language Thesaurus Dictionary* bibliography, aligned at entry and, partially, at meaning level. The purposes of the Essential Romanian Lexicographic Corpus are: to achieve a scanned corpus, with the reference dictionaries of DLR, aligned at entry and, partially, at sense level, to obtain a medium of programs that allow an interactive consultation, to develop a quasi-exhaustive list of words for Romanian language starting with the aligned corpus.

2. Principles of Development

Through this project it can be clearly seen the necessity of creating a missing bridge between two very different directions of scientific research. The ways of approach, development and solutions to the scientific problems differ significantly between the Computer Science area and Lexicography. However, the interdisciplinary cooperation can lead to surprising results, not only for the involved parties, but also for the general research.

The informatics part within this project aims at achieving the dictionaries from the established bibliography, in electronic format to process them (by OCR – optical character recognition – the conversion from image to text), to store them in a database, segmenting the text at entry level and then to process these data by aligning them at word level and, where it is possible and if it exists the necessary information, to achieve the alignment at meaning level.

All the stages of processing comply with the above enounced order having as reference the dictionary in work. The interoperability of stages from various dictionaries comes to support the adaptation and the optimization of the working process depending on the encountered particularities. However, the most viable solution should cover a great area of problems because an individual treatment for each dictionary based on the specific features would lead to waste of time and to a significant effort.

A basic principle of this project is the opportunity to extract partial data with an important degree of coherence. In this respect, all modules process and store information that can be used regardless the completion stage of the processes.

The software tools developed in this project respect the principle of portability and free access through Internet service.

All results obtained by this approach can be used via query of the database, which eliminates the physical consultation of any of the 100 dictionaries. There is also the possibility of achieving a complex search that increases the degree of interactivity in terms of utility and by providing all these components via Internet various access obstacles to the information source are removed.

3. Storing and Securing Data

All processed data have a different legal character. Some of these dictionaries from the CLRE bibliography are protected by the national patrimony law[1] or by copyright law[2]. As such, few dictionaries have from this „point of view" a free character. For this reason we must secure and limit the access to the data like: folders containing scanned corpus, text obtained after optical character recognition, segmented entries resulted from the processed dictionaries, aligned definitions. This data is stored on a data platform based on SQL.

1 Law no. 182/2000 regarding the protection of movable national cultural heritage, republished in 2008.

2 Law no. 8/1996 regarding copyright and related rights modified by Law no. 285 of June 23, 2004 and Urgency Ordinance 123 of 1st September 2005.

The web access can be done through the project address http://85.122.23.90 and the primary stored data are available at ftp://85.122.23.90. As a security measure the password is encrypted in order to protect the personal information. Also access to computer tools that can affect the database is divided into levels of rights. However, to support the user and simplify the access to the software platform, the access to services is done with the same username and password, depending on rank of the user account.

4. The Process of Dictionary Digitization

Digitization is the process by which, using a scanner, the document (the book) is transposed from physical (paper, manuscript, book, volume) into electronic format (pdf., tif., jpg. files). This method has the advantage of facilitating the access to information, which can be consulted online, and, at the same time, the specific document can be accessed simultaneously from several locations. Moreover, it represents the means of distributing rare documents, whose physical consulting may cause its deterioration.

4.1.The Acquisition of Dictionaries (Scanning)

This first stage is achieved with the help of a professional planetary scanner which uses the technique of photographing pages in a controlled environment. For an optimal quality of the captured images, the following settings must be taken into account:

- white cold light from auxiliary lamps;
- we do not use the light produced by camera flashes because it makes characters brighter, thus lowering the contrast between the letters and the background, and in other cases a blurring effect is caused, which seriously reduces the quality of the captured image;
- photo cameras are on manual mode;
- the environment is not exposed to any other source of light;
- the exposure times are between 1/15 and 1/30, depending on paper quality;
- auto focus;
- the ISO level is set to 200.

One of the major problems in the scanning process is represented by the quality of the paper. For example, thin sheets of a book affect this stage because of transparency, which allows the capture of letters from the reverse and the following pages. This effect can be minimized by interposing a black matt board under the page.

More details about that can be found in the tutorial offered by E-BOOK ENLIGHTENMENT[3].

4.2.Capture Editing

The captured images are edited with the application BookDrive Editor Pro[4]. In this stage, the images are cut, in order to separate the content of the dictionary page from the

3 http://en.flossmanuals.net/e-book-enlightenment/scanning-book-pages/
4 http://www.atiz.com/bookdrive-editor-pro/

rest of the capture. Another task is represented by the process of pivoting images in portrait orientation. Thus, line of the text should display a right angle with the bottom of the page.

In the next step the new images in tif. format are saved with dynamic compression, in black-and-white. In this stage, some pages require superior editing, which implies setting the contrast, the level of primary colors, or reducing noises.

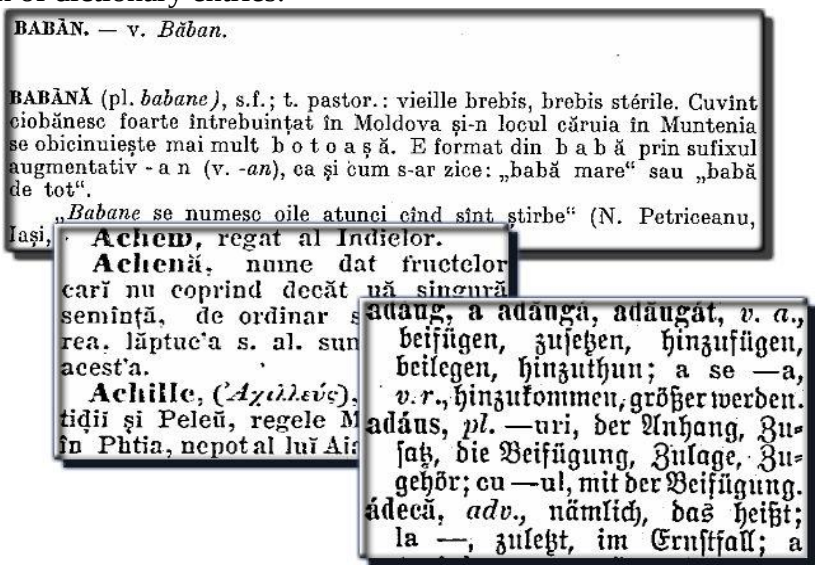
4.3.The Process of Character Optical Recognition (OCR)

The images acquired in the previous stage were edited and indexed within a database. The editing meant a process of character recognition (OCR), using The Abbyy Fine Reader 9.0 library set for each page.

The information was thus stored in the data base in two ways: as image of the dictionary page and as text format. For reasons related to paper quality, the conditions in which dictionaries were stored, the passing of time, or the typing quality, the efficiency of the OCR may be affected. Thus, in the recognized text some errors may appear due to the OCR process.

5. The Primary Editing of Data (Breaking in Entries)

The unit of indexing information within dictionaries is the entry, the lexicographic definition of a word. The digitization process is followed by the stage of identification/ segmentation of dictionary entries.


BABÀN. — v. *Băban.*

BABÂNĂ (pl. *babane*), s.f.; t. pastor.: vieille brebis, brebis stérile. Cuvînt ciobănesc foarte întrebuințat în Moldova și-n locul căruia în Muntenia se obicinuiește mai mult b o t o a ș ă. E format din b a b ă prin sufixul augmentativ - a n (v. *-an*), ca și cum s-ar zice: „babă mare“ sau „babă de tot“.

„*Babane* se numesc oile atunci cînd sînt știrbe“ (N. Petriceanu, Iași,

Achem, regat al Indielor.

Achenă, nume dat fructelor carĭ nu coprind decăt ună singură semințĭ, de ordinar s rea, lăptuc'a s. al. sun acest'a.

Achille, (*Ἀχιλλεύς*), tiḑii și Peleŭ, regele M în Phtia, nepot al luĭ Ai

adaug, a adăugá, adăugát, *v. a.*, beifügen, zufetzen, hinzufügen, beilegen, hinzuthun; a se —a, *v. r.*, hinzukommen, größer werden.

adáus, *pl.* —uri, der Anhang, Zufatz, die Beifügung, Zulage, Zugehör; cu —ul, mit der Beifügung.

ádecă, *adv.*, nämlich, das heißt; la —, zuletzt, im Ernſtfall; a

Figure 1: Types of dictionary

The analysis of formatting styles and the position within the page is the most viable solution, since the accuracy of the recognized text after the OCR-isation process, when compared to the original, leads to the conclusion of not using the information contained within the text, as there is no certainty in the validity of the processed data.

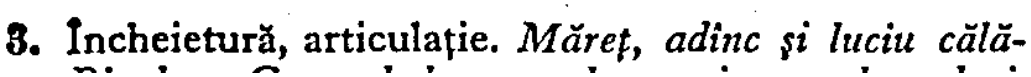
Though, this solution is impeded by the diversity of stylistic ways of formatting.

Practically, there are dictionaries whose titles of entries:

- are aligned before, after, or on the same level with the definition;

- are preceded or not by alphanumeric series or punctuation marks;
- have the same formatting style as the body of the definition;
- are written in lower case, upper case letters, or combination of these.

To answer these problems, an algorithm was created, which analyses each page of the dictionary and, based on both word formatting and its position within the page, the algorithm identifies the title terms that appear in the text. This is a goal approach, which aims at finding the common element of each set of features that characterize dictionaries.




```
...
<text>
    <p l="197" r="1073" t="280" b="351" len="117" li="9" ri="0" ls="35" a="3">
        <lines>
            <l l="249" r="1073" t="280" b="316" ci="0" C="57">
                <words>
                    <w d="false" ci="0" len="2" id="275489">3.</w>
            <w d="true" ci="3" len="12" id="275490">încheietură,</w>
            <w d="true" ci="16" len="12" id="275491">articulație.</w>
            ...
                </words>
        </l>
        <l l="197" r="1072" t="318" b="351" ci="57" C="60">
                    ...
        </l>
        </lines>
        <chars>
        <c l="249" r="266" t="284" b="310" c="45" s="false" sid="3">3</c>
        <c l="268" r="276" t="302" b="310" c="55" s="false" sid="3">.</c>
        <c l="276" r="298" t="280" b="310" c="100" s="false" sid="3" />
        <c l="298" r="310" t="280" b="310" c="19" s="false" sid="3">î</c>
        <c l="312" r="329" t="293" b="309" c="74" s="false" sid="3">n</c>
        <c l="332" r="346" t="292" b="308" c="59" s="false" sid="3">c</c>
        <c l="347" r="364" t="284" b="308" c="63" s="false" sid="3">h</c>
                                                    ...
```

Figura 2: Fragment of a scanned page with a OCR-ized text

The extracted text from the scanned page is stored in the XML format. The indices regarding the position in the page of each character, its text line, whether it is a capital letter or not, the font, the font size, whether it is bold, italic, or underlined must be maintained. All these parameters, used in various combinations, define the multitude of formatting styles.

The multitude of possible title-words is made up of the first tokens on each line of the page. The limit was established by starting from the presumption that some dictionaries contain signs, numbers, or alphanumeric series indicating the existence of the dictionary entry. We remove from this group the token whose formatting styles are equivalent to the most common style used within a specific page. This step is justified in the case of the OCR-ed pages, where at least two formatting styles have been identified.

The action of removing the words with common formatting is justified by the numeric existence of more words than titles of dictionary entries, these being different from most of the lexemes in the dictionary as they have their own formatting style (in most of the cases).

In this new lot, the tokens are ordered based on the frequency of formatting styles. By analyzing all the parameters previously mentioned, the algorithm classifies the words at the beginning of the lines into four categories of words, namely:

Title words;

Possible title words;

Terms that can be used as title words but with very little probability;

Terms that cannot be used as title words.

The page where the analysis of the formatting styles was made is sent to a linguist to be validated. The specialist has the ability to confirm or reject the options presented by the algorithm.

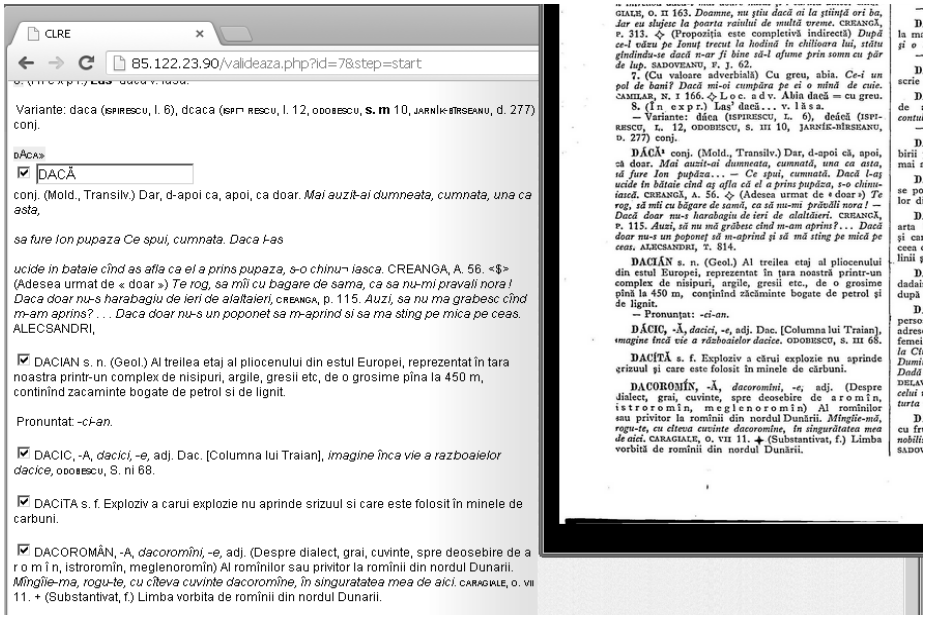


Figura 3: Validation tool for segmentation of dictionaries

The work method is very simple, the linguist has to compare the image of the page with the extracted text in order to mark or unmark the title words where it is the case. The title-words are suggested by the segmentation algorithm. This tool offers four types of options:

- definitely title-word – with the box marked;
- possible title-words – yellow marked tokens;
- ordinary tokens with a little probability to be title-words;
- impossible title-words – tokens which can't be selected.

Furthermore, there is the possibility to correct the recognized form of the defined term. A special activity performed by lexicographers was to manually introduce into the interface for validation the entries from the dictionaries written in the Cyrillic or in the translational alphabet. This operation is necessary on account of the errors that occur in the process of automatic recognition, as, yet, there is not any OCR software available for old Romanian texts. Since the Romanian Cyrillic writing varies depending on its date and source, it is virtually impossible, at least for now, to do an automatic recognition of an old Romanian text from an image format to a text format; this is why a manual validation has been preferred, made by lexicographers, who have practically

attached a "label" with the transcription of the title words in Latin alphabet. Moreover, the old dictionaries will be aligned on the level of image, not on that of the text, only the area of the searched word being displayed in the image format.

Following this stage, the alignment of the title words to their definition is processed. This new parameter is introduced in the analysis of a new page. At the same time, information is kept on the same format that is used for the title words on the validated page. This information is used in the analysis of a new page only if it refers to one of the format existing styles. This solution was applied as there were cases in which the types of the format styles did not appear in subsequent pages after the process of automatic character recognition.

6. Secondary Editing of Data (alignment on the entry level)

One of the essential aims of the project is represented by the alignment on the entry level of the dictionaries in the bibliography. The method does an automatic grouping of the common lemmas. After this stage, some problems may arise regarding the title words rendered in different orthographic systems of the Romanian language. It is known that in 1993, the Romanian Academy urged that the words containing*"î"* in the middle should be written with *"â"*. Another problem of this alignment is caused by the existence of various forms and variants in dictionaries. For all the exception previously mentioned, the linguist has the ability to add the correct lemma.

7. Conclusions

The CLRE project aims at achieving a useful instrument for both lexicographers and computer scientists and for other users interested in the study of the Romanian language. Given the alignment of the 100 dictionaries with eDTLR, the completion of the CLRE corpus will mean the re-evaluation of the results achieved on the project of digitization of the Thesaurus Dictionary of the Romanian Language. Last, but not least, this corpus of dictionaries may be aligned with lexicographic corpora of other languages, which will increase the visibility and accessibility of the Romanian linguistic resources in electronic format.

References

Atkins, B.T.S., Rundell, M. (2008). The Oxford Guide to Practical Lexicography. Oxford, Oxford University Press.

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007). The Digital Form of the Thesaurus Dictionary of the Romanian Language, *Advances in Spoken Language Technology* (Corneliu Burileanu, Horia-Nicolai Teodorescu eds.), Bucureşti, Editura Academiei Române, 195-206.

Dănilă, E., Clim, M.-R., Catană-Spenchiu, A. (2011). Towards a Romanian Lexicographic Corpus. „*Philologica Jassyensia*", An VII, Nr. 2 :14, 191-198.

Haja, G., Dănilă, E., Forăscu, C., Aldea B.-M. (2005). Dicţionarul limbii române (DLR) în format electronic. Studii privind achiziţionarea, Iaşi, Editura Alfa, www.consilr.info.uaic.ro.

Tamba Dănilă, E., Clim, M.-R., Patraşcu, M., Catană-Spenchiu, A. (2012). The Evolution of the Romanian Digitalized Lexicography. The Essential Romanian Lexicographic Corpus, *Proceedings of the 15 th EURALEX International Congress*, Oslo (Ruth Vatvedt Fjeld & Julie Matilde Torjusen eds.) , Press Reprosentrales, UiO, p. 225, *in extenso* (pe suport electronic) 1014-1017; http://www.euralex.org/proceedings-toc/euralex_2012/.

SUGGESTIONS FOR THE CLASSIFICATION OF TEXTS

CĂTĂLINA MĂRĂNDUC

'Iorgu Iordan – Al. Rosetti Institute of Linguistics' of the Romanian Academy, Bucharest – Romania

catalina_maranduc@yahoo.com

Abstract

The Department of lexicology and lexicography of our institute possesses a collection of texts in electronic format. In the first part of the paper, we use a qualitative and quantitative perspective to present the content of this collection, which includes approximately 2,000 documents, the manner it is structured, the way it was obtained and it is currently used. We also point out certain deficiencies and instances of incoherence in its management.
In the second part we present suggestions regarding the most efficient way of classifying the texts. We start from a pragmatic perspective on the text, seen as a macro-message addressed by the sender to the receivers within a situation of communication and as a bearer of an intentionality that has both a locution and per-locution character. The texts in our collection are selected according to the criterion called success and thus represent patterns of efficient fulfillment of the addresser's intentionality, which we can infer from the texts' enveloping form and manner of structure. The texts should be classified according to several criteria, present in the communicative enveloping form (the header or the bibliographic record of the text), annotated, in such a way that a computer may select them depending on one or more than one of these criteria.

Keywords: collection of documents in electronic format, operative criteria for text classification, principles of text classification from a pragmatic viewpoint

1. The collection of electronic documents at Bucharest Institute of Linguistics

The Department of lexicography of ILB possesses a collection, which at 01.02.2013 included 1,980 texts in electronic form.
The first documents of this collection were obtained by typing the texts. This remains the only method of rendering in electronic format the texts written in Cyrillic alphabet for which there are no new editions elaborated by specialists in the history of literary language. The collection of typed texts contains approximately 200 documents and only 40 of these are typed from texts written in the Cyrillic alphabet.

During the years 2007-2008 we obtained a grant, CNCSIS 1215/2007, which aimed to collate these typed texts and bring them to an optimum form, superior to the one presented by the OCR-rendered form. Not all the 200 texts could be collated.

The activity of scanning and OCR rendering of texts began in 2008 under the guidance of computer specialists who participated together with us at the eDTLR project.
Until September 2011 we worked with only one scanner and only one person scanning the texts, namely myself. We scanned for the eDTLR project works which could only be found in the ILB library. In the report for the managing of the project it was specified that a proportion of 4 percent of the vast bibliography of DA and DLR was accomplished in Bucharest.

We achieved another 10 percent by downloading the texts scanned in Iaşi and those made available by the researchers from the other centers from Bucharest and Cluj, centers which took part in the afore-mentioned project. In the end we presented to our colleagues a slip pattern with 1,000 quotations which had been extracted simultaneously from 10 books in a month using the program called Lucon, a concordance, whose author was Cătălin Mititelu[1].

This example led to the expanding of the activity of scanning and OCR rendering in the department of lexicology of ILB; subsequently, five scanners were used, and thus, in the last two years, the number of scanned texts was trebled. The documents are brought in the txt format with Unicode UTF 8, in order to keep the diacritics, by using the ABBYY FINE READER 9 program.

With the help of the concordancing program, we can open a great number of such text documents simultaneously, all of them from our collection or only a part of them. We can find the number of occurrences of each word or those of a word form, and these were used successfully in order to find the first attestation of words. This collection of documents is continuously increasing, but its management is not always rational. The collection is not the only one in ILB.

The Department of Literary Language, headed by dr. Al. Mareş, possesses another collection of about 200 old texts, in a doc. format, which were obtained by OCR rendering and then collated by data operators specialized in typing. However hard we tried, we did not manage to convince them to keep the scanned materials from which they obtain the OCR rendered texts. The Department of literary language does not possess the scanned texts, although scanning the texts also represents a way to preserve the books. Being able to access both collections of texts of the two departments, one department was able to access some of these texts from the other department, despite the differences in the format and conventions of marking the page number.

Other departments within ILB also possess texts that have been scanned or possess other electronic formats about which we have but scant information. We offered help to whichever researcher asked for help in copying of documents from our collection or scanning some of the most widely used books from the common library.

The solution for the optimal management of texts would be for all the resources of the departments to be located on an ILB server, one that all the researchers had access to, in order not to double the activity of scanning and OCR rendering.

[1] The program is available on the 6.0 variant at the internet address: http:/sourceforge.net/projects/lucon.

Returning to the collection of the documents of the department of lexicography, this was built on the basis of the DA and DLR bibliographies, which are extremely vast. Texts of the editions comprised in these bibliographies written in the Cyrillic alphabet were typed, and afterwards they were collated.
But in order to resume and modernize the first letters of the thesaurus comprehensive dictionary of Romanian, as the DA + DLR is sometimes referred to, a radical modification of the bibliography was initiated, in such a way that the texts for which an impressive amount of work was done and was financed also by the CNCSIS 1215/2007 were left behind.

The principle which functioned in the case of the old bibliography was that of using the first editions of books. Currently this has been replaced by the principle of using the most recent critical editions. Thus the texts written in the Cyrillic alphabet are no longer quoted in our transcription, but in the editors' transcription, since they are specialized in this field. However, the editors generally use the interpretive transcription, distancing themselves from the original form, instead of using the diplomatic transcription which has so far aimed to keep as close to the source text: thus the reader is denied anything but an extremely mediated access to it.

The quotations excerpted previously that will be used in the new volumes will have to be transposed in the new editions with the help of the computer, and to do this the electronic format of both the edition from the old bibliography and the edition of the new bibliography will be used.

The work of elaborating the new bibliography and of rigorously establishing the date of the texts lasted a few years, and during this time it happened many a time that an edition was scanned just to have another specialist recommend a different edition, considered as superior, then scan this edition too. These are the very reasons for which it is not a rare occurrence that the texts are present (with or without justification) in our collection in several editions.

To be able to preserve the old bibliography, some documents were split in smaller documents. For example, a particular novel from the old bibliography was now to be found in the *complete works edition* of that particular author, we decided to split that edition smaller parts that render the original bibliography. Sometimes only the OCR rendered format was split, but other times the pdf scanned format was also split, a fact that is regrettable as we ended up with more documents.

Our collection does not contain annotated corpus of Romanian. In (Barbu, 2008) suggested that we annotate manually at the header of the text, marking differently the title, the author, the publishing house and other information. This suggestion has not been put into practice.

In fact, these data represent some important textual marks for the classification of the text, they are a textual enveloping form we cannot do without, albeit it should be mentioned that they do not belong to the text proper.

The documents are placed in folders that contain the same text in several formats, namely at least two: pdf from tif images and txt with Unicode UTF 8, obtained by OCR rendering or by typing. To this is sometimes added the doc format and, starting from 2012, also the OCR rendering saved in a pdf format, which allows looking up the word or the quotation on the page numbered similarly with that of the pdf from the scanned texts.

The presence of the scanning and of the txt format in the folder is a rule that is not without exceptions (though these are quite infrequent). There are folders in which only the scanned text exists. The texts written with Cyrillic spelling (or rather with a transition spelling, combining Latin letters with Cyrillic letters, which we include in the same category) cannot be OCR rendered. In some cases we even have typed texts for these documents. There also exist folders from which the scanning is absent, for which the scanned form was lost, and thus it is impossible to check the correctness of the texts.

The looking up of documents in this collection is done manually. To simplify the work with this collection of documents we included in the same folder the books published in several volumes and then we placed a number of folders of this kind in a surrounding one in case of books belonging to the same author. At the present moment, the folders obtained in this manner are deposited in five sections.

2. The structure of the collection of electronic documents

2.1. Books

The first and the richest collection comprises a number of 1,316 volumes that belong to the old or the new bibliography for DLR and DA and confer these prestigious dictionaries a certain continuity; they may be looked up or indexed with the concordancing program that we possess.

2.2. Dictionaries

The second section comprises documents that should be aligned with another program, one that should index only the words that the paragraphs begin with, namely the title words of the lexicographical definitions. This section comprises 133 documents, a number which includes in fact only 75 dictionaries, some of which have several volumes. Some dictionaries are very old, are written with Cyrillic or Gothic (old German) spelling and cannot be OCR rendered. The dictionaries are monolingual and explanatory, bilingual, multilingual, etymological or belong to special domains. Certain dictionaries are for other languages, Turkish, English, Italian, Romani, as these are useful for establishing the etymologies. Until the present moment, we have not possessed a program which allows us to align the dictionaries.

2.3. Journals

The third section contains periodical publications. We possess 501 issues belonging to 11 publications; some of these have a single issue, while the magazine entitled Viaţa Românească is present with approximately 300 issues. If we were to consider the specific type of the journalistic genre, the question arising here is why this category does not also include the volumes that contain journalism writings by a certain author.

2.4. Manuscripts

The fourth section includes the scans of 2 books in manuscript form, of which one has 3 volumes, amounting to a total of 4 documents. They are used for consultation reasons, in order to check the fidelity of the published editions. Note that libraries do not possess (or do not allow access to) scans of manuscripts.

2.5. Literature studies

In the last section, which has recently been founded and comprises 26 documents, we included texts that do not belong to the bibliography of the dictionary we work on, but are useful for other type of research apart from the lexicographical activity. They are written in various widely-used languages.

3. Operative criteria for text classification

As we have quite an impressive collection of texts, one starts to ponder about their classification according to various criteria, in such a way as to make their consultation operational.

As one may notice in the afore-mentioned presentation, the section that raises difficult problems of ordering and classification according to rigorous principles is the first section, which includes a number of 1,316 documents. It comprises:

– Fiction, novels, volumes of novellas, sketches, stories, poems, plays. More often than not these are written by a single author, but the collection includes also books that belong to a specific trend (for instance the School of Ardeal) or were written by members of a family (the Văcăreşti Poets);
– Critical work and history of literature;
– Memorial literature, journals, memoirs, travelling journals;
– Correspondence;
– Journalism work by certain authors, collections of articles published in the press;
– Speeches made by parliamentarians;
– Legislative texts, codes of law;
– Religious literature;
– Folklore, dialectal texts;

– Manuals or scientific treatises from various fields: philosophy, logics, aesthetics, linguistics, history, mathematics, physics, anatomy, medicine, pharmacy, botany, zoology, economy, agro-technical engineering, etc.

Not all the fields of human knowledge are illustrated in treatises; some of them are not illustrated or are illustrated only by special dictionaries included in the second section of our collection. The reticence for scientific texts stems from the obsolete conception according to which scientific and technical terms should not be included in the word-lists of general comprehensive dictionaries published under the aegis of the Academy.

As a matter of fact, nowadays when high-school education has become the prevalent tendency, each person needs to learn the notions and the basic terminologies of all the domains of human knowledge and thus certainly needs terminological clarifications.

A first criterion according to which documents should be classified would be their year of publication. As a consequence, the 1,316 texts of this section should first be ordered depending on the century they were used originally, then according to their scientific, monograph or fiction character.

The ordering of texts does not only have the function of allowing users to look these texts up. Using the concordancing program on a very great number of texts is a slow and tedious process. It is recommended only for looking up certain rare words. Very good results may be obtained if we use the concordancing program to open a folder that contains a smaller number of books, selected according to certain criteria, chosen depending on the objective of the research: either all the texts are written by a certain author, or all are documents published within a certain period or the documents having a specific theme. This way a small number of relevant quotations would be selected for a word that seems novel, though this manner was adopted and used pretty frequently at a specific time. At the same time, the labels referring to the register and frequency (rare, unusual, familiar, obsolete, colloquial, regional) would be introduced in the dictionary depending on precise criteria (number of attestations).

The question arises whether we shall be able to use criteria that are sufficiently clear and precise to classify these texts, focusing on criteria that are simple to be used by both knowledgeable and less knowledgeable people, also by the computer.

It would be ideal to have some exact criteria when creating a corpus of texts and also some criteria for its organization that would help advanced searches. The header (or the bibliographic record) of any text should be annotated according to all these criteria, so that the software be able to read and can give the user whatever he/she needs. We have not made this collection according to our classification principles, but constrained by dictionary references.

Other corpora are organized differently, according to their needs. For example, the Oxford Corpus[2] is based on texts found on the World Wide Web, and its structure is different than ours (for more details see the link below). While most of our corpus consists in fictional texts, masterpieces of Romanian literature, fiction accounts for only 0.3% in the Oxford Corpus. Maybe our corpus needs a more modern conception, because the formation of a new corpus requires human and financial resources;

[2] http://oxforddictionaries.com/words/the-oec-composition-and-structure.

therefore, it is desirable to use what we have not only for the Academic dictionary, but also for other research.

4. General principles of text classification from a pragmatic viewpoint

The distinction between the text and a simpler series of signs, sentence, complex sentence, or paragraph cannot be accomplished by means of grammar, which is a science that studies human communication and the relationships between signs. Likewise, this distinction cannot be made by means of semantics, a science that studies human communication and the correlation between signs and the signified or the possible world these designate. The text may be circumscribed only by means specific to pragmatics, a science that studies the relationships that the communicative message establishes between its users in a certain communication situation. Text classification is thus a problem that also belongs to the field of pragmatics.

Each text is a (macro-)message that the author addresses one or several receivers with certain intentionality. The condition of success of the text is the fulfillment of the addresser's intentionality.

The property of textual character of a number of signs is circumscribed by Beaugrande and Dressler (1981) to seven standards: (grammatical) cohesion, (semantic) coherence, intentionality, acceptability, information-providing character, situationality and inter-textuality.

We note that the theory can answer the necessity of refining the model outlined by Plett (1983) by clarifying that which the author used to call internal premises. We will notice that some standards refer to the activity of textualizing of the addresser (intentionality, information-providing character), others to the activity of decoding the receiver (acceptability and expectancy).

According to Rastier (1989) we might establish the classification of texts according to the point of view of each of the six poles of communication in the Jacobson (1963) model: sender, receiver, referent, channel, code, text-message.
If we place the receiver in the centre of classification, the main criterion will be expectancy, namely the satisfying of his expectations regarding the quantity of information (possibly correlated with that of dimension), the expressivity and the coherence of the text.If we start from the sender, the criterion of classification will be the type of emotion, the sentiment, the feeling, the information that he intends to communicate through the text.

If we start from the referent, we will notice that for a text it represents not an object, like in the case of simpler linguistic signs, but a configuration of objects, namely a state of things; thus we will classify the text depending on the value of truth of the state of things it refers to. To apply this criterion of text classification, we will observe if the system of reference according to which we establish the value of truth of each type of text is or is not the real world. If we only consider the real world, then it is necessary to establish the value of truth not only in the multitude consisting of two elements, true and false, but we also need to resort to the system of polyvalent logics (Vasiliu, 1990), where the function assumes values in a multitude of several elements, or the system of modal logics, which allows four modals (Vasiliu, 1978) such as: possible, necessary, permissible, mandatory.

If we start from the code, we will classify the texts depending on the stylistic variant of the language they are issued in. If we start from the channel, we will classify the support by which the text reaches the receiver.

Finally, a classification that would have at its centre the text-message should be made according to aesthetic and cultural criteria, apparently more appropriate for literary texts, but also applicable to other types of texts. We consider that all these manners of text classification are relevant depending on the problem studied by the researcher (Mărănduc, 2005).

The pragmatic vision of research calls for a classification from a functional perspective. The text cannot fulfill its function, consisting in the addresser's intentionality to produce an effect on the receiver, unless the expected receiver recognizes this intentionality.

The textual sense cannot be decoded by neglecting the communication situation. The typical situations determine types of text. That is why we shall notice that the types of text are circumscribed less by defining features that exhibit a linguistic character and more by features that have a cultural, social or even strictly practical character.
In the communication situation there exist correlations between the addresser, the receiver, the referent that are not mediated by the text. That is why in order to accomplish a typology of texts we ought to found a typology of situations in which these texts have been produced.

In countless papers of linguistics, text classification is done by using the term of genre, a term borrowed from classic rhetoric which is nevertheless ascribed a much wider meaning and thus becomes applicable not solely to literary texts. In (Rastier, 1989), the genre is not a program of prescriptions and licenses that regulate the generating and interpretation of the text: these have no linguistic character, but are the resulting entities of a social codification.

The same idea is to be found in Van Dijk (1985), where it is stated that there are differences between the various types of discourse, also called genres. Although the general principles are valid for each type of discourse, there may exist constraints added over the local and global coherence or specific semantic properties, valid for certain typical discourses.

As we deal with a theoretical foundation of text typology, Beaugrande and Dressler (1984) estimate too that the traditional criteria of linguistics cannot be used to found a typology of texts. One needs to appeal to data of the socio-cultural encyclopedia to determine a dominant function of the text: this simultaneously fulfills several such functions, but should be classified depending to its own dominant.
In order to systematize the classification, we shall first notice that there exist texts with a pronounced functional syncretism (poly-functional texts) which are less connected to a certain opportunity or necessity of being produced, and we shall place these texts in an A group, and specialized types of texts, emphatically marked by a unique practical functionality (mono-functional texts), which can be placed in a B group.

The poly-functional texts may be predominantly cognitive, such as the scientific or the didactic texts, differentiated according to the type of receivers they envisage, or predominantly expressive, as is the case of literary texts. Journalistic texts seek both to confer dissemination of information and to produce a strong impression on the receiver.

The texts in group B are emphatically marked by a kind of functionality. The receiver may receive a piece of information with a practical character via the text, in cases such as announcements or inscriptions made on an object. The useful and practical character is even more emphatic for the booklets informing people how to use various house utilities, drug prescriptions, or in the case of advertisements and invitations.

The receiver may be offered a reasoning that attempts to make him change his/her opinion, in the case of political discourses and pleading during court trials, which can be placed in the category of texts of an argumentative type.

The receiver is let know an obligation or an order in the case of legislative texts, books of instructions and norms concerning work protection, citations or summons, the success of which is obtained provided the receiver acknowledges this situation that has a compulsory character and the consequences he/she will face if they ignore it; these texts at times contain the mention of the means of sanctioning the person concerned if they choose to ignore the directive.

For our collection it is not necessary to sort the texts according to the success of the addresser's intentionality, since a selection has been made of the most representative or outstanding texts, texts of success which are in fact true models for the accomplishment of textual standards.

When we evaluate the situation of communication it is nevertheless necessary to refer to the specific standards of the historical period the texts were written in and not to judge old texts according to the norms of contemporary texts, considering them as naïve, obsolete or uninteresting.

In terms of the informative function, this is more efficiently fulfilled by the texts belonging to the scientific style, as these are characterized by specific terminology, the mono-semantic character of terms and language economy, as the quantity of information compared to the typographical space is ample.

The category of texts that exhibit a great economy of language should also comprise dictionaries; the users expect these to contain the definitions of as many words as possible within a frame of reasonable dimensions. If we refer to the value of truth correlated with the real world, which is specific to these types of texts, and if we use a bivalent logic this can be said to be correlated with, if not conditioned by, the system of knowledge mankind has access to in the historical time the texts were written in.

From the same category of non-fictional texts we can cite diaries, memoirs, memories, but these texts respect only optionally the criterion of language economy and almost never the criterion of mono-semantic character, since they use the figurative sense of terms, just like fictional texts.

The texts of a fictional type fulfill aesthetic functions and convey the addressers' ideas, sentiments, physical or mental states and are not informative. These texts do not respect the criterion of language economy and are characterized by the terms' polysemy, figurative senses that display an unlimited symbolical character and inventiveness.

The texts from the argumentative category cannot be qualified, either according to the criterion of information dissemination power or to the criterion of expressiveness (they have these functions too, but they are not dominant) but rather according to the criterion of coherence and efficiency of argumentation. Parliamentarian discourses belong to this category, but also the articles which express the opinion certain journalists have on various topics.

Similar characteristics are displayed by philosophical or religious texts, as these texts refer to the elements of a possible world that are not accessible to human knowledge and cannot be verified.

The normative texts, the codes of laws belong to the category of mono-functional texts that can be said to have a per-locution value because their very publication leads to modifications of the real world.

5. Conclusions

In order to computer-process the collection of texts, it would be necessary to annotate the information of the header or the bibliographic record of the text, in such a way as to achieve a classification of the texts according to as many indicators as possible, author, theme, field, year of elaboration, literary, scientific, familiar, colloquial, regional functional style; then it would be necessary for texts to carry a label of being placed in the category of poly-functional or mono-functional texts, another label concerning the degree of language economy, the mono-semantic or polysemous character of terms, the fictional or non-fictional, argumentative, coercive character, the per-locution effect.
A search engine for the collection of texts should be able to sort out, at the request of the user, through the multitude of texts according to one or several criteria.

It could be possible in the near future to study the characteristic of each text type, to formalize these characteristics so that software would be able to recognize to what particular type of text a certain fragment belongs to.

Acknowledgements. The author is grateful to Professor Dan Cristea that proposed for the first time in the eDTLR project to scan the entire bibliography of dictionary. Even if this endeavour seemed impossible, we have already realized half of the impossible. We also want to thank to the Romanian National Research Council for cutting down the eDTLR funding – we thus realized how strong we are as a team.

References

Beaugrande, R. A. de, Dressler, W. U. (1981). Introduction to Textlinguistics. *Longman Linguistics Library*, London, New-York.

Beaugrande, R. A. de, Dressler, W. U. (1984). Text production. Toward a Science of composition. *Advances in Discourse Processes, ABLEX Publishing Corporation*, New-Jersey.

DA – Dictionary of Romanian. Tome I. Part I: A-B. Bucharest, Socec & Comp. Bookshops and C. Sfetea, 1913; Tome I. Part II: C. Bucharest, Printing House of the 'Universul' Newspaper, 1940; Tome I. Part III: Fascicle I: D – de, Bucharest, Bucharest, Printing House of the 'Universul' Newspaper, 1949; [Fascicle II: de – destina; under print, 1989]; Tome II. Part I: F-I, Bucharest, Bucharest, Printing House of the 'Universul' Newspaper, 1934; Tome II. Part II. Fascicle I: J – lacustru, Bucharest, Printing House of the 'Universul' Newspaper S. A., 1937; Tome II. Part II. Fascicle II: Ladă-lepăda. Bucharest, Printing House of the 'Universul' Newspaper S. A., 1940; Tome II. Part II. Fascicle III: lepăda-lojniţă. Bucharest, Printing House of the 'Universul' Newspaper S. A., 1948.

Dijk, T. van A. (ed) (1985) Handbook of Discourse Analysis vol. I-IV, Academic Press, London, Orlando, San Diego, New-York, Toronto, Montréal, Sydney, Tokio.

DLR – Dictionary of Romanian Language. New Series. Tome VI. Letter M: 1965-1968; Tome VII. Part I. Letter N: 1971; Tome VII. 2nd Part. Letter O: 1969; Tome VIII, Letter P: 1972-1984; Tome IX. Letter R: 1975; Tome X. Letter S: 1986-1994; Tome XI. Part I. Letter Ş: 1978; Tome XI. 2nd and 3rd Part. Letter T: 1982-1983; Tome XII. Part I. Letter Ţ: 1994; Tome XII. 2nd Part. Letter U: 2002; Tome XIII. 1st and 2nd and 3rd Part . Letter V and letters W, X, Y: 1997- Tome XIV. Letter Z: 2000; Tome, 3rd Part, Letter D, 2006-2008, Academy Publishing House, Bucharest.

Jacobson, R. (1963). Essais de linguistique générale, Éditions de Minuit, Paris.

Mărănduc, C. (2005). Norme de corectă formare a textului – din perspectivă pragmatică [=Norms for the Correct Formation of Texts – A Pragmatic approach], Publishing National Foundation for Science and Art, Universitas collection, Bucharest.

Plett, H. F. (1983). Ştiinţa textului şi analiza pe text. Semiotică, lingvistică, retorică [Science of Text and Text Analysis. Semiotics, Linguistics, Rhetoric], Univers Publishing House, Bucharest.

Rastier, F. (1989). Sens et textualité, Éditions Hachette Supérieur, Paris.

Vasiliu, E. (1978). Preliminarii logice la semantica frazei [=Logical Premises of the Semantics of Complex Sentences], Scientific Publishing, Bucharest.

Vasiliu, E.(1990). Introducere în teoria textului [=Introduction In Text Theory]. Scientific Publishing, Bucharest.

RELYING ON LANGUAGE

DAN S. STOICA

"Alexandru Ioan Cuza" University, Faculty of Letters, Iaşi – România
dstoica_ro@yahoo.com

Abstract

Discourses are social interactions which leave traces in our lives. Thus, we feel close to the CDA approches on the anaysis of such interactions, mainly to the detection of the social profile of the actants, and the tool that seem to be most appropriate to our goal are, in our opinion, the pragmalinguistic markers present in discourses. As a new kind of medium - the Internet - is gaining over the others (or it only seems to gain!), we will therefore pay attention to what happens there, too. Being a study trying to determine the level of chances one could have to establish patterns in such a dynamic activity like the discourse, we don't think of identities, but of generic profiles, types of individuals, as they come on the basis of their habits in linguistic communication.

Keywords: pragmalinguistics, automatic profiling, critical discourse analysis, Internet forums

1. Introduction

Studying discourses seems to be an endless occupation, mostly because of the different approaches to the concept of 'discourse', but also because of the ever changing media and contexts where discourses take place. Our attention has always been drawn to the idea that discourses are social interactions which leave traces in our lives. Thus, we feel close to the critical discourse analysis (CDA) approaches on the analysis of such interactions, mainly to the detection of the social profile of the actants, and the tools that seem to be the most appropriate to our goal are, in our opinion, the pragmalinguistic markers present in discourses. As a new kind of medium – the Internet – is gaining over the others (or it only seems to gain!), we will therefore pay attention to what happens there, too. Being a study trying to determine the level of chances one could have to establish patterns in such a dynamic activity like the discourse, we don't think of identities, but of generic profiles, types of individuals, as they come on the basis of their habits in linguistic communication. As it is not (yet!) an applied study, there will be no corpus exploited in it, and the method will largely resemble a contemplation of common sense observations already made in discourse analysis of all kinds and in the field of pragmalinguistics. We are just at the beginning of an attempt to verify whether such (discursive) social profiles as those described above can be extracted by means of automatic tools (NLP tools) from discursive interactions taking place online. Journey with us and be prepared to share our joy as to whether the conclusions will be satisfactory!

They say the utterer is always right, because s/he knows why s/he has made the choices that s/he has made from the infinite offer of the language to express her/his thoughts or intentions. Anyone of us humans is entitled to believe they are right, because they know better than anyone else what they meant to say. We could then consider that one can always rely on her/his own sayings, on the linguistic expression of their respective thoughts or intentions. People can unveil or dissimulate whatever they want in their verbal interventions. They have control over their own expressions. Or haven't they? Nevertheless, pragmalinguistics and the CDA show that there are tools to detect more than it is said in a given verbal expression (oral or written). Markers in discourse are giving out lots of information no-one would ever agree to unveil of themselves for different reasons. From all the elements that describe linguistic competence, pragmalinguistics deals with the concept of "choice". The choices we make are describing us as social actors in different contexts. Observing this could offer ideas for new approaches in the study of language use, with or without automatic tools to help us in our efforts.

2. The reality of language use and some theoretical foundations

Let us start from a reality of all languages, because it is a language universal: there are no such things like identical words (semantically speaking) in one and the same historical language, there is no perfect synonymy. This brings us directly to the issue under discussion: trying to express ourselves, we make choices. The vocal of Simply Red does not say "I hope you'll understand", but "I know you'll comprehend", because it is not about reasoning, coping with some already existing bit of information, but capturing and processing new information in order to make it part of the existing knowledge. Rational vs. cognitive processing. Apparently, any average native speaker of English would assume it's the same thing. And it is. Up to a point! The above distinction between the semantics of the two words can stand for an example for what I am trying to talk about in the present study. People would say: potato-potato, but some specialist could find not necessarily the semantic distinction as such, but what it could make us think about. Some of us would say "comprehend" is a mark of presumptuousness, while "understand" is the "normal" way of speaking. Some would speak of the need of belonging (to some elite who use fancy words to say even trivial things). There will also be people for whom the meaning of "comprehend" is unknown, so they will simply not understand a thing. Looking to the facts everybody could remark in the everyday life, we can add situations such as the one where we have an individual using a "fashionable" word without knowing its exact meaning. I knew someone who used to say (frequently!) "mass media" instead of "average people", and we all know at least one person who says "fortuitous" (casual, accidental) with the meaning of "forced", just trying to use a fancier word and by that impress her/his audience.

This is not all. Let us now think of faking a mistake, for fun or for educational purposes. And then let's think of those kinds of situations where the utterer knows what s/he is saying, but in the end, s/he realizes s/he made the wrong choice of words with respect to the interlocutor. Here we have touched a fundamental idea related to the discourse: it takes (at least) two to have one! There is more: each of the two interlocutors has to try their best to get a representation of the other as near as possible to the reality of the other.

Getting back to the core issue of this study, we have to accept the idea that even adequacy is a question of choice (which could be good or not so good). We shall look now to some possible conclusions an analyst could extract from discursive production of known but more than this of unknown people. We are talking of some kind of second degree inference, and we have to bear in our minds that in doing so we use (almost) everything, from the physical context of the production of the discourse, to the common history of the two and the contract of communication they have agreed upon for that particular instance of discursive activity. The scariest part in all this is what Ivan Preston captured in his phrase saying that "meanings are in people, not in messages"[1].Almost everything can be recomposed – a set, a town, the world itself, even the climate! – but there is no such thing like the recreation of a state of human nature. Not only are we different from one another, but no individual can ever repeat his same self. This spoils the context, and makes it more than difficult to have rigorous patterns of human behaviour. And yet! Psychologists talk about the fact that where there is human activity, there are patterns. Obviously, they settle for less, but this is the only chance to make some progress in interpreting human behaviour. This is the chance also for all IT approaches on the study of language uses. As approximate as they can be, the patterns reproduce that part of the language uses which remains somehow the same in different situations (which have not to be so different, but seemingly similar!). Looking back to the examples above, we can see that there can be more than one interpretation of the use of some word ("comprehend") instead of its (helas!, imperfect) synonym "understand". Going further, the question would be whether we can "capture" the personality of the utterer of a phrase like the one in the song by Simply Red, by the simple fact that s/he made a choice of words. Could that option be a sign of ignorance? Or a sign of a good knowledge of the English language? Or maybe a desire to make some impact on the listener? Or…? Yes, yes and yes! All of the above. And more.

3. Important authors to keep in mind

In one of his published books (Marcus, 2006), Solomon Marcus talks about imprecision in language use. I dare anyone to find a better term to name the phenomenon! One could find in that book discussions on many kinds of imprecision, such as: the paradox, the antinomy, the independence, the undecidability, the incompleteness, the axioms of choice, the fractals, the plausibility (with its special case, the possibility) and indeed many others. The presence of an imprecision is difficult enough to manage in trying to get the meaning of some discourse, but Solomon Marcus invites the reader to think of even more complicated cases of interference of different imprecision: as significant examples, we have the way random interferes with negligible, the way chaotic interferes with negligible, or the random with the gradual. It is quite alarming. We could ask ourselves how do we still manage to communicate using language, when we cannot really rely on what we are saying or on what we think we understand from what others tell us. It seems that using language is one of the riskiest activities, but we continue to pretend to understand each other in this way. There is more: we like to communicate with the help of this marvelous tool, and we love to play with it, to ex-press ourselves, or to dissimulate ourselves using it. Using language is pretending. It's sending around

[1]Ivan Preston, „Understanding Communication Research Findings", in *J. of Consumer Affairs*, vol. 43, no. 1, 2009, pp. 170-173.

hints to direct our fellows humans in their effort of understanding what we are expecting them to do. It works. Not always, not perfectly, not with everybody, not in every circumstances. The imprecision make the beauty of the language, but they also make us look like victims of our own "habit" to speak.

The aim of the present study is to look a little longer at the later aspect: us as victims, in this particular sense of the language unveiling to the analysts things we wouldn't want to go public. The main target at the present time is the discourse on the Internet and the perspective is that of the analyst. The challenge is bigger, because the method should be new, the environment is new (and still not quite known or understood), the people interfering there are numerous (not actually, but as declared identities), the manner of using language is at too many levels of (im)perfection, and above all the freedom is unlimited. To keep this tremendous freedom and to really enjoy it, people tend to hide their identity. Looking closer at the phenomenon we could remark that this is not such a new behaviour. As we play roles all the time in society, we pretend. It's true that in the actual world, where we can bump into one another if we want to, getting to know who you are interacting with is easier and more accurate than in an environment which is based on make believe. To give a name to this problem, we could say that it is an issue of representation and metarepresentation, which is the psycho-social foundation of human communication of every kind. The better you can represent your interlocutor in your head, the faster you will be able to tune your verbal intervention to match his/her expectations, and consequently your discourse will be more effective. We know it's difficult enough to do that in the actual world, and that we often fail. So, what are the chances to do it right on the Internet? Practically zero! Should we just give up? No! We will just change our focus: we won't go after individuals, but after types of individuals, and this is possible because individuals – in actual life or in cyberspace – copy each other, they follow fashions, they imitate. So, they have become to resemble one another, and one could create groups of profiles describing common behaviour of various groups of people. It is a way of creating typologies in a world that refuses to give us enough information, but which is too important for either of us to ignore. All the analyst should have to do is to pick up elements of discourse that could speak of the speaker. The fundamental tool will come from pragmalinguistics (Veltman, Steiner, 1988), and it will be, as shown above, the choice any speaker makes in constructing their discourses. It resembles to the logical syllogism: you choose that, you are like this (All those who choose X are of the class Y; the individual speaker N chose X; then N is of the class Y).

Classes (or types, or generic profiles) are to be predetermined, in a focused research, starting from the interest we focus on in the study. If the interest is just to "see" who is there, on the Net, classes could form along the way. Let's think, for example that it was never hard to discriminate between those with something like a real name (say, Malcolm Gladwell) and those calling themselves "angry bird", or "big boy", or something of the sort. We shall never really know that "Malcolm Gladwell" is the actual name of the one who is using it on the Internet, but isn't this a problem even in the concrete world we live in?! So therefore we will be more inclined to believe that an individual who used a "real" name is in fact the person having that particular name. Can we now form a class with "real names" and another with nicknames? Yes, we can! This was just an example, but it is enough to prove the point. So, starting from the declaration of their respective identities, the forums' users can be broken down in classes like: complete real name, partial real name and nickname (here, we could have

many subclasses as the nickname varies from highbrow to low brown). One could also imagine more classes and subclasses, but the point is that we can distinguish patterns in the behaviour of the writers on the Internet from this point of view. What we have to avoid will be the temptation to transform patterns into rules!

Then, we can have as a criterion the capacity to stay with the idea of the article, to keep close to the topic the journalist has proposed in his column. A study of isotopes (surface ones as well as deep ones) would show how close to the article topic is a comment or a comment of a comment. Based upon this, classes of profiles could be drawn: on topic, close to the topic, vaguely close to the topic (with many subtypes: those who do not understand what the article is about, those who understand, but who would want to change the topic by keeping the same context with the same interlocutors, those who understand a part of the topic, but who are more interested in the author), not even close to understanding, wishing just to be present online, but also some kinds of *ad persona*, meaning they do not argue on the topic but attack the journalist or the writer of a previous comment. Then, we have the educated ones vs. the not so educated ones, the aligned ones vs. the neutral ones, the rational ones *vs*. the emotive ones and many other couples of opposite profiles.

As we can see, a lot of information can be extracted from the actual expression a writer gives to his/her thoughts or intentions. We are talking of information on the writer, something giving him away. Jean-Pierre van Noppen is talking of *tenors*, which are to be taken as characters or patterns of something (van Noppen, 2009). Such a tenor can tell what kind of person the author is or is pretending to be, what kind of people the expected audience are, what the relationship between them is (or what relationship the authors is presenting it as). Tenors could be: relative status (equality or inequality between interlocutors), social distance (eg. familiarity, friendliness), personalization, standing (or how much the author comes across as an authority on the subject), stance (broken down further into attitude and modality, which can be epistemic or deontic).

Starting from those tenors, we can analyze and construe lexical and grammatical choices of a given author and conclude to some pattern s/he could belong to. Following the description of that pattern, a class of individuals can be considered as opposed to other classes with different descriptions. I will mention an example from the book by Jean-Pierre van Noppen, to illustrate this idea: Explaining the tenor called *stance*, the author presents several kinds of attitudes, more or less explicit, and therefore more or less easy to detect. Let's have a look at the paragraph:

" Attitudes […]:

- Asserted attitudes are attitudes which are mentioned quite openly, a typical reader is aware of them and is free to disagree with them: " The government's behaviour was disgraceful".

- Assumed attitudes are attitudes which are mentioned as if they were truths accepted by everyone, on which an argument can be built: "After nine years of the government betrayal, …" (Main argument follows). A typical reader will feel less free to disagree with assumed attitudes.

- Triggered attitudes aren't mentioned at all, but a typical reader will infer them. Example: "Even though Fred's father is very old, Fred only visits him once a year". This triggers a negative attitude towards Fred. […] the syntax employed ("even though … only") encourages a typical reader to infer from the facts a negative attitude towards Fred.

One way of triggering attitudes is the manipulation of agency and affectedness in a text: wording material processes in such a way that certain entities appear as actors (and therefore come across as responsible for what happened) while others appear as goals (and therefore come across as more or less victims of what happened). " (van Noppen, 2009).

4. More authors and practical discussions

When it comes to expressing extreme nationalism or xenophobia, an author can put it bluntly in the text or it can simply make some unnecessary precisions like the French journalist who reported a crime committed somewhere in Paris mentioning that the perpetrator was a young French citizen of Moroccan extraction. If he would have simply said that it was a young man living in Paris, the missing accent on the ethnicity of the criminal wouldn't have created the possibility to infer that immigrants are a problem for the native French people. It goes similarly for precisions like "a Roma Romanian was caught while she…". An analyst can easily understand that the crime itself is not the major preoccupation of the journalist, but the immigrants are. The same goes for the comments' authors or for the authors of comments to other comments. We could imagine worste: what if the journalist (or some commentator on the forum) says the same thing in another form (like: there is no point to ask ourselves who the perpetrator is, because it has become common knowledge that Maghrebians/Roma Romanians do these kind of things to us)? What are we going to feel is the obvious conclusion? Would we be able to remember the crime? Maybe, maybe not. But all our stereotypes and our prejudices will push us to infer something negative about those immigrants, who are the source of all the bad things what happen in our country.

We can see how Paul Watzlawick was right when he said that in any act of communication there is a content and a relation, and that the relation is classifying the content, thus it is a metacommunication. It is quite the same finding as the one of the British scholars working in the field of the systemic functional grammar, some time later, when they were talking of the metafunctional organization of the language (Halliday, 1977). For Halliday, a social situation is a "semiotic structure" which can be described using several elements:

1. A field of social action**:** something that is actually going on, and that has attached a meaning within the social system;

2. A tenor of role structure: the participant relationships; as in the theory of van Noppen, they include the speech roles, those that come into being through the exchange of verbal meanings;

3. A mode of symbolic organization**:** the particular status that is assigned to the text (we would say *discourse*!) within the situation.

Language will always convey more than the information or the intention of the speaker. The text will always remain as an evidence of the meaning of the discourse in the particular situation where it took place. Confirmations of this view can come from anywhere, and in any form. Let's look to what Jay Lemke says:

"I hope that in all our work on how meanings are made with text we will remember that the text is a product and a record of meaning making processes which are essentially dynamic. These processes are social semiotic practices, the signifying practices of a community. It is these practices that make texts and make sense of texts, dynamically, dramatically, moment to moment, word by word, enacting meaning by words, in moments whose meanings the words make and change[2]. "

Parts of the text structure will reveal to the eye of the analyst even the most secret thoughts of the author. It's frustrating to see how difficult this can be to express oneself in words, but it can be even more frustrating to see how much more the interlocutor can infer from what you say.

5. Conclusion

What did I mean by saying this (Lemke, 1991)? Exactly what I said! No inference required! Can I rely on you for this? Because I cannot rely on the language.

References

Halliday, M.A.K. (1977): Text as Semantic Choice in Social Contexts in Grammars and Descriptions (Studies in Text Theory and Text Analysis). Eds. Teun A. van Dijk and Janos S. Petofi. New York: Walter de Gruyter.

Lemke, J.L. (1991): Text production and Dynamic Text Semantics, in In E. Ventola (Ed.), *Functional and Systemic Linguistics: Approaches and Uses*, Berlin: Mouton/deGruyter (Trends in Linguistics: Studies and Monographs 55).

[2] Jay L. Lemke, "Text production and Dynamic Text Semantics", in In E. Ventola (Ed.), *Functional and Systemic Linguistics: Approaches and Uses*, Berlin: Mouton/deGruyter (Trends in Linguistics: Studies and Monographs 55). 1991, pp. 23-38.

Marcus, S (2006): Paradigme universale II. Pornind de la un zâmbet, Bucureşti, Paralela 45.

Noppen, J.-P. van (2009): English Synchrony. Developing Pragmastylistic Competence, Part II, Bruxelles, Presse Universitaire de Bruxelles.

Veltman, R., Steiner, R. (eds.) (1988): Pragmatics, discourse, and text,London: Frances Pinter.

CHAPTER 2
TEXT PROCESSING

ROMANIAN-ENGLISH STATISTICAL TRANSLATION AT RACAI

TIBERIU BOROȘ[1], STEFAN DANIEL DUMITRESCU[1], RADU ION[1], DAN ȘTEFĂNESCU[2], DAN TUFIȘ[1]

[1] *Research Institute for Artificial Intelligence"Mihai Drăgănescu", Romanian Academy, Bucharest - Romania*

[2]*University of Memphis, USA*

[1] *{tibi, sdumitrescu, radu, tufis}@racai.ro*

[2] *dstfnscu@memphis.edu*

Abstract

During the last years the NLP group at RACAI developed the technologies necessary for statistical machine translation and conducted extensive experiments on the Romanian-English language pair. With support from the ACCURAT and MetaNet4U EU projects, we managed to refine our processing tools and build appropriate language resources for a large set of experiments and evaluations. We describe experiments on domain adaptation as well as the results on self-improvement of translation quality by learning from mistakes made during the training phase. We found that a statistical baseline system can be re-used to automatically learn how to (partially) correct translation errors, i.e. to turn a "broken" target translation into a better one. Without any additional data, what we called a "cascaded" SMT system shows a sensible quality increase in terms of BLEU scores, for both translation directions (English to Romanian and Romanian to English).

Keywords: cascaded translation, domain adaptation, statistical machine translation, translation quality evaluation

1. Introduction

The results of the translation experiments we describe in this article refer to the Romanian-English language pair. However, we strongly believe that the methodology and general conclusions we draw for these experiments are applicable to any language pair. The language pair is challenging because Romanian (a Romance language) and English (a Germanic language) have different morphological productivity and syntactic structures. We collected parallel texts in various domains, from highly specialized to general (encyclopedic) language:

- DGT (juridical domain): the accumulated legislation, legal acts, and court decisions constituting the body of European Union law (Steinberger et al., 2012); 18.5 million words, 1.05 million sentences.
- EPL (juridical domain, different register than DGT): proceedings of the European Parliament (Koehn, 2005); 9.5 million words, 377,000 sentences.
- LIT (literary domain): the "1984" novel written by George Orwell, parallel Romanian-English version; 87,000 words, 6,200 sentences.
- MED (medical corpus): parallel corpus created from documents from the European Medicines Agency (Tiedemann, 2009); 3.5 million words, 220,000 sentences.
- NWS (news domain): SETimes (Tyers, 2010), a parallel corpus of news articles in the Balkan languages, originally extracted from http://www.setimes.com; 2.4 million words, 98,000 sentences.
- SPK (free-speech domain): parallel corpus created from translated TED speeches (Cettolo, 2012); 2.5 million words, 142,000 sentences.
- WIKI5 (Wikipedia domain): parallel corpus extracted from Romanian and English Wikipedia articles using our LEXACC tool (Ştefănescu et al., 2012); 5 million words, 240,000 sentences. Recently an even larger parallel corpus has been extracted from Wikipedia, as reported in (Ştefănescu & Ion, 2013), but here we report experiments on the earlier version WIKI5.
- ALL ("general" domain): concatenation of all seven corpora mentioned above.

Altogether, the seven concatenated corpora (ALL) contain more than 2.16 million parallel sentences and almost 41.5 million words.

All the experiments presented in this article have been performed with the Moses (Koehn et al., 2007) open-source translation software. The phrase-based translation approach has overcome several drawbacks of the word-based translation methods and proved to significantly improve the quality of translated output. We experimented with different factored models that include surface form, lemmas and different part of speech tag sets in various combinations. The first challenge of our research was to identify the optimal feature combination ensuring the best baseline translation system. The second objective of our work was to validate our intuition that a statistical baseline system can be re-used (cascaded) to automatically learn how to (partially) correct its own translation errors, i.e. to turn an initially "broken" translation into a better one.

2. Corpora preprocessing

Before starting any translation experiment, the available resources have to be preprocessed. This is a multi-step process. Initially, we attempt to increase the quality of the Romanian side of the parallel corpora and then annotate both the Romanian and the English side with various types of information.

As such, we performed automatic text normalization on the Romanian side of our corpora. Due to the fact that in earlier versions of the Windows operating system the letters *ş* and *ţ* were initially written as ***ş***, ***ţ*** (with a cedilla underneath – old, incorrect style), only later being correctly written as ***ș***, ***ț*** (with a comma underneath), we currently have several resources with incompatible diacritics for these two letters. The first preprocessing task involved changing all ş and ţ letters to the correct writing style. The second correction made is required due to the 1993 Romanian orthographic reform which re-established the orthography used until 1953. The main effect reflected in our corpora is that the inner letter "î", not preceded by a prefix, has been replaced by "â".

Texts have thus been corrected to the current orthography using an internally developed tool that uses a 1.1 million word lexicon of the Romanian language, backing-off a rule-based word corrector in case the lexicon might not contain some words. The third correction made concerned texts that do not have diacritics. Restoring diacritics is a difficult and error-prone task, as a misplaced or missing diacritic can change the part of speech of a word up to making an entire sentence loose meaning. Using the DIAC tool (Tufiș and Ceaușu, 2008), we were able to carefully restore diacritics where they were missing.

Having corrected the Romanian side of our parallel corpora, the next step of the preprocessing phase was the automatic annotation of both Romanian and English sides. For this task we used TTL (Ion, 2007), with which we annotated every word with its lemma, and with two types of part-of-speech tags: morpho-syntactic descriptors (MSDs) and a reduced tag set (CTAGs), as well as different combinations of them. The tags themselves follow the Multext-East lexical specifications (Erjavec & Monachini, 1997) and the tiered tagging design methodology (Tufiș, 1999).

For example, for the English sentence "*We can can a can.*" TTL provides the following annotation:

We|*we^Pp*| *Pp1-pn*|*PPER1*
can|*can^Vo*| *Voip*|*VMOD*
can|*can^Vm*| *Vmn*|*VINF*
a|*a^Ti*| *Ti-s*|*TS*
can|*can^Nc*| *Ncns*|*NN*
.|.^*PERIOD*|*PERIOD*|*PERIOD*

The first of the four factors for each word is the word itself (e.g. *can*). We label it by #0. The second factor, labeled by #1, is the lemma of the word, linked by the "^" character, to its first two positions in the MSD tag (grammar category and type; e.g. *can^Nc*). The third factor, labeled #2 is the MSD (e.g. *Ncns*) and the fourth factor is the CTAG, labeled #3 (e.g. *NN*). The TTL tool correctly recognizes that the first two *can*'s are verbs (the former being a modal verb while the latter an infinitive) and the last *can* is a noun. Table 1 shows a pair of aligned sentences, as annotated by TTL.

Table 1: EN-RO annotated sentence pair

English	**Romanian**
#0 | #1 | #2 | #3	**#0 | #1 | #2 | #3**
Store|store^Vm|Vmn|VINF	**A**|a^Qn|QN|Qn
in|in^Sp|PREP|Sp	**se**|sine^Px|PXA| Px3--a--------w
the| the^Dd| DM|Dd	**păstra**|păstra^Vm|VN| Vmnp
original|original^Af|	**în**|în^Sp|S|Spsa
package|package^Nc| NN|Ncns	**ambalajul**|ambalaj^Nc| NSRY|Ncmsry
.|.^PERIOD|PERIOD|PERIOD	**original**|original^Af| ASN| Afpms-n
	. |.^PERIOD|PERIOD|PERIOD

Next, we perform true-casing on the corpora. True-casing means lower-casing the first word in every sentence, where necessary (for sentences that are not all-caps). A model is trained on available data, learning what words should not be lower-cased, as acronyms or proper nouns, and applied back to the data. True-casing benefits automatic machine translation when building both the translation model and the language model by reducing the number of surface forms for each possible word (for many words, there will no longer be two identical annotations with the exception that one form is capitalized and one is not).
The last step of the preprocessing phase was to apply Moses's cleansing script. This script removes blank lines, sentences that are longer than a specified number of words (we limit the maximum sentence length at 60 words) and that have significantly different lengths. In addition to this script, we also removed duplicate lines. By the end of the corpora cleansing we retained about 78% of the original data.
From each domain corpus we withheld 1000 test sentences. The final sizes of the domain corpora on which we trained the systems are: DGT – 636,835 sentences, EPL – 363,738, LIT – 4,210, MED – 209,577, NWS – 95,719, SPK – 140,379, WIKI5 – 234,879, and the combined ALL corpus – 1,685,337 sentences.

3. Factored models overview

We had a few different design choices in the building of the baseline: what language model to use, what translation model to use, what decoder type to use (and associated parameters).

For the language modeling (LM), both English and Romanian, we used the data contained in the respective English/Romanian side of our concatenated parallel corpus (ALL). This would give us a domain-independent language model that we can use for all the systems we experiment with. Due to the size differences between our domains,

the resulting language model is not truly domain-independent, but it is as close as possible without penalizing its size. We differentiate between *lexicalized language models* (LLM) and *grammatical language models* (GLM) the only difference between them being the nature of the tokens considered in the n-grams: in LLM the tokens are the surface forms (factor #0 in the output of TTL), while in GLM the tokens are corresponding morpho-syntactic descriptors (factor #3).

The ALL language models (both LLM and GLM) were built with SRILM (Stolcke, 2002) by interpolating the 5-ngram LMs using the Knesser-Ney smoothing method. As one would expect, the size of GLM was significantly smaller than the size of the LLM (e.g. English LLM was 435 MB and contained around 20 million n-grams while the English GLM was 77.5 MB with almost 2.2 million n-grams). We further built LMs (both LLM and GLM) for each domain, all with the same parameters (5gram, interpolated Knesser-Ney).

At the end of the LMs building phase, we had an LLM for each domain in part (DGT, EPL, LIT, MED, NWS, SPK, WIKI5) and one domain independent LLM (containing all the domain corpora put together, named ALL). Similarly, we have a GLM, for each domain and for all of them combined.
Regarding the translation model for the baseline system, we used the most common translation method – direct surface-to-surface translation, meaning that the translation tables and the language models contained only surface forms (the #0 factor in TTL's output). For the word alignment, we used the lemma of the words adjoined by the word's reduced Part-of-Speech tag (factor #1 in TTL's output). This yielded a better word-to-word alignment due to the slightly reduced word space, compared to simply using the surface forms. The same factor #1 is used, instead of the surface form for the reordering model.

The last design choice was what decoder type to use. Moses offers a number of options, from which the most used are the default decoding and the alternative cube-pruned search. The default decoding offers good performance, but it is rather slow. It has two tunable parameters that allow adjusting the performance and the translation time: stack size and beam search width. The cube pruning decoding offers greater speed, but slightly penalizes performance. It also offers an adjustable stack size. Both decoding options produce an n-best list, the top scoring translation being always picked. This behavior can be changed specifying to the decoder that minimum Bayes risk should be used to pick the translated sentence. This option picks the most similar candidate (to the other candidates) from the top n-best translation choices. It offers a compromise between speed and performance, sometimes surpassing the default translation itself. For the baseline system, we chose Moses's default decoder with the default stack size and beam search values.

The initial baseline model for our translation system involves a surface-to-surface translation model (denoted by t0-0 since the surface factor is labeled #0) and a lexicalized language model (denoted by m0). Other configurations for our translation experiments are shown in Table 2.

Table 2: Translation flows variants

Model #	Details	Comments
#1	t0-0 m0	baseline translation flow
#2	t1-1 g1-0 m0	translates lemmas and reduced MSDs in language A to lemmas and reduced MSDs in language B (t1-1) and then employs a generation step to generate surface forms in language B from lemmas and reduced MSDs in language B (g1-0). Fluency coercion is performed by a lexicalized LM (m0). Default reordering model.
#3	t1-1 g1-3 t3-3 g1,3-0 , m0m3	translates lemmas and reduced MSDs in language A to lemmas and reduced MSDs in language B (t1-1), followed by a lemma and reduced MSD to MSD generation in language B (g1-3), a translation of MSDs in language A to MSDs in language B (t3-3) and finally generating surface forms from the previously translated lemmas and MSDs in language B (g1,3-0). Fluency coercion is performed by both a lexicalized LM and a grammatical model (m0m3). Default reordering model.
#4	t1-1 g1-3 t3-3 g1,3-0 , m0m3 r0	Same as model #3 but using a reordering model based on surface forms.
#5	t1-1 g1-3 t3-3 g1,3-0 , m0m3 r3	Same as model #3 but using a reordering model based on lemmas.

Table 3 presents the BLEU scores (Papineni et al., 2002) obtained testing the five proposed models. For the Romanian → English direction, model #3 was the best performing of the five, with a BLEU score of 57.01. For the English → Romanian direction, scores were a bit lower, model #2 having the highest 53.94 BLEU points.

Table 3: BLEU scores for various translation flows

RO → EN		EN → RO	
Model #	**BLEU**	**Model #**	**BLEU**
#1	56.31	#1	52.43
#2	56.49	**#2**	**53.94**
#3	**57.01**	#3	49.97
#4	56.79	#4	49.12
#5	56.89	#5	48.70

The next step was to estimate the translation time of the ALL corpus. Moses offers two different translation options: the default translation search and the cube pruning search

algorithm. There are two adjustable parameters: the stack size and beam search. These parameters have been manually specified to obtain insights about their influence on translation speed and quality. We present only model #3 for the RO→EN direction.

The translation time includes language model and translation/generation tables loading time. The test machine is a dedicated 16 core (8 physical + 8 virtual, running at 2.6GHz), 12 GB RAM server.

Table 4: Model #3 RO-EN: Parameter variation, translation time and BLEU scores.

Stack Size Param.	**Beam Search Param.**	**Translation Time (s)**	**BLEU Score**
(default)	(default)	3074	**57.01**
100	(default)	1611	56.69
50	(default)	831	56.05
20	(default)	391	54.97
15	(default)	307	54.36
10	(default)	229	53.16
5	(default)	144	51.35
(default)	100	83	39.17
(default)	10	83	43.29
(default)	2	87	47.17
(default)	1	93	49.63
(default)	0.5	151	51.80
(default)	0.1	169	55.84
100	1	106	49.63
Cube pruning algorithm with stack size 2000		**167**	56.29

Table 4 shows measurements for the translation times and BLEU scores (RO→EN direction) of the test files (1,200 sentences), for different settings of the Stack Size and Beam Search.

Even though the best performing translation was achieved using the default parameters (BLEU score: 57.01), due to the very long translation time, we found that the best compromise was to use the cube pruning algorithm with the stack size 2,000 that obtains a marginally lower BLEU score of 56.29. When using the cube pruning algorithm, we found that, for our test set, increasing the stack size to more than 2,000 does not generate any noticeable score improvements.

4. Cascaded translation

In order to improve the quality of texts automatically translated, they are usually post-edited by human experts. Trying to speed-up the process of post-editing (Ehara, 2011) presented their EIWA ensemble which is based on a commercial rule-based MT (specialized in patent translation) for the first step and a MOSES-based SMT for the second phase (named statistical post-editing).

In (Tufiș and Dumitrescu, 2012) we introduced the notion of cascaded translation using the same SMT system trained on different parallel data. Except for the training data and the different parameter settings, the two systems are incarnations of the same basic system. The first system S1, trained on parallel data $\{C_A, C_B\}$ learnt to produce draft

translations from L_A to L_B. The second translation system S2, trained on the "parallel" data {S1(C_A), C_B}, learnt how to improve the draft translations.
There are several other methodological differences between our system and the one described in (Ehara, 2011). EIWA does not work in real time because before proper translation of a text T, the SMT post-editor is trained on a text similar to T. The similar text is constructed from a large patent parallel corpus (3,186,284 sentence pairs) by selecting for each sentence in T an average number of 127 similar sentences. Contrary to Ehara (2011), we found that setting the distortion parameter to a non-null value improves the translation quality. Translation of a new, unseen text is achieved in real time (no retraining at the translation time).
Based on the experiments reported in previous chapter, we have used the two best performing models (model #3 for the RO→EN direction and model #2 for the EN→RO direction) with the cube pruning search algorithm to translate each side of the ALL parallel corpus {C_{RO}, C_{EN}}. We obtained two new corpora: for the RO→EN direction we obtained a translated corpus in English paralleled with its reference translation {$T_{S1}(C_{RO})$,C_{EN}}, and for the EN→RO direction, a translated corpus in Romanian paralleled with its reference translation {C_{RO},$T_{S1}(C_{EN})$}. After the translations, the two newly obtained "parallel" corpora were processed as discussed in Chapter 1. Using the same NLP tool we used to annotate the original corpus we annotated the translated corpora with lemma, CTAGs and MSDs.
Each of the two "parallel" corpora was used as training material for a second layer of the translation architecture with the purpose of validating our intuition that a cascaded translation system may improve its translation accuracy by learning from own mistakes.

4.1. Second layer translation system (S2)

Translating from *broken* language L into a better version of L (L being either English or Romanian), we trained 9 models to see which one would perform best. Table 5 shows the models chosen (the notations used in the *Details* column have the same meanings as in Table 2) and Table 6 shows the translation and BLEU scores using the cube pruning and default translation algorithms. The same models were used for both translation directions.

Table 5: Translation flows variants for the second translation system

Model	Details
#1	t0-0 m0
#2	t1-1 g1-0 m0
#3	t1-1 g1-2 t2-2 g1,2-0 m0,m2
#4	t1-1 g1-3 t3-3 g1,3-0 m0,m3
#5	t1-1 g1-3 t3-3 g1,3-0 m0,m3 r3
#6	t1-1 g1-2 t2-2 g2-3 t3-3 g1,3-0 m0,m2,m3
#7	t0,1-0,1 m0
#8	t0,1,2-0,1,2 m0,m2
#9	t1,2-t1,2 m0,m2

The S2 translations (in both directions) were performed using the cube pruning search with stack size 2,000. The reordering model is the Moses default, with the only difference that in model 5 we have used MSDs as the reordering factor.

For testing S2 we used the same test files as for S1, as they were translated with the best S1 models: the model #3 for RO→EN direction and the model #2 for the EN→RO direction (see Table 3). The reference translations for the two directions were T_{EN} and T_{RO} respectively (1,200 sentences each).

For the RO→EN direction the BLEU translation score of the S1+S2 system has been improved from the best S1 model (57.01) to a new BLEU score of 60.90.

The fact that S2 translation based on model #7 (surface form & lemma with reduced MSD to surface form & lemma with reduced MSD using only the surface language model) was the fastest and most accurate is not surprising: we "translated" from partly broken English into presumably better English.

Generation steps in models #2, #3, #4, #5, #6 were more detrimental than useful but the information on the lemma eliminated some candidates from the search space. That observation suggests that there were few inflected word forms to be corrected and most error corrections came from a more precise retrieval of the translation equivalents. Interestingly, the translation time the using default Moses parameters is very close to the cube pruning search (because the chosen model has just phrase translation and no generation component), but yields approximately 0.14 BLEU point increase.

Table 6: RO→EN: S2(S1(T_{RO}))

Model #	**Transl. time (s) with cube pruning**	**BLEU with cube pruning**	**Transl. time (s) with default params.**	**BLEU with default params.**
#1	195	60.42	257	60.65
#2	186	59.59	4745	60.12
#3	175	55.68	4129	56.12
#4	281	55.50	3994	56.18
#5	221	55.45	4104	56.20
#6	244	55.16	5016	55.98
#7	**108**	60.74	143	**60.90**
#8	144	58.50	254	58.61
#9	136	58.50	249	58.61

Table 7 shows that for the EN→RO direction, the S2 system models #7 and #8 have a similar performance, increasing the BLEU score from the original 53.94 points to 54.44 (0.5 BLEU point net increase). As with the RO→EN direction, the S2 models that employ generation steps actually slightly decrease the score.

Table 7: EN →RO: S2(S1(T_{EN}))

Model #	Transl. time (s) with cube pruning	BLEU with cube pruning	Transl. time (s) with default params.	BLEU with default params.
#1	254	54.41	154	54.42
#2	1443	52.14	556	52.55
#3	1051	53.50	594	53.50
#4	543	53.59	798	53.59
#5	530	53.59	613	53.59
#6	805	53.56	997	53.56
#7	282	54.43	**167**	**54.44**
#8	417	54.41	287	**54.44**
#9	403	54.40	280	54.42

Another interesting result was to evaluate the simple cascading systems without feature models, that is (S1=#1)+(S2=#1) and compare their performances with the direct translations and the best feature-models cascaded systems. The results are shown in Table 8.

Table 8: S2(S1(T_{source}))

RO → EN'→EN		EN → RO'→RO	
Model #	**BLEU**	**Model #**	**BLEU**
#1+#1	60.47	#1+#1	54.29
#3+#7	60.90	#2+#7	54.44

The increased accuracy due to various feature combinations versus the baseline system has been apparent from Tables 6 and 7 compared to the results in Table 3. Table 8 shows that the direct translations (S1 with any model) for both directions have BLEU scores lower than the cascaded system (S1+S2) even when feature models were not used (model #1+#1). Thus, we can support the statement that the morphological features and the cascading idea are beneficial to the overall accuracy of translations (at least between Romanian and English). Finally, we took the cascading idea one step further by repeating the entire train-translate process (step 2), obtaining S3(S2*(*S1*(*T_{source}*)))*. We observed that the translation stabilized, with very few sentences being changed (around 1%), and with the changes being minor (increasing or even decreasing the BLEU score by less than ~0.05 points). We concluded that further cascading would not bring significant improvements.

Overall, we obtain a 3.89 BLEU point increase for the RO→EN direction and a smaller 0.5 BLEU point increase for the more difficult EN→RO direction using our cascaded system. In (Dumitrescu et al., 2013) we showed that the cascaded translation is beneficial for translating both *in-domain* and *out-of-domain* input texts.

5. Analysis of errors in cascaded translation

We were interested to see which the most distant translations from the reference were, assuming that these were bad translations. We computed for each sentence *I* the similarity scores SIM between its translations and the reference translation. These scores were computed with the same BLEU-4 function used for bitexts. Similarly to the BLEU score applied to a bitext, 100 means perfect match and 0 means complete mismatch. Thus, we obtained 1,200 pairs of scores SIM^I_{S1} and SIM^I_{S1+S2}. We also compute the average similarity scores as $\frac{1}{1200}\sum_{I=1}^{1200} SIM^I_{S\alpha}$ where S_α is S1 or S1+S2. As expected, the average SIM scores make the same ranking as the BLEU scores, although they are a bit higher (ex: 61.18 for S1 and 63.58 for S1+S2 for the RO→EN direction).

We briefly comment on the results of this analysis for the Romanian-English translation direction. We manually analysed the test set translations. We identified 3 sentences with their translations having a zero SIM score for both systems. The explanation was that the reference translation was wrongly aligned to the source sentence.

S1 produced 72 perfect translations (score 100) while S1+S2 produced 105. Only 57 perfect translations were common to S1 and S1+S2, meaning that S1+S2 actually deteriorated a few of the original correct translations. By analyzing the 15 translations that were "deteriorated" we noticed that they were identical, except that unlike S1+S2, S1 and Reference translations either had a differently capitalized letter that marginally lowered the score or had multiword units joined by underscores (e.g. *as well as* vs. *as_well_as*). This was a small bug which has been removed and which, overall, brought a 0.05 increase in the BLEU score.

The capitalization and punctuation are other sources of lower scoring against the reference. All these examples show the sensitivity of the BLEU scoring method, especially for very short sentences.

Another important variable to note is the amount of change from one layer to the other: out of all sentences, around 37% had a BLEU increase while around 20% had a BLEU decrease (but see the comment on the underscore difference), the rest 43% have not been changed in any way. The table below shows some examples with differences between the standard translation system and the cascaded one.

Table 9: Comparison between the standard SMT and the cascaded SMT

S1 BLEU	S1 Sentence	S2 BLEU	S2 Sentence	Reference
0.36	the area is situated in the Golful normand-breton , in the southern part of the Mânecii	0.83	the area is situated in the Normano-Breton Gulf , on the south side of the Mânecii	the area is situated in the Normano-Breton Gulf , on the south side of the English Channel .
0.58	other information , provided that they have a suitability and reliability can be reasonably demonstrated . '	0.84	other information , provided that its suitability and reliability can be reasonably demonstrated . '	other information provided that its suitability and reliability can be reasonably demonstrated . '
0.3	(13 februarie 1934) is an American actor , film and television .	0.58	(February 13 , 1934) is an American film and television actor .	(February 13 , 1934) is an American film , stage and television actor .
0.46	Speer was made available to historians and other scholars .	0.35	Speer was made available to historians and scholars .	Speer made himself widely available to historians and other enquirers .

We can see that in general, sentences are improved. This usually happens in three distinct ways:

- S1 fails to translate words which are subsequently translated by S2. While counter-intuitive, because basically no new information is added to the system, this happens because in S1's phrase table some phrases are automatically pruned, leaving for example unigrams that are found in the training corpus but do not exist in the phrase table. S2, on the other hand, does not miss these unigrams and therefore, it translates them as is presented in the first sentence in table 9.
- Better word ordering. In the third sentence in table 9 we can see that S2 translates February and then reorders the day and month to match the English date format.
- General phrase substitution. Sometimes there are more appropriate phrases, as it is shown in the second sentence. While the S1's translation might be accurate from a Romanian word-for-word perspective, S2 manages to find a better phrase than '… provided that they have a ...'

However, the system sometimes degrades sentences, usually by shortening them. In example 4, we see that 'other' from S1's translation is removed by S2.

6. Experiments on translating in-domain versus out-of-domain texts

The astute reader might have noticed that our evaluations used **in-domain** data. The text data have been randomly extracted from the ALL corpus. Although, not seen during the training phase, the test data qualifies as in-domain data. In (Dumitrescu et al., 2012) we provided a detailed analysis of experiments with several translation systems,

corresponding to the 7 distinct domains (see *Introduction* section) plus the system trained on the concatenation of the 7 domain specific corpora (ALL corpus). As these types of experiments are very time consuming we considered only one translation direction (RO-EN).

Depending on the way the training was performed, we obtained 48 different performing MT systems: 16 baseline (see Table 2, model #1: t0-0 m0) and 32 factored translation systems. Eight baseline systems were generated as surface phrase-based systems from the domain specific corpora using the same language model built from the ALL corpus and their performances are shown in Table 10. The other eight baseline systems were generated only from the respective domain specific corpora and their performances are shown in Table 11.

Table 10: Translation results using the baseline systems with the domain-independent LM

		Test domain						
		DGT	EPL	LIT	MED	NWS	SPK	WIKI5
Model trained on domain corpus	DGT	***51.45***	31.98	9.80	34.16	27.74	15.45	21.1
	EPL	37.73	***40.97***	13.13	29.97	31.98	23.34	22.74
	LIT	8.31	8.76	*14.09*	12.44	9.01	11.49	12.48
	MED	25.76	18.7	6.97	***54.54***	15.85	12.05	15.81
	NWS	26.15	31.82	10.47	25.83	***40.07***	20.02	22.21
	SPK	20.21	28.08	13.75	26.96	24.33	***27.95***	22.7
	WIKI5	30.67	31.59	13.35	31.66	32.2	22.11	*29.51*
	ALL	51.43	40.89	**18.00**	53.46	39.31	26.73	**29.95**
	(diff)	*0.02*	*0.08*	*-3.91*	*1.08*	*0.76*	*1.22*	*-0.44*

Table 11: Translation results using the baseline systems with the domain-dependent LM

		Test domain						
		DGT	EPL	LIT	MED	NWS	SPK	WIKI5
Model trained on domain corpus	DGT	***51.94***	25.24	7.55	22.17	19.16	10.98	16.51
	EPL	26.99	*40.85*	10.96	19.34	24.51	19.89	18.51
	LIT	6.93	6.33	*14.33*	10.34	7.60	8.79	12.3
	MED	17.24	12.11	5.18	***55.34***	12.60	8.39	13.58
	NWS	15.64	24.31	8.3	16.48	***40.23***	15.24	19.05
	SPK	10.42	18.91	11.49	15.34	16.24	***28.65***	16.9
	WIKI5	18.10	21.99	11.07	21.03	24.08	17.83	***29.99***
	ALL	51.43	**40.89**	**18.00**	53.46	39.31	26.73	29.95
	(diff)	*0.51*	*-0.04*	*-3.67*	*1.88*	*0.92*	*1.92*	*0.04*

The 32 factored translation systems were constructed as described above, in Chapter 3, selecting the best performing model flow for our language pair and direction (RO->EN). As shown in Table 3, this was the model #3: t1-1 g1-2 t2-2 g1,2-0 m0,m2. Depending on how the language models m0 and m2 were built and used in the respective factored translation systems we generated the following ones:

- 8 systems using LLM (m0) and GLM (m2) generated from ALL corpus; their performances are shown in Table 12.
- 8 systems using LLM (m0) generated from ALL corpus and GLM (m2) generated from the domain specific corpora; their performances are shown in Table 13.
- 8 systems using LLM (m0) and GLM (m2) generated from the domain specific corpora; their performances are shown in Table 14.
- 8 systems using LLM (m0) generated from the domain specific corpora and GLM (m2) generated from ALL corpus; their performances are almost identical to those shown in Table 14 and are not discussed here.

Table 12: Factored translation results using domain-independent lexical and grammatical LMs

		Test domain						
		DGT	EPL	LIT	MED	NWS	SPK	WIKI5
Model trained on domain corpus	DGT	***46.43***	30.72	10.02	30.91	25.43	15.95	21.16
	EPL	33.32	***39.12***	12.86	27.05	28.62	22.67	22.47
	LIT	9.3	9.64	*14.31*	15.8	10.57	13.32	13.44
	MED	23.45	18.87	7.1	***48.65***	15.76	13.04	16.07
	NWS	25.51	30.7	11.04	24.42	**38.03**	19.9	22.83
	SPK	19.71	26.83	13.28	24.28	22.9	***26.66***	21.69
	WIKI5	28.21	29.59	13.19	29.43	29.65	21.7	***28.55***
	ALL	45.63	38.08	**16.11**	45.72	35.77	25.64	28.08
	(diff)	*0.8*	*1.04*	*-1.8*	*2.93*	*2.26*	*1.02*	*0.47*

Table 13: Factored translation results using domain-independent lexical LM and domain dependent grammatical LM

		Test domain						
		DGT	EPL	LIT	MED	NWS	SPK	WIKI5
Model trained on domain corpus	DGT	***46.51***	30.55	9.94	30.35	24.73	15.58	20.91
	EPL	32.98	***39.06***	12.82	26.12	28.06	22.48	21.96
	LIT	9.13	9.4	*14.43*	15.67	10.42	12.98	13.28
	MED	23.1	18.41	6.71	***49.55***	15.64	12.54	15.91
	NWS	25.1	30.66	10.92	24.17	***38.49***	19.33	22.61
	SPK	19.32	26.4	13.12	23.45	22.58	***26.82***	21.53
	WIKI5	27.88	29.12	13.17	28.64	29.48	21.28	***28.77***
	ALL	45.63	38.08	**16.11**	45.72	35.77	25.64	28.08
	(diff)	*0.88*	*0.98*	*-1.68*	*3.83*	*2.72*	*1.18*	*0.69*

Table 14: Factored translation results using domain-dependent lexical and grammatical LMs

		Test domain						
		DGT	EPL	LIT	MED	NWS	SPK	WIKI5
Model trained on domain corpus	DGT	***46.31***	26.79	8.85	23.91	20.35	13.1	18.16
	EPL	27.11	***38.85***	11.79	21.48	24.93	21.28	20.02
	LIT	7.91	7.71	*13.74*	13.9	9.17	10.98	12.55
	MED	17.84	14.19	5.55	***49.48***	13.56	10.83	14.38
	NWS	18.09	26.38	9.45	19.07	***38.6***	17.39	20.38
	SPK	13.37	20.27	11.65	17.45	18.42	***26.51***	18.54
	WIKI5	19.43	23.75	12.02	22.47	25	19.68	***28.82***
	ALL	45.63	38.08	**16.11**	45.72	35.77	25.64	28.08
	(diff)	*0.68*	*0.77*	*-2.37*	*3.76*	*2.83*	*0.87*	*0.74*

We also performed experiments on language model adaption (Dumitrescu et al., 2013). This meant, in essence adding more sentences to an existing LM and testing to see whether the new, larger LM increased translation scores. We extracted the new sentences from a large monolingual English corpus that belongs mostly to the news genre. The sentence selection was based on each sentence's individual perplexity versus the domain-independent language model. We experimented by adding batches of 500,000 sentences to the domain-independent language model and re-testing the ALL baseline translation system on each domain to see the variation of the BLEU score. We showed in (Dumitrescu et al., 2013) that the translation accuracy does not necessarily improve with the increase of monolingual data for building language models (see also Table 15 below).

Table 15: Evaluation of the optimal extension of the LM training data for the Baseline translation system (ALL) on various domains

Test domain	**Baseline BLUE score**	**Maximum BLUE score (*variation*)**	**LM size (number of additional sentences)**
DGT	51.43	52.76 *(+1.33)*	7,500 K
EPL	40.89	41.73 *(+0.84)*	13,500 K
LIT	18.00	18.10 *(+0.1)*	4,500 K
MED	53.46	53.46 (0; never increased)	0 K
NWS	39.31	40.10 *(+0.79)*	14,000 K
SPK	26.73	27.89 *(+1.16)*	7,000 K
WIKI5	29.95	30.00 *(+0.05)*	14,000 K

What do the results summarized in Tables 10-15 demonstrate?

- Domain-specific SMTs, either surface-form based or factored with LLM and GLM, ensure always better quality translation (measured in terms of BLEU score) only for *in-domain* input texts. If the texts expected to be translated are *out-of-domain* the quality of translations deteriorates seriously. The deterioration is more significant as the register of test data is more distant from the one of the training data.

- The surface based SMTs are usually performing better than factored SMTs both on *in-domain* and *out-of-domain* test data **only if the training data is sufficiently large** (in the range of several hundred thousands of parallel sentences). This is not necessarily true when the target language has a highly inflectional morphology. In the case of small training data sets, the factored SMTs are a much better approach.
- While parallel data may not be available in sufficient quantities[1] for under-resourced language pairs, monolingual data in either languages may exist in much larger quantities. Building language models as large as possible pays off, especially when using factored translation models.

Trying to rank various ways for building SMT systems, according to our experiments, we would say that using factored-based approaches with domain independent language models (LLM and GLM) is the most promising way for dealing with arbitrary texts. The least recommended solution (for translating arbitrary texts) is using an SMT built only from a specific domain.
However, the ideal solution would be to build a large domain ontology and translation systems for each category of the ontology. Text classification is a task which already gets very high precision figures. Such a text classifier might direct the input text of a recognized domain to the appropriate domain MT system. Texts for which the classification is uncertain (either because the text is too short or because there is no matching category in the domain ontology) should be dealt with a global general system (of the type *ALL* in our experiments). Such a solution would work more satisfactorily than any present ones.

7. Conclusions

We presented various architectures for SMT systems, both surface and factor-based and commented on their behavior in translating in-domain and out-of-domain texts, showing that, according to our experiments, using factored-based approaches with domain independent language models (LLM and GLM) are most promising for dealing with arbitrary texts. The least recommended solution (for translating arbitrary texts) is using SMT systems built only from a specific domain. We addressed the issue of language model domain adaptation for the quality of produced translations arguing that building language models as large as possible pays off, especially when using factored translation models. The novel concept of cascaded translation architecture (Tufiș and Dumitrescu, 2012), has been shown to be able to learn from its own errors and sensibly improve the translation quality for both in-domain and out-of-domain texts.

Acknowledgments. This work has been supported by the ACCURAT project (www.accurat-project.eu/) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under the Grant Agreement n° 248347.

[1] With the general increased interest in exploiting comparable corpora, new tools for extracting quasi-parallel data showed efficient solutions to compensate the lack of really parallel data. See for more details (Ștefănescu et al. 2011), (Tufiș, 2012) and the site http://www.accurat-project.eu

References

Avramidis E., Koehn, P. *(*2008). Enriching morphologically poor languages for statistical machine translation. *Proceedings of Association for Computational Linguistics / HLT*, 763–770, Columbus, Ohio,.

Cettolo, M., Girardi, C., Federico, M. (2012). WIT3: Web Inventory of Transcribed and Translated Talks. *Proc. of EAMT*, Trento, Italy, 261-268.

Dumitrescu, Ș. D., Ion, R., Ștefănescu D., Boroș, T., Tufiș, D. (2013). Language and Translation Models Adaptation for SMT. *Towards Multilingual Europe 2020: A Romanian Perspective. Romanian Academy Publishing House,* 205-224, Bucharest, Romania (Tufiș, D., Rus, V., Forăscu, C. eds.)

Ehara, T. (2011). Machine translation system for patent documents combining rule-based translation and statistical postediting applied to the Patent MT Task. *Proceedings of NTCIR-9 Workshop Meeting*, Tokyo, Japan, 623-628.

Erjavec, T., Monachini, M. (1997). Specifications and Notation for Lexicon Encoding. Deliverable D1.1 F. Multext-East Project COP-106. http://nl.ijs.si/ME/CD/docs/mte-d11f/,

Tyers, F. M., Alperen, M. (2010). South-East European Times: A parallel corpus of the Balkan languages.

Habash, N., Dorr, B., Monz, C. (2006). Challenges in Building an Arabic-English GHMT System with SMT Components. *Proceedings of AMTA'06*, Cambridge, MA, USA.

Ion, R. (2007). Word Sense Disambiguation Methods Applied to English and Romanian, PhD thesis (in Romanian). Romanian Academy, Bucharest.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit*.

Koehn, P., Hoang, H. (2007). Factored Translation Models. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, 868–876.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, demonstration session, Prague.

PAPINENI, K., ROUKOS, S., WARD, T., ZHU W.J., *BLEU: A Method for Automatic Evaluation of Machine Translation*. In Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, Philadelphia, 2002.

Ștefănescu, D., Ion, R., Hunsicker, S. (2012). Hybrid Parallel Sentence Mining from Comparable Corpora. *Proceedings of the 16th Conference of the European Association for Machine Translation* (EAMT 2012), Trento, Italy, 137-144.

Ștefănescu, D., Ion, R. (2013). Parallel-Wiki: A Collection of Parallel Sentences Extracted from Wikipedia. *Proceedings of the 14th International Conference on*

Intelligent Text Processing and Computational Linguistics (CICLING 2013), University of the Aegean, Samos, Greece.

Steinberger R., Andreas E., Szymon K., Spyridon P., Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. *Proceedings of the 8th international conference on Language Resources and Evaluation* (LREC'2012), Istanbul.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. Spoken Language Processing*, Denver, USA.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. Recent Advances in Natural Language Processing (vol V), (N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov eds.), 237-248.

Tufiş, D., Ceauşu, A. (2008). DIAC+: A Professional Diacritics Recovering System. *Proceedings of LREC 2008*, ELRA - European Language Resources Association, Marrakech, Morocco.

Tufiş, D. And Dumitrescu, S.D. (2012). Cascaded Phrase-Based Statistical Machine Translation Systems. *Proceedings of the 16th Conference of the European Association for Machine Translation* (EAMT 2012), Trento, Italy, 129-136.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. *Text, Speech and Dialogue LNCS* vol. 1692, Springer-Verlag Berlin Heidelberg (F. Jelinek & E. Nth eds), 28-33.

Tufiş, D. (2012). Finding Translation Examples for Under-Resourced Language Pairs or for Narrow Domains; the Case for Machine Translation. *Computer Science Journal of Moldova*, Academy of Sciences of Moldova, Institute of Mathematics and Computer Science, ISSN 1561-4042, vol.20, no.2(59), 227-245.

STATISTICS ON DERIVATION AND ITS REPRESENTATION IN THE ROMANIAN WORDNET

VERGINICA BARBU MITITELU

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest, Romania

vergi@racai.ro

Abstract

In this article on the Romanian Wordnet we focus on the derived words it contains and on the derivational relations that have been marked in the network. We briefly present the methodology adopted for adding such morphologic relations and their associated semantic labels. The main part of the article contains statistics on the morphological links, on the number of semantic labels, correlations with the data on affixes productivity and their semantic values found in linguistic studies and interpretation of the differences.

Keywords: derivational relations, morphological links, lexical semantics, Romanian Wordnet

1. Introduction

Work on derivation (i.e. the language-internal means of creating new words by attaching affixes to another word) is of constant interest for linguists. The mechanisms of this morphologic process do not change throughout time. However, the constant interest is justified by two phenomena manifested in diachrony: on the one hand, there are variations in the productivity of (certain) affixes and even in their stylistics; on the other hand, the inventory of affixes slightly changes due to the fact that borrowings contribute not only new words to a language, but sometimes new affixes, as well.

When talking about the derivation types two criteria are considered: the mechanism by which the derived word is obtained from the root and the position in the word where it occurs. According to the former, there are three types of derivation:

- progressive derivation – when the derived word is obtained by adding affixes (one or more at a time) to the root: *bucătar<bucată+-ar*;
- regressive derivation (usually called back-formation in the English linguistics literature) – when the derived word is obtained by removing an affix from the root: *cais<caisă*;

- zero derivation – when the derived word and the root have the same structure and, thus, no addition or removal of affixes occur: *fermecător* (adverb) < *fermecător* (adjective). Although in the English literature this morphologic phenomenon is called zero derivation, in Romanian linguistics it is called conversion or changing of the morphologic category and considered a distinct language-internal means of enriching the vocabulary. This is also our understanding here.

According to the latter criterion, there are three types of derivation in Romanian (in other languages one can find other types, as well, such as infixation):

- suffixation – when a string of letters (i.e. the suffix) is added to the end of the root: *bucătar<bucată+-ar*;
- prefixation – when the string (i.e. the prefix) is added to the beginning of the root: *dezlipi <dez- +lipi*;
- parasynthetic – when a string is attached at the beginning and another at the end of the root in the same derivational step: *împăduri< în-* (with its phonetic allomorph *îm-*) + *pădur[e] + -i*.

Semantically, a derived word inherits some of or all the semantic content of the root and benefits from a semantic contribution of the affix(es), too. Just like words, affixes display mono- or polysemy. Linguists focusing on the study of affixes have also described their semantic values. For example, the Romanian suffix *-tor* can help create nouns designating: jobs and the persons doing that job (*muncitor*) and instruments (*sucitor*) (Pascu, 1916).

Speakers have knowledge of the derivational mechanisms available in their language and the instruments used by them. A proof for this is the ease with which ad-hoc derivatives are created both by adults and by children.

Words derived from the same root make up a word family. Psycholinguistic studies (Marslen-Wilson, 2007) have proved that derived words are separate lexical entries in the mind lexicon, so they are not derived every time the speaker needs them. A proof for this is the fact that derived words have sometimes developed secondary meanings from the original meaning resulted from the attachment of the affix to the root. All members of a word family are somehow connected in the mental lexicon. Thus, a resource like the wordnets, which aim at representing lexical knowledge in the same way it is represented in the speaker's mind, needs to also include the relations between members of word families.

Research and work for adding such relations in the wordnets have been reported for Turkish (Bilgin et al., 2004), for English (Fellbaum et al., 2007), for Czech (Pala & Hlaváčková, 2007), for Polish (Piasecki et al., 2009), for Estonian (Kahusk et al., 2010).

2. Types of Relations in RoWN

Just like any wordnet aligned to the Princeton WordNet (PWN) (Fellbaum, 1998), the Romanian Wordnet (RoWN) (Tufiş et al., 2004) contains two types of relations: semantic and lexical. The former have a conceptual nature and are established between

synsets. Such relations are: hyponymy, meronymy, troponymy, lexical implication, cause. They are represented as relations between two synsets. The interpretation is that the relation holds between whichever two literals, each from a different synset in the relation.

Lexical relations are established between literals. They are: synonymy, antonymy, and derivative relations. Synonymy is established between any two literals within the same synset, while antonymy and derivative relations are established between two literals each from a different synset. Synonymy is represented by mere enumeration of words (and associated senses) in the same synset. Antonymy is represented as a semantic relation in RoWN, although in PWN it is lexical in nature. When transferred from PWN, it was named *near-antonymy*. In PWN antonymy is represented as a relation between literals, which are defined by their position in the synset in which they occur. Derivational relations in RoWN are represented at the level of literals specified by their position in the synset.

A word can have more meanings. All relations, irrespective of their type, are established at the sense level, not at the level of the word as a complex of meanings, that is not all synsets in which a word occurs enter the same (types of) relations. This is an important remark, especially in the case of derivation. On the one hand, there is no distinction between polysemous words and homonyms and, on the other, secondary meanings of a word can be too distant to its primary meaning. Once one has identified the pairs root-derived word, there appears the need to hand validate all the possible results of the Cartesian product of the sets of synsets in which the root and derived word occur. If a semantic relation can be established between two meanings of the pair root – derived word, then in our wordnet we define a lexical relation between the respective meanings and we can add a semantic label to this relation. This label can be safely generalized to hold at the level of the synsets to which the two senses belong due to its semantic nature. For example, for the synsets {*conduce, şofa*} and {*conducător, şofer*} we mark derivational relations between *conduce* and *conducător* and between *şofa* and *şofer*; the label AGENT corresponds to these relations at the synset level: for a sentence like "Un şofer conducea maşina cu o alcoolemie de 1.59 mg/l." *şofer* must be interpreted as the AGENT of the verb *conducea*.

Such semantic labels can be considered further semantic relations in the wordnet. In our opinion, provided the languages idiosyncrasies, these labels can be safely transferred from one language into another, if the wordnets for the two languages are aligned to each other (usually via PWN). This transfer is very common in the case of all semantic relations. At the morphologic level, in the synsets between which we have such labels there may be words derivationally related or not: see the pairs *bucătărie – bucătar* (in Romanian) and *kitchen – cook* (in English).

3. Characteristics of derivational relations

Derivational relations are established between a derived word and its root. They have the following properties:

- symmetry – if a word A is in derivational relation with the word B, then B is also in derivational relation with A. This allows for a uniform and easier treatment of derivation and back-formation.

- transitivity – if a word A is in derivational relation with the word B and B is in derivational relation with the word C, then A is also in derivational relation with C; thus, a derivational chain is formed, containing the root and the subsequently derived words; for example, due to this property of derivation, one can identify in RoWN a derivational chain such as: *pădure – împăduri – reîmpăduri – reîmpădurire – nereîmpădurire*.
- non-reflexivity – a word A is not in derivational relation with itself (so we do not treat zero derivation as a type of derivation, as already mentioned in section 1 above).

Derivational relations are language specific. They cannot be automatically transferred from one language into another. For example, in Romanian there is a derivational relation between *bucătărie* and *bucătar*, but, although the equivalent terms exist in English as well, they are not morphologically related: *kitchen* and *cook*. So, we can transfer the semantic label (as presented in section 2), but not the derivational relations.

Note that the properties above hold for derivational relations, not for the semantic labels as well.

4. The Method for Adding Derivational Relations to the RoWN

In order to add derivational relations to the RoWN, they need to be identified first. For this, we made use of the list (called LL) of simple literals (i.e. we disregarded literals made up of at least two words, such as *floarea-soarelui* or *lumină_intermitentă*) from the RoWN and a list of Romanian affixes (called LA). We identified the pairs of literals that obey the following formula:

$$\text{literal}_1 + \text{affix} = \text{literal}_2$$

where $\text{literal}_1 \in$ LL and $\text{literal}_2 \in$ LL, with $\text{literal}_1 \neq \text{literal}_2$, while affix $\in$ LA and can be either a prefix or a suffix.

The Romanian Explanatory Dictionary contains etymologic information. We could have opted for retrieving the root-derived pairs from it. However, there are a couple of inconveniences of this method: on the one hand, RoWN contains more literals than the dictionary, so some cases would have been missed; on the other hand, we aimed at treating similarly words derived in Romanian and borrowings that can be analyzed, so that have a parallel structure and meaning to other cases that were derived in Romanian.

Thus, we identify the pairs derived-root in which the derived word is created either by progressive or regressive derivation. So, we did not deal with parasynthetic derivation and neither with the words presenting alternation(s) in the root via this method. However, afterwards, such cases were extracted from the dictionary and marked in the RoWN.

This automatic identification was followed by manual validation, in which pairs such as *abate – abator*, *veni – deveni* were discarded.

The following table contains some data about this phase of our work:

Table 1: Pairs of root-derived words

	LA	**LL**	**Pairs**	**Correct pairs**	**Per cent of correct pairs**
Prefixes	83	31872	2862	1990	70%
Suffixes	409		13556	8452	62%
TOTAL	492		16418	10442	64%

In our approach, we consider that there is a derivational link between two literals iff there is a morphological relation (of the type derivation) between the literals AND there is a semantic similarity between the literals. As a consequence, the previously validated pairs must undergo further validation, this time at the sense level.

Thus, for each pair, we extracted from RoWN all synsets in which each of its members occurs and thus populated a set for each literal. Then we calculated the Cartesian product of the sets corresponding to the two literals in a pair. We obtained the data in the following table:

Table 2. Annotated pairs

	Validated pairs (word level)	**Subject to annotation (word sense level)**	**Annotated pairs**	**Per cent**
Prefixes	1990	30132	3145	10.43
Suffixes	8452	25717	13916	54.11
TOTAL	10442	55849	17061	30.55

The semantic condition imposed on the pairs of derived-root words and the regularities that are obvious among these pairs allow us to add a semantic label to the morphologic relation. Such labels have been used in other wordnets (PWN – see Fellbaum et al., 2007; Turkish – see Bilgin et al., 2004; Czech – see Pala and Sedláček, 2005; Estonian – see Kahusk et al., 2010). When establishing the list of semantic labels to use, we tried to ensure the following:

- a high degree of generality;
- a different name from the relations already marked in the wordnet;
- a reduced number of labels;
- multilingual character – whenever available, we made use of labels already defined by other teams, which is meant to facilitate further comparisons among languages.

We have defined a set of 57 labels: 16 for prefixed words (TOGETHER, SUBSUMPTION, OPPOSITION, MERO, ELIMINATE, ITERATIVE, THROUGH, REPEAT, IMPLY, SIMILITUDE, INSTEAD, AUG, BEFORE, ANTI, OUT, BACK) and 41 for suffixed ones (SUBSUMPTION, MEMBER_HOLO, MEMBER_MERO, SUBSTANCE_HOLO, SUBSTANCE_MERO, INGREDIENT_HOLO, HOLONYM, PART, AGENT, RESULT, LOCATION, OF_ORIGIN, JOB, STATE, PERIOD, UNDERGOER, INSTRUMENT, SOUND, CAUSE, CONTAINER, VEHICLE,

BODY_PART, MATERIAL, DESTINATION, GENDER, WIFE, DIM, AUG, OBJECT_MADE_BY, SUBJECT_TO, BY_MEANS_OF, CLOTHES, EVENT, ABSTRACT, COLOUR, TAX, MAKE_BECOME, MAKE_ACQUIRE, MANNER, SIMILITUDE, RELATED). It is worth mentioning that the set is enriched throughout annotation, whenever necessary. We do not define labels for which we lack examples in the set of pairs.

5. Statistics

5.1. Number of simple literals – number of derived words

In the RoWN (the version we worked with contained 59869 synsets) there are 31872 simple literals. Among them we could find 16418 pairs of root-derived words, distributed into the types of derivation as in Table 1. Note that one literal can occur even in more than one pair: for example, *reîmpărudi* occurs in the pairs *împăduri – reîmpăduri* and *reîmpăduri – reîmpădurire*. After subjecting the pairs to validation, we found that only 10442 of them were correct. This means that at least one third of the simple literals in RoWN are derived or analyzable words.

Considering the distribution of derived words into the types of derivation we have worked with, it is worth mentioning that almost one fifth of the derived words are prefixed and the rest are suffixed. This can be correlated with the linguists' "axiom" that suffixation is more productive than prefixation (Puşcariu, 1940; Avram, 1989), although we were not able to find in the literature an indication of the ratio of prefixed and suffixed words in Romanian.

Table 3: Prefixation versus suffixation

	Correct pairs	**Per cent**
Prefixation	1990	19%
Suffixation	8452	81%
TOTAL	10442	

5.2. Affixes and their representation in RoWN

As can be noticed in Table 1, the total number of Romanian affixes that we could find is 492. However, only 261 of them occur in the derived words in our RoWN, distributed as follows:

Table 4: Affixes productivity

	Total in the literature	**Found in RoWN**	**Per cent**
Prefixes	83	64	77%
Suffixes	409	197	48%
TOTAL	492	261	

The (19) prefixes not occurring in the literals in RoWN are: *antre-*, *cis-*, *do-*, *en-*, *ento-*, *intra-*, *iz-*, *întru-*, *o-*, *ob-*, *pod-*, *poi-*, *se-*, *spre-*, *tă-*, *tra-*, *tră-*, *vă-*, *ză-*. About three of them we know that they appear only dialectally in Banat: *do-*, *iz-*, *ză-*, while the others occur in some borrowings but have never been productive in Romanian (Avram, 1978). Moreover, these borrowings are not frequent in the language and they do not designate important concepts, thus explaining their absence from the RoWN and, implicitly, of the prefixes.

The most frequent prefixes are *ne-* (467 occurrences), *re-* (262), *in-* (180) and *de-* (121). It is interesting to note that the prefix *ne-* occurs in our electronic explanatory dictionary (EXPD, which has around 70000 entries) in 218 words, while in RoWN it occurs in 467 words. One possible explanation is the fact that lexicographs usually do not include all words derived with very productive affixes (such as *ne-*, *re-*) in dictionaries. Their contribution to the meaning of the derived word is very predictable and, formally, they do not raise any problems either. Another explanation is the fact that borrowings, although analyzable, are not analysed in EXPD, but their foreign etymon is indicated, while we do not distinguish between words created internally and analysable borrowings.

Stoichiţoiu-Ichim (2007) noticed the increased productivity of negative prefixes and of the repetitive *re-* in contemporary Romanian, which can also explain, partially, our bigger number of derived words with these prefixes, while they are not yet listed in dictionary, due to their novelty.

The most frequent suffixes are: *-re* (1822 occurrences), *-ie* (238 occurrences), *-ic* (207), *-ist* (182), *-ism* (176), *-ător* (137), *-ar* (124), *-eală* (109), *-os* (118), *-or* (107), *-a* (655), *-iza* (196), *-i* (189), *-e* (198 occurrences). The very high number of occurrences of the suffix *-re* can be explained by the high number of deverbal nouns in RoWN. Most of the words suffixed with *-re* have been labeled as EVENT. The distribution of the verbal suffixes *-a* and *-i* is the expected one: the former is more frequent than the latter, as a consequence of the higher productivity of the first conjugation as compared to the forth.

5.3. Semantic labels in RoWN

In the following table we present the number of occurrences of each label in the RoWN:

Table 5. Semantic labels for prefixes and their frequency in RoWN

Label	**Occurrences**	**Example**
BACK	2	reflux-flux
TOGETHER	29	întreţese-ţese
AUG	5	supraabundenţă-abundenţă
OUT	1	epidermă-dermă
SIMILITUDE	61	reţine-ţine
IMPLY	26	desconsidera-considera
THROUGH	5	răzbate-bate
MERO	17	suprafaţă-faţă
BEFORE	14	preambalare-ambalare
OPPOSITION	792	neesenţial-esenţial
REPEAT	305	reaprinde-aprinde
SUBSUMPTION	363	subclasă-clasă
ANTI	10	anticolinesterază-colinesterază
INSTEAD	6	vicepreşedinte-preşedinte
ITERATIVE	2	răsfoi-foaie
ELIMINATE	9	deşela-şale

Table 6. Semantic labels for suffixes and their frequency in RoWN

Label	Occurrences	Example
RELATED	1294	călduros-căldură
SOUND	163	bufnealã-bufni
STATE	284	îndoială-îndoi
DESTINATION	5	patentant-patenta
AUG	1	grăsan-gras
SIMILITUDE	115	încărcătură-încărcare
PERIOD	43	bătrâneţe-bătrân
JOB	179	semănător-semăna
PART	12	optime-opt
MEMBER_MERO	17	orăşean-oraş
BY_MEANS_OF	104	opreliște-opri
CAUSE	19	umezeală-umezi
MEMBER_HOLO	37	soldăţime-soldat
RESULT	227	tencuială-tencui
SUBJECT_TO	19	chinui-chin
ABSTRACT	490	cerinţă-cere
SUBSUMPTION	42	căpetenie-cap
OF_ORIGIN	29	sătean-sat
EVENT	699	împărtăşanie-împărtăşi
INSTRUMENT	84	ondulator-ondula
INGREDIENT_HOLO	1	sticlărie-sticlă
TIME	1	cătănie-cătană
MANNER	436	primejdios-primejdie
MAKE_ACQUIRE	110	îndigui-dig
CONTAINER	17	afişier-afiş
HOLONYM	26	pieptar-piept
DIM	50	căsuţă-casă
OBJECT_MADE_BY	50	chinezărie-chinez
CLOTHES	1	pieptar-piept
SUBSTANCE_HOLO	2	cerat-ceară
AGENT	394	linguşitor-linguşi
LOCATION	87	cărămidărie-cărămidă
MATERIAL	4	îndulcitor-îndulci
UNDERGOER	47	setos-sete
COLOUR	19	cenuşiu-cenuşă
GENDER	13	călugăriţă-călugăr
SUBSTANCE_MERO	1	ricină-ricin
MAKE_BECOME	89	caricaturiza-caricatură

Whenever a semantic relation was found between the synsets containing the literals in the pairs of the Cartesian product, we did not add a further semantic label. There is only one exception to this rule, namely the adding of the label DIM even if there is a hyponymy/hypernymy relation between the two synsets. For example, *clopoţel* (ENG30-04146504-n) is a hyponym of *clopot* (ENG30-02824448-n) and we further marked the former as a diminutive of the latter.

5.4. Same PoS – cross PoS relations

Adding derivational relations to a wordnet contributes to the increase of density of relations in the network, in general, and of the cross-part of speech relations, in particular. In general, semantic and lexical relations are established between words of the same part of speech. To a high extent, derivational relations link words of different parts of speech, as obvious from the table below:

Table 7. Distribution of derivational relation on PoS

	Same PoS - %	**Cross PoS - %**
Prefixes	97	3
Suffixes	15	85
AFFIXES	38	62

6. Conclusions

The more relations there are in a wordnet, the better resource it is, so the better results can be obtained from its use in various applications. Until now there are almost 7000 derivational relations in our wordnet, most of them semantically labelled. A further concern will be how to mark derivational relations for words that are newly introduced in the resource.

The data we presented in this paper about the content of the RoWN seem to be convergent with the linguists' remarks on the distribution of vocabulary related facts, which proves the sensible decisions made when choosing the synsets to be implemented in the RoWN throughout time.

Acknowledgements. The work reported here was supported by the Sectorial Operational Program Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number SOP HRD/89/1.5/S/59758.

References

Avram, M. (1978). *Originea prefixelor româneşti*. Al. Graur, M. Avram (eds.), *Formarea cuvintelor în limba română*, vol II. Bucharest: Editura Academiei, 300-304.

Avram, M. (1989), *Introducere în studiul sufixelor*. Al. Graur, M. Avram (eds.), *Formarea cuvintelor în limba română*, vol III. Bucharest: Editura Academiei.

Bilgin, O. et al. (2004). Morphosemantic relations in and across wordnets: A study based on Turkish. P. Sojka, K. Pala, P. Smrz, C. Fellbaum, P. Vossen (Eds.) *Proceedings of GWC*, Brno.

Marslen-Wilson, W. D. (2007). Morphological Processes in Language Comprehension, in M. Gareth Gaskell, Gerry Altmann. *The Oxford Handbook of Psycholinguistics*. Oxford University Press.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fellbaum, C. et al. (2007). Putting Semantics into WordNet's "Morphosemantic" Links. *Proceedings of the 3rd Language and Technology Conference*.

Kahusk, N. et al. (2010). Enriching Estonian WordNet with Derivations and Semantic Relations. *Proceeding of the 2010 conference on Human Language Technologies – The Baltic Perspective*.

Pala, K., Hlaváčková, D. (2007). Derivational relations in Chech WordNet. *Proceedings of the Workshop on Balto-Slavonic natural Language Processing*, Prague, 75-81.

Pascu, G. (1916). *Sufixele româneşti*. Bucureşti: Editura Academiei Române.

Piasecki, M. et al. (2009). *A Wordnet from the Ground up*. Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej.

Puşcariu, S. (1940). *Limba română*, vol. I *Privire generală*. Bucureşti: Fundaţia pentru Literatură şi Artă „Regele Carol II".

Stoichiţoiu-Ichim, A. (2007). *Vocabularul limbii române actual. Dinamică, influenţe, creativitate*, Bucureşti: Editura BIC ALL.

Tufiş,D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. (2004). The Romanian Wordnet.

Romanian Journal on Information Science and Technology. Special Issue on BalkaNet 7, 105-122.

INSTANTIATING CONCEPTS OF THE ROMANIAN WORDNET

STEFAN DANIEL DUMITRESCU, VERGINICA BARBU MITITELU

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest, Romania

{sdumitrescu, vergi}@racai.ro

Abstract

The authors present a new development for the Romanian WordNet (RoWN): a Java Application Programming Interface (API) with which it is possible to programmatically edit RoWN. Using this free API, the authors extend RoWN by extracting entities from external sources and linking them to RoWN concepts. The API has several other uses described in the paper such as performing queries, adding new relations and classes, easily obtaining RoWN statistics, etc.

Keywords: Romanian WordNet, API, Java, concepts, editing WordNet

1. Introduction

The Romanian WordNet (RoWN) is a reliable semantic network mirroring and extending on Princeton WordNet (PWN) (Fellbaum, 1998). While a mature resource, RoWN is and will continue to be developed and expanded, because it is used in various applications.

Given the implementation approach we followed for the RoWN, some problems arise when using it to language specific tasks. Many English classes have instances represented, which are, usually, specific to the English-speaking world. When creating the RoWN starting from PWN, we added many such instances to our wordnet. But they are simply useless when we deal with Romanian text and we need to know, for example, that Craiova is a town. Thus, in order to improve the quality of our work, we envisaged a method for adding new Romanian instances of some of the existing concepts in the wordnet.

The development of the RoWN started in the BalkaNet project (Tufiş et al., 2004). In one of the final stages of the project there was an interest for adding concepts specific to the Balkan area. Besides concepts, instances have also been introduced. For example, all the Romanian counties and their capitals were introduced in RoWN during BalkaNet and assigned a specific synset ID, with a specific format.

Previous versions of PWN did not distinguish between instances and classes. However, starting from PWN 3.0 instances are marked differently than concepts. As RoWN is aligned to PWN 3.0 this distinction is made in it, too.

The development of RoWN has always been oriented towards serving the interests of our team in various projects in which we have been involved and was done manually, using only an editing tool (WNBuilder, see Tufiș & Barbu, 2004) or editing the underlying xml file directly. In order to be able to at least semi-automate the editing task for enriching RoWN independently of PWN, we present an Application Programming Interface (API) that we developed and used in this article.

2. Related Work

RoWN followed the structure characteristics of PWN as the RoWN developing strategy consisted in transferring the backbone of PWN, namely the organizing semantic relations. However, there are some points in which RoWN diverts from PWN:

- allowance of empty synsets – the concepts existing in English but lacking lexicalization in Romanian are represented as empty synsets, marked with a special tag in the xml file;
- a different sense numbering – we borrowed the sense numbering from our EXPD (Romanian explanatory dictionary in electronic format, covering the official Explanatory Dictionary, DEX) (which keeps track of the evolution of word senses), as opposed to PWN, in which the sense numbers are assigned according to their frequency in a semantically disambiguated corpus;
- random order of literals in a synset – in PWN literals are ordered in the synsets in reverse order of their frequency in a semantically disambiguated corpus; for Romanian we lack such a corpus, so we randomly wrote the literals in the synsets;
- lexical relations that were transferred from PWN were marked as semantic relations, so holding between synsets, not between literals and renamed by adding *near* in front of the name of relation (e.g., near-antonymy). This is due to the lack of correspondence between literals in the two wordnets; the focus is on implementing meanings, not on giving equivalents for literals;
- it was enriched with SUMO/MILO[15] and DOMAINS[16] labels and with SentiWordNet[17] scores.

Currently, a large number of APIs exist that provide a programmatic interface to PWN, written in various languages. For Java, there are more than 10 such interfaces

[15] http://www.ontologyportal.org/
[16] http://wndomains.fbk.eu/
[17] http://sentiwordnet.isti.cnr.it/

(http://wordnet.princeton.edu/wordnet/related-projects/#Java), each having different strengths and weaknesses: JWNL/extJWNL[18], RitaWN[19], JAWS[20], WNJN[21], JWNL[22], JWI[23], etc. Some of them have the PWN already included; others need the dictionary files as input. JWI is an interesting and very versatile interface, but is rather complex, and thus rather difficult to extend. RitaWN on the other hand is very simple to use, but lacks some of the functionality we would like to have, such as the ability to perform breadth-first searches in the noun hypernym tree. However, the main issue with most of the current interfaces is that they are focused on PWN, both in the input/output functionalities (e.g., PWN is usually either embedded in the interface or is distributed as "dict" files, whereas RoWN is distributed as a single XML file) and in the data structures within the interface, making it difficult to programmatically adapt. (RoWN synset data differ from PWN synset data: with a couple of exceptions, RoWN contains all PWN data fields and some more, as described above).

We need to find a very simple, basic API that is easy to extend, and in which we can code some very specialized functionalities very quickly. Given that our survey of existing interfaces has yielded no clear winner in terms of simplicity of usage and extension, we considered that writing a new API from the start was a better choice.
The project GeoWordNet (http://geowordnet.semanticmatching.org/) integrated PWN with GeoNames (http://www.geonames.org/). The latter is a geographical database covering place names from the entire world. Romanian entities are also included both in GeoNames and in GeoWordNet. GeoWordNet was used in geographical information retrieval (Buscaldi & Rosso, 2008) tasks. However, while GeoWordNet has added the vast majority of geonames.org entities as instances to PWN thus creating a very large repository (the added instances heavily outnumbering the concepts), we intend to create a Romanian WordNet core that contains the most important entities, such as major cities, mountains, rivers, etc. Moreover, searching through the Romanian entities in GeoNames, we considered that we need to have them more organized in RoWN: for example, mountains grouped in larger units (ranges), the peaks associated with the mountains, etc. If needed, we could create an extension having only geonames.org entities (and entities from other sources as well) that could reside in a different file than RoWN itself, as, for example, a word-sense disambiguation application usually works only with concepts and not entities.

3. Developing RoWordNetLib

RoWordNetLib is thus an API interface written in Java (due to its widespread usage). It is compiled in a "jar" format, exposing classes that read, process and write to and from a WordNet source. We have the following classes, divided into packages by their function:

[18] http://extjwnl.sourceforge.net/
[19] http://rednoise.org/rita/wordnet/documentation/index.htm
[20] http://engr.smu.edu/%7Etspell/
[21] http://wnjn.sourceforge.net/
[22] http://sourceforge.net/projects/jwordnet
[23] http://projects.csail.mit.edu/jwi

Table 1: Main RoWordNetLib classes

Package	**Class**	**Function**
data	Synset	Data structure of a Synset
	Literal	Data structure of a Literal (substructure of a Synset)
	Relation	Data structure of a Relation (substructure of a Synset)
	RoWordNet	Main data structure containing all synsets
io	IO	Generic Input-Output functions
	XMLRead	Class implementing read functions from an XML source
	XMLWrite	Class implementing write functions to an XML destination
	SQLRead	Class implementing read functions from an SQL source
	SQLWrite	Class implementing write functions to an SQL destination
op	Operation	RoWordNet operations such as intersection, union, etc.

We will shortly describe each data structure class. The central element of a wordnet is the Synset structure. It contains 10 fields: id (a unique identifier of the synset), pos (the part of speech of the literals in the synset), definition (of the meaning of the literals in the synset), usage (examples of the occurrences of literals), stamp (the name of the developer), literals (at least one, but no literal is also possible), relations (at least one, but no relation is also possible especially with adverbs), domain (from mappings to DOMAINS 3.2), sumo (from mappings to SUMO), sentiwn (from SENTIWORDNET). The fields id, definition, stamp, domain are of type String, sumo is of type 2 Strings, sentiwn of type 3 Strings, while usage is a String Array and literals a Relation Array.

The Synset class provides getters and setters for all its data fields. It also has a custom hash code function for quick indexing in hash maps and a custom equals method. A Synset is enforced to have an id, a definition, and a part-of-speech.

The synset contains one or more literals, which are represented as Literal structures in our API.

A Literal class contains a word (the literal itself) and its sense number. There are custom hash and equal functions, as for example, one can search for the literal “capitală” with sense “1.1” or for all literals “bancă” without specifying a particular sense, yielding the literal “bancă” with different senses.

The Synset also contains a Relation data structure, encoding the relation of this synset to another.

The class RoWordNet is the main operating class of RoWordNetLib and represents a complete wordnet. It contains an array of synsets (for ordered synset access) and a hash map of the same synsets, indexed by their IDs (for O(1) *contains* and *get* operations). It also contains four smaller hash maps, each similarly indexed the synset by their IDs but grouped by each of the four distinct parts of speech. Table 2 presents a few of the functions RoWN offers.

Table 2: Examples of RoWN functions

Method name	**Function**
getNewID	Obtains a new, unused ID. As IDs can have different formats, the getNewID function is aware of the prefix and suffixes of the requested IDs.
addSynset	Adds a new synset to the collection.
getSynset	Returns the synset that has a specific ID.
getIdFromLiteral	Obtains a set of synsets that contain a specific literal. The search can also find all synsets that contain a literal with a specific sense.
getRelatedSynsetIDs	Returns a set of synsets that are connected to a parent synset given a relation (or "*" to specify that any relation is acceptable).
getLeastCommonAncestor	The function returns the closest common ancestor of two synsets. The function is used to perform distance calculations and find connections between synsets in the RoWN noun hierarchy.

The *io* package handles data input-output. At present we have the XML reader/writer fully working and the SQL reader/writer in an alpha version.

The XML reader/writer is implemented as two static classes, very easy to use:

```
RoWordNet rown = new RoWordNet(XMLRead.read("d:\\RoWordNet_3.0\\wn.xml"));
```

The code above will create a new RoWordNet object named rown that will contain all synsets stored in the wn.xml file. The XMLRead.read function takes only the path of the Romanian WordNet file and provides a fully constructed RoWordNet object. The XML reader is implemented as a SAX parser and as such is fully extensible with new, unknown tags.

The XML writer is implemented as an XMLStreamWriter. The function takes four parameters: the RoWordNet object, the output file path, whether to overwrite any existing file and whether to write the XML as a compressed format or not.

```
XMLWrite.write(rown,"d:\\RoWordNet_3.0\\out.xml",true, false);
```

The above example will write all synsets in the *rown* object to file out.xml, overwriting it if the file already exists, using a non-compressed format ("pretty" formatting).

The SQL reader/writer is a working alpha version, capable of reading and writing in MySQL databases. We will extend the SQL reader/writer capabilities with a PostgreSQL and MSSQL interface. A SQLite interface for local database access is also in the future development plan.

Finally, the *op* package contains the Operations class where we have implemented several functions such as Intersect, Difference, Union, etc. These functions view RoWN as a set of synsets, and given two such sets operations such as union become possible. With these functions there are two ways of editing the RoWN: 1. standard editing on a single RoWordNet object (adding, modifying and removing synsets directly from the object) and 2. using set functions between two RoWordNet objects to obtain a third (e.g., creating a new, empty RoWordNet, adding new synsets and then performing a

Union operation between this new object and an existing object to obtain a single RoWN that automatically does not contain duplicates).

4. Using RoWordNetLib

The goal of creating the RoWN API is to enable easy usage and editing of the RoWN. As such, we present our experiments of extending and improving RoWN using our API.

4.1. Instantiating concepts with Geonames.org entities

Geonames.org is a very large collection of geographical data spanning the entire world. The data is organized as a shallow, two-step hierarchy of concepts that are instantiated by entities. At the time of this writing, geonames.org states that "it contains over 10 million geographical names and consists of over 8 million unique features whereof 2.8 million populated places and 5.5 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes" (http://www.geonames.org/about.html).

Roughly, when adding new synsets to a wordnet one has to take care the following requirements are obeyed:

1. The synset has a unique ID.
2. The new synset has a gloss.
3. All its possible relations are specified, thus preventing it being a dangling node in the network.

None of these is a trivial task. The third point was solved quite easily because we started from a list of concepts for which we wanted to add instances. As such, we obtained the geonames.org dump file corresponding to Romanian entities. The "RO.zip" file was downloaded from http://download.geonames.org/export/dump/ which contains the tab-delimited 3MB RO.txt. We created a Java class named GeonamesWrapper that functions as a dictionary containing entities extracted from any geonames.org dump file. The GeonamesWrapper object based on the RO.txt file yielded an indexed dictionary of 25461 entities. Each entity is composed of several data fields: a unique ID, a UTF-8 name, an ASCII name, geographical coordinates, county name (optional), a feature class and a feature code, etc. We keep all these features in the GeonamesWrapper object.
Having available the list of Romanian entities from Geonames.org, we then selected which classes of entities we wanted to add to the WordNet. For each entity class we had to obtain a parent class from RoWN. The added entities will thus be instantiations of the parent class. For example, each entity having feature class AIRP (airport) will be an instantiation of the WordNet synset:

```
Synset: id=ENG30-02692232-n, pos=Noun, definition=Teren special amenajat pentru
decolarea, aterizarea și staționarea avioanelor, cuprinzând și instalațiile, asistența
tehnică etc. necesare activității de zbor, domain=town_planning
   Literal [literal=aerodrom, sense=1]
   Literal [literal=aeroport, sense=1.x]
   Relation [ENG30-02692232-n hypernym ENG30-02687992-n]
   Relation [ENG30-02692232-n part_meronym ENG30-02687821-n]
   Relation [ENG30-02692232-n part_meronym ENG30-02693246-n]
   Relation [ENG30-02692232-n part_meronym ENG30-03098959-n]
```

Table 3 describes the entity classes chosen and the RoWN concepts they instantiate (their parents).

Table 3: Geonames.org entity classes chosen to instantiate in RoWN

Geonames.org entity class	**Parent ID**	**Parent synset**	**Count**
PPLA	ENG30-08524735-n	oraş, metropolă, centru_urban	35
ADMF	ENG30-03449564-n	clădire_administrativă	2
AIRP	ENG30-02692232-n	aerodrom, aeroport	17
CAVE	ENG30-09238926-n	cavernă, grotă, peşteră	8
DLTA	ENG30-09264803-n	Deltă	1
ISL	ENG30-09316454-n	Insula	36
MSTY	ENG30-03781244-n	mănăstire	15
PASS	ENG30-09386842-n	pas, strâmtoare, trecătoare	20
PLAT	ENG30-09453008-n	platou, podiş	9
PLN	ENG30-09393605-n	şes, câmp, câmpie	12
		TOTAL	155

The total number of synsets added to RoWN with the help of this API is 155 geographical instances.

Every added entity will be linked to its parent synset through the *instance_hypernym* relation. Also, the reverse relation will link the parent to the instantiation using the *instance_hyponym* relation. As an example, after adding all geonames.org AIRP entities, the {aerodrome, aeroport} synset will contain 17 *instance_hyponym* relations pointing to each of the 17 added entities, while from each of these instances an *instance_hypernym* relation points to the {aerodrome, aeroport} synset. So, for the 155 instances 310 relations have been added.

The 155 instances belong to 10 different classes of Geonames features. For each class we have manually written a definition that will become the gloss for every instance of that class. We can also add a placeholder (*) that will be automatically filled at run-time. For example, for the CAVE class, all instances received the gloss: "Peşteră în judeţul *", where * is the automatically retrieved county name that most entities in Geonames possess. So, for instance "Peştera_Polovragi" received the following gloss: "Peşteră în judeţul Gorj".

All added instances received an incremental unique ID obtained using the *getNewID* function of the API. Using the prefix RI which stands for "Romanian Instance", we have obtained 155 new ids: instance Peştera_Polovragi received the RI-00000059-n unique id. The count started from 0 because there were no previous synsets with the RI prefix. The chosen prefix differentiates the new synsets from the previously existing ones: those implemented starting from PWN have the ILI prefix, while those specific to the Balkan area (developed during BalkaNet) have the BILI prefix.

We present the RoWN entry for the newly created "Peştera_Polovragi" instance.

```
<SYNSET>
  <ID>RI-00000059-n</ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>Peştera_Polovragi<SENSE>x</SENSE></LITERAL>
  </SYNONYM>
  <DEF> Peşteră în judeţul Gorj</DEF>
  <ILR> ENG30-09238926-n<TYPE>instance_hypernym</TYPE></ILR>
</SYNSET>
```

The name of an administrative building is often used to refer to the institution whose headquarters it houses or to the head of that institution: "Guvernul a anunțat anterior participarea premierului Ponta la CSAT joi, de la ora 14.00, însă ulterior *Palatul_Victoria* și-a retras anunțul […]" (http://cursdeguvernare.ro/ponta-a-anuntat-sedinta-csat-inaintea-lui-antonescu-privind-atentatul-din-bulgaria.html, accessed on 26th February, 2013). Thus, there appears the need to add further metonymic synsets in such cases, given that there already are such metonymies in wordnet: see *White_House:1* {the chief executive department of the United States government} and *White_House:2* {the government building that serves as the residence and office of the President of the United States}.

Counties and their capitals were already in RoWN due to a BalkaNet initiative of adding some area specific concepts to the wordnets in the project. What is important is that from these toponyms we can create the name of the inhabitants of the places or of the persons born in those places. This can be an automatic task: given the list of toponyms and a list of specific suffixes (i.e. suffixes that create names of inhabitants or of people born in a place), we can automatically attach the suffixes to the toponyms. The results are automatically validated on the web, establishing a threshold for the number of occurrences of a word above which we validate the form. Whenever manual intervention is necessary we appeal to it. The same approach to validate automatically derived words in Romanian was adopted by Petic (2012). These synsets designating names of people can be automatically attached a gloss, too. This remains future work.

4.2. Using the API to improve the current RoWN

Besides extending RoWN with new instantiations, while developing RoWordNetLib we have identified some errors in the current version of the wordnet and gained some interesting insights. We further present only a few of them:

One technical error identified was the use of xml-reserved characters in the definitions and in the sumo tag "TYPE". For example, the tag was written as "<TYPE>>=</TYPE>". The character ">" is not allowed in the content between tags, as it usually means the end of a tag. Instead, the character should be represented as ">". Using the SAX XML parser we were able to identify these minor errors and correct them.

Another issue we have discovered during the development of RoWordNetLib was that some of the synsets did not have a DEFINITION tag. 53 of the almost 60000 synsets were missing a definition. The synsets have been corrected by adding the missing definitions using RoWordNetLib programmatically.

Having all synsets indexed means that we can perform different counts to obtain some interesting insights and statistics. For example, before adding the geonames.org entities, we had 41063 noun synsets, 10397 verb synsets, 3066 adverb and 4822 adjective synsets. We can also obtain a relation frequency table. For example, we have 3889 instance_hypernym relations, meaning there are 3889 instantiations of wordnet concepts. While RoWN contains 48316 hypenym relations, it also contains only 3 near_participle and only one near_domain_topic relation. In total we have 37 different relations. Other statistics can be obtained, such as the synsets with most instantiations, the synsets with most hyponyms (or any other relation), etc.

5. Conclusions

This article presented another step in the development of the Romanian WordNet. We have written a Java API specifically to allow easy RoWN access and editing. Also, the API itself was made to be as easy to use as possible, and also easy to extend with new functionalities as RoWN itself evolves. The API offers basic access to RoWN, whether in the original XML format or stored as an SQL database, basic operations to manipulate synsets, set operations (like union, difference, intersection) applied to RoWN objects, as well as more complex operations like breadth first searching on the noun graph.

The development of the API has allowed us to identify and correct a few small inadvertencies in the wordnet, and also to obtain insights in its structure and entity distribution using the embedded statistics functions.

The main achievement using the API was to actually expand RoWN by semi-automatically instantiating concepts with entities extracted from an external source (GeoNames – geonames.org). We focused on geographical entities belonging to the classes: administrative buildings, airports, caves, delta, islands, monasteries, passes, hills, and plains. It is worth mentioning that geography was chosen because of the big amount of such information available on the Internet. However, this is only the starting point in this experiment and other domains will also be covered. Some of the enumerated classes are of utmost importance for the domain, while others are relevant for the administrative domain and occur frequently in news of national concern, in which they offer the local coordinates of the event. All these are different categories in GeoNames, thus we manually specified for each of them the correspondent in the wordnet.

For future work, we intend to make the API freely available, on a public platform such as sourceforge or github. Currently, while perfectly usable, the code needs optimizing and lacks commenting on each of its functions. Also, usage examples need to be given to help users understand the structure of RoWordNetLib an how to use the code.

Also as a future work we intend to use the developed API to create an extension of RoWN containing most of the Romanian entities in geonames.org, as well as entities from other sources. This extension will be optional when loading RoWN with RoWordNetLib.

Mountains and rivers are important geographical entities. In the case of the former, for example, we consider it is not enough to simply include them as instances of mountain. In geography they talk about groups, about ranges, mounts and peaks. We want to capture the same information in our network. That is why we decided to part from

GeoNames in such cases and use other available sources of such information to start from when adding them to RoWN. This will be future work.

References

Buscaldi D., Rosso P. (2008). Using GeoWordNet for Geographical Information Retrieval. *Revised Selected Papers CLEF-2008.* Springer-Verlag, LNCS(5706), 863-866

DEX - Coteanu, I., L. Seche, M. Seche eds. (1996). *Dicţionar explicativ al limbii române* (DEX). Bucureşti: Editura Univers Enciclopedic.

Fellbaum, C. (ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Petic, M. (2011). *Automatizarea procesului de creare a resurselor lingvistice computaţionale*, PhD Thesis. Institutul de Matematică şi Informatică al AŞM.

Tufiş, D., Barbu, E. (2004). A Methodology and Associated Tools for Building Interlingual Wordnets. *Proceedings of the 5th LREC Conference*, 1067-1070.

Tufiș, D., Cristea, D., Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. *Special Issue on BalkaNet of the Romanian Journal of Information Science and Technology*, 7:(1-2), 9-43.

STEPS TO A NEW DTD AND SCD-BASED DICTIONARY ENTRY PARSER. OPTIMIZING RECURSIVENESS IN SENSE DEPENDENCY HYPERGRAPHS

NECULAI CURTEANU[1], ALEX MORUZ[1,2], SVETLANA COJOCARU[3]

[1]*Institute of Computer Science, Romanian Academy, Iaşi Branch, România*
[2]*Faculty of Computer Science, "Al.I. Cuza" University, Iaşi, România,*
[3]*Institute of Mathematics and Computer Science, Chişinău, Republic of Moldova*

ncurteanu@yahoo.com, alex.moruz@gmail.com, svetlana.cojocaru@math.md

Abstract

In previous papers we developed the dictionary-entry text version for the *parsing method* of SCD (Segmentation-Cohesion-Dependency) *configurations*, which was applied to *six* largest (Romanian, French, German, and Russian) thesaurus-dictionaries, with outstanding efficiency and portability results. In SCD method, the Dependency Hypergraph (DH) describes, for a dictionary, the specific pre-established dependency relations between the sense marker classes of that dictionary. The DH of a dictionary is akin to its *fingerprint*. The present paper solves the following ***problem***: transforming the sense DHs with non-embedded cycles and / or troublesome (*e.g.* disconnected) hypernodes, into DHs having *only* structurally embedded recursive cycles and linearly connected hypernodes. The DH optimization is based on a *total ordering* of *literal enumeration*(*s*) within sense marker classes, obtaining linearly embedded cycles for *all* DHs that represent an SCD parsing level. This solution opens the effective possibility to construct the *least upper bound* (LUB) of several optimized DHs, the associated *parametrized grammars* of such LUB DHs yielding the formal descriptions of a sound DTD and a general SCD-based parser for very large dictionaries.

Keywords: sense marker-depending renaming of the literal enumeration, total ordering of sense levels, parametrized grammar

1. Introduction

In (Curteanu et al., 2008, 2010, 2012) we applied the method of *Segmentation-Cohesion-Dependency* (SCD) *configurations* to model and parse the following *six*, sensibly different, Romanian, French, German, and Russian largest thesaurus-dictionaries: **DLR** (The Romanian Thesaurus – new format), **DAR** (The Romanian Thesaurus – old format), **TLF** (Le Trésor de la Langue Française), **DWB** (Deutsches Wörterbuch – GRIMM), **GWB** (Göthe-Wörterbuch), and **DMLRL** (Dictionary of Modern Literary Russian Language). Parsing a dictionary entry means to identify its *lexicographic segments* (the first SCD configuration – SCD-Config1), to extract its *sense tree* (SCD-Config2), and to parse the *atomic sense definitions* (on SCD-Config3).

When applied to dictionary entry parsing, the method of SCD configurations merges the following sequence of (at least) *three* specific *configurations*, *i.e.* lexical-semantics sense levels: **(a)** The *first one*, abbreviated hereafter SCD-*config*1, performs the segmentation and dependencies for the *lexicographic segments* (Hauser & Storrer,1993), (Erjavec et el., 2001) of each dictionary entry. **(b)** Stepping down into the lexicographic segments of a thesaurus-dictionary entry, the *second* SCD *configuration* (SCD-*config*2) usually parses the lexicographic segment of *sense description*, extracting its *sense tree* structure (Curteanu et al., 2008, 2010). Actually, the SCD-*config*2 parses the entry sense definitions of larger lexical-semantics granularity in the sense description segment: primary, secondary, and literal / numeral enumeration senses. **(c)** The *third* SCD *configuration* (SCD-*config*3) continues to refine the sense definitions of SCD-*config*2, parsing each node in the generated sense-tree for obtaining the atomic definitions / senses, *i.e.* finest-grained meanings of the dictionary entry.

In (Curteanu et. al., 2010, 2012), we gave a structural analysis of the *dictionary entry parsing* (DEP) process, comparing the classical approaches to DEP with the method of SCD configurations, applied to dictionary entry text. In the classical (called also standard) DEP, relying largely on *lexicographic formal grammars*, *e.g.* (Neff & Boguraev, 1989), (Tufiș et al., 1999), the *LexParse system* in (Hauser & Storrer, 1993) and (Lemnitzer & Kunze, 2005), the main problem is that the sense tree construction of each entry is recursively embedded and mixed within the definition parsing procedures. The formal *proof of optimality* for the parsing method of SCD configurations compared to the *standard* DEP (Curteanu et al., 2012) shows that the latter DEP strategy contains *three embedded cycles*, corresponding to the parsing processes of *lexicographic segment recognition*, *sense tree extraction* (for entry senses defined by explicit marker classes), and *atomic definition parsing*. The main power and novelty of the SCD-based *parsing method* is that it succeeds to *separate* and *run sequentially*, on *independent* parsing levels (*viz.* SCD-configurations), the three above mentioned parsing cycles / processes.

Since the process of *sense tree construction* in the method of SCD configurations could be made *completely detachable* from that of *parsing the atomic sense definitions*, the whole DEP process with SCD-based method is much more *efficient* and *robust*, actually *optimal* (Curteanu et al., 2010). There are (at least) two *distinct* and *specific features* of the SCD parsing method: **(a)** the *breadth-first search* of *all* the sense markers of an entry, and **(b)** working directly on the sense marker sequence(s), and *only* on them (for the SCD configurations of larger semantic granularity senses), to compute the *sense tree* of the entry. These properties of the new parsing method of SCD configurations have been effectively tested by in-depth parsing experiments on *six* largest Romanian, French, German, and Russian thesaurus-dictionaries (Curteanu et al., 2010, 2012). Remarkably, the proposed *parsing method* is a completely *formal grammar-free* approach, with the parsing program for a new thesaurus readily adaptable in weeks-time (depending on its lexicographic modeling task, specific to each dictionary), thus providing an outstanding *portability* of the parsing program, both on linguistic and computational grounds.

As a major computational component of the SCD parsing method, the Dependency Hypergraph (DH) of a thesaurus-dictionary embodies (by the SCD lexicographic modeling) the pre-established hierarchy between the sense marker classes of *that* dictionary, being actually its true semantic "fingerprint". The study of DHs for various

thesaurus-dictionaries has a special importance for both the lexicography and parsing technology: **(a)** DHs have to reflect the regular (and irregular) representations of the sense dependencies. **(b)** The comparison between various specific DHs is the best opportunity to simplify, regularize, and standardize (b1) the dictionary sense marker classes, (b2) the rules encoding the sense definition markers, and (b3) the dependencies that can be established between the sense / subsense marker classes of the dictionary. The careful analysis of DHs for various thesaurus-dictionaries, based on the parsing method of SCD-configurations, have important consequences within the cross-linguistic *lexicography-lexicology research*: to establish better, standard and optimal design rules for the dictionary sense markers, and entailing optimized DHs of sense dependencies for the most complex thesaurus-dictionaries, either existing or designed ones.

We strongly emphasize that this paper does not deal with effective parsing experiments of one or several dictionaries but with the formal representation and optimization of DHs, as computational objects in the parsing method of SCD configurations, with important consequences on the design of a new, procedural DTD and of a formal, general SCD-based parser for very large thesaurus-dictionaries.

2. The Problem of Non-Embedded Call Cycles in Dependency Hypergraphs

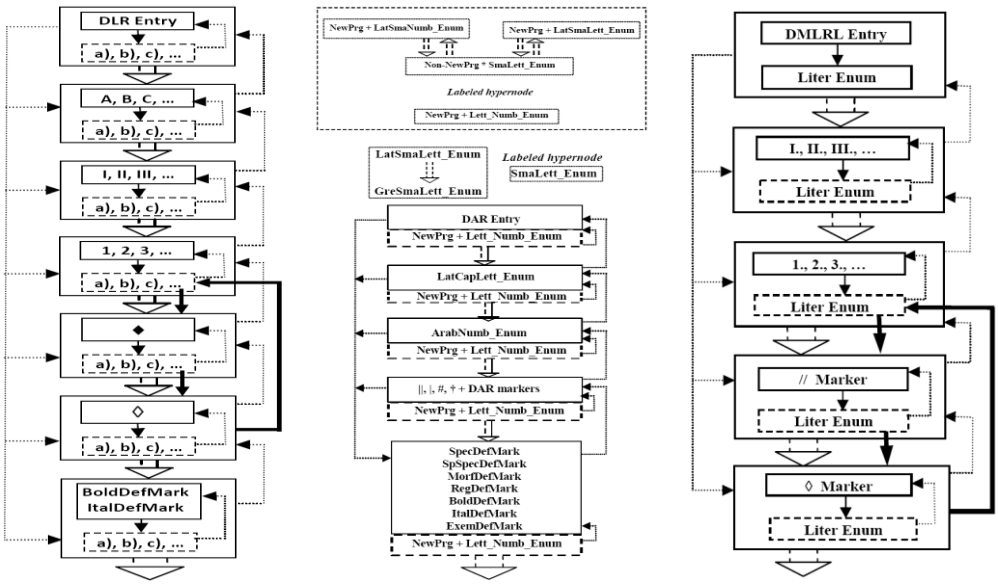


Figure 1: DHs for **DLR**, **DAR**, and **DMLRL** (Curteanu et al., 2012)

The project of a new, *procedural* DTD for dictionaries, based on the formalization of the SCD parsing method, is exposed in (Curteanu & Moruz, 2012b), based on *parametrized* (*par-*)*grammars* encoding the sense DHs that correspond to the SCD parsing levels (SCD configurations) in a dictionary entry. Two par-grammars for **DLR** are enclosed, as a typical sample from a larger package of combined grammars for the six, above mentioned, dictionaries. This package should be constructed as the "*least upper bound*" (LUB) of all the par-grammars, written for the parsed dictionaries on the SCD configurations (parsing levels). Such a package of par-grammars should represent the DTD description of a SCD-based formally defined *parser* for large dictionary

entries, and thoroughly extends the current DTD in the standard (XCES TEI P5, 2007). The *soundness analysis* of sense structure definitions in thesaurus-dictionaries, achieved in (Curteanu & Moruz, 2012a), revealed the special problems raised by the recursive calls between secondary sense markers (*i.e.* filled and empty diamonds ♦ and ◊, or their sense marker equivalents) and the literal enumerations (*i.e.* **a)**, **b)**, **c)**, …) in certain special entries, presented in Section 3.

For understanding the problem whose solution we propose in this paper, a final element has to be explained. The **main problem** we are dealing with in this paper is the following: **how to transform** the DHs of the kind in Fig. 1 (with non-embedded recursive call cycles and non-connected hypernodes) into **linearly recursive** DHs, with **completely embedded call cycles**, such as the DHs in Fig. 3. The optimized DHs displayed in Fig. 3 are suitable to LUB-computing (by unification-matching algorithms), the *LUB*(DH_i) being that DH in Fig. 4, whose par-grammar is devised in Section 6 as the new, procedural DTD representing the primary and secondary sense marker classes on the SCD-config2 parsing level.

3. Atypical Entries Generating DHs with Non-Embedded Cycles

The in-depth analysis in (Curteanu & Moruz, 2012a) discussed the cyclicity calls between *secondary sense markers* and *literal enumerations*, and pointed out examples of such atypical entries in **DMLRL**, **DLR**, and **DAR**, where the recursive calls for literal enumeration are mixed with secondary sense markers (filled and empty diamonds, or their correspondents). These entries, *viz.* "**LUMÍNĂ**" in **DLR**, "**CAL**" in **DAR**, and "**БЫ**" in **DMLRL** (Curteanu & Moruz, 2012a) (see the excerpts below).

(Ex. 3.1) The entry "**LUMÍNĂ**" in **DLR** Romanian thesaurus-dictionary (excerpt):
LUMÍNĂ s.f. **A.** (Predomină sensul concret de radiaţie; în opoziţie cu î n t u n e r i c)
I. (Adesea cu determinări calificative) Radiaţie care face corpurile vizibile.
1. (Ca atribut al universului, al naturii ambiante; componentă a lumii înconjurătoare) *Lăudaţi-l toate stealele şi … … … gonească Cât va fi câmp de gonit Şi lumină de zărit".* ALECSANDRI, O. I, 8.
a) (Ca radiaţie solară, element al peisajului diurn) *Voi întoarce lumira soarelui de cătră voi, de va fi întunrearecu* (a. 1600). CUV. D. BĂTR. II, 49/9. *Lumina soarelui face dzua.* PRAV. 141. … … … … … … *Deopotrivă se găseşte-n toate Amestecată umbră şi lumină.* ISANOS, V. 281. ♦ L o c. a d j. **De lumină** = **a)** luminos, sclipitor; s p e c. (despre ochi) strălucitor. *Deunăzi ... mă simţii cufundat ca într-un nor întunecos ... Ancuţo! tu ai prefăcut acel nor … …* ODOBESCU, S. I, 143. … … …*Ochi de lumină avea fiul lui Ieronim, privirea lui în noapte fulgera.* ROMÂNIA LITERARĂ, 1970, nr. 93, 17/3; **b)** (despre un spaţiu, un loc) în care pătrunde lumina **(A I 1),** plin de lumină *Acest loc ... era pe atunci, în 1650, un ochi de lumină în mijlocul marelui codru al Căpoteştilor.* IORGA, C. I. II, 5; **c)** (despre plante) care trăieşte la lumină **(A I 1).** *După o fază de 2-3 ani cu floră de buruieni de lumină, urmează faza de fâneaţă cu ierburi cu rizomi.* CHIRIŢĂ, P. 71. ♦ L o c. a d v. **Pe** (sau, rar, **la) lumină** = în timpul zilei **(I 2),** de … … … … … … … ARHIVA R. I, 87/20. *A înviat din morţi ..., Lumina ducându-o Celor din morminte!* EMINESCU, O. IV, 359. *Zâmbetul sfânt al martirului care-ntrevede ... lumina vieţii eterne.* CARAGIALE, O. II, 64. … …… …… … … … … … … … …… … …… … …… … …… … … …
b) (Ca radiaţie reflectată de lună; element al peisajului nocturn) *Luna, ... fire are lumina ce iase den ea să turbure udăturile trupului.* CORESI, EV. 81. … … … … … … … … … … … … … … … … …

(Ex. 3.2) The entry "**CAL**" in **DAR** Romanian thesaurus-dictionary (excerpt):

NewPrg **CAL** s.m. *Cheval.*
NewPrg 1^0. Numele generic al speţei cavaline; s p e c. individ masculin...
…
NewPrg *Adecă amù cailoru zăbalele în gură lă*…
… {a large block of definitions and *DefExems* of the entry **CAL**}

NewPrg În compoziţii:
NewPrg a.) (Entom.) **Cal-de-apă** = o specie a c a l u l u i - d r a c u l u i, numită…
…
NewPrg **Calul-dracului** = a.) insectă cu corpul lung… | (De aici) Babă rea… ; –b.) = **cal-de-apă**…
…
NewPrg **Calul-popii** = a.) c a l u l-d r a c u l u i…; –b.) = cal-de-apă… Insectă lungă şi cu aripile pătate…
NewPrg **Cal-turtit** = c a l u l-d r a c u l u i…
NewPrg b.) (Zool.; la românii din A.-U.) **Cal-de apă** s. (după germ. Nilpferd) **–cal-de-Nil** = h i p o p o t a m LB., BARCIANU …
…
NewPrg 2^0. P. a n a l. (Mor.) *Caii* cu spetezele ţin coşul şi alcătuesc…
…

(Ex. 3.3) The entry "**БЫ**" in **DMLRL** Russian thesaurus-dictionary (excerpt):

… … … … … …
2. В придаточной части сложного предложения обозначает действие, обусловливающее собой то, о чем сообщается в главной части. *Когда б разбойника облавою не взяли, То многие еще бы пострадали.* Михалк. Бешен, пес
3. Обозначает различные оттенки желаемости действия; **а)** Собственно желаемость. *Учился бы сын.* … … … ◊ Если бы, когда бы, хоть бы и т. п. О, *если бы когда-нибудь Сбылась поэта сновиденья!* Пушк. Посл. к Юдину. [Николка:] *Хоть бы дивизион наш был скорее готов.* Булгаков, Дни Турб. ◊ С неопр. ф. глаг … … … … *Вот бы поймать!* А. Остр. Не было ни гроша … … … … *Искупаться бы!* Купр. Бел. пудель. // Употр. для выражения опасения по поводу какого-л. нежелательного действия (с отрицанием). *Не заболел бы он.* ◊ С неопр. ф. глаг., имеющей перед собой отрицание. **—** *Гляди, — говорю, — бабочка, не кусать бы тебе локтя!* … … … Леск. Воительница. ◊ Только бы (б) не. **—** *По мне жена как хочешь одевайся, .. только б не каждый месяц заказывала себе новые платья, а прежние бросала новешенькие.* Пушк. Арап Петра Вел. … … … … **б)** Пожелание. *Условие я бы предпочел не подписывать.* Л. Толст. Письмо А. Ф. Марксу, 27 марта 1899. ◊ С неопр. ф. глаг. *Поохотиться бы по-настоящему, на коня бы денег добыть, — мечтал старик.* Г. Марков, Строговы. ◊ В сочетании с предикативными наречиями со знач. долженствования, необходимости, возможности. … … *Вслед ему косились плешивые повытчики: «Потише бы надо, ... … … … …* **в)** Желание-просьба, совет или предложение (обычно при мест. 2л.). [Марина:] *И чего засуетился? Сидел бы:* Чех. Дядя Ваня. …
— Ты бы, Сережа, все-таки поговорил ... ...: Пришв. Кащ. цепь. **г)** Желаемость целесообразного … … … ◊ С неопр. ф, глаг. *Вам бы вступиться за Павла-то! ... ...* М. Горький, Мать. … … … …

4. Total-Ordering for Sense Marker Classes in Dependency Hypergraphs

For building linearized recursive DHs, *i.e.* DHs without non-embedded cycles between the sense marker classes (which is the problem enounced in Section 2), we propose the following informal description for the solution to this problem (see also Fig. 4):

(a) To the marker classes of ***primary senses*** there are assigned increasing scores accordingly to their decreasing priority, actually to their decreasing semantic granularity of each sense meaning. For instance, to the **four** primary senses in the **DLR** thesaurus-dictionary (**root senses** and the sense marker classes **A.**, **B.**, **C.**, …; **I.**, **II.**, **III.**, **IV.**,…, and **1.**, **2.**, **3.**, …) one can assign as priority scores the numbers 2, 4, 6, and 8.

(b) The first level of ***literal enumeration***, *i.e. p***a).**, *p***b).**, *p***c).**, …assigned to *all the primary senses* in **DLR**, receives the score $p = 9$, thus greater than *all* the scores allocated to the primary senses in **DLR**. Whether in a dictionary, the first level of literal enumeration is refined by further literal enumerations (*e.g.* in German **DWB**), encoded by **2a).**, **2b).**, **2c).**, … and **3a).**, **3b).**, **3c).**, … , these two additional levels of literal enumerations receive the priority scores of 10 and 11, respectively (see Fig. 4).

(c) The ***secondary senses*** and their markers are treated as *a second package of senses*, playing a distinct role compared to the *package of primary senses*, since secondary senses are considered to be endowed with a (substantial) smaller lexical-semantic granularity than the primary ones. This is why we assign to them **special priority scores**, correlated with the literal enumerations that are used within their levels. Namely, an example of allocated priorities is the following: the filled and empty diamonds ♦ and ◊ may receive the scores 12 and 14, respectively. The literal enumerations associated with secondary senses, let they be denoted by **♦a).**, **♦b).**, **♦c).**, … and **◊a).**, **◊b).**, **◊c).**, …, may receive as priority scores the numerals 13 and 15, respectively. If necessary, several layers of literal enumerations can be added to the basic level, as shown at the point **(b)** above, together with the corresponding codification of the additional enumeration refinements. These score allocations allow the sense recursive calls in the sample entries displayed in (Ex. 3.1-3.3) to be represented by the DHs in Fig. 3. Thus the proposed solution supports linearized recursive DHs, eliminating the non-embedded call cycles in the DHs. This fact allows for DH representation with par-grammars and tractable LUB-computing of par-grammars as DTD for SCD-config2 parsing level (of primary and secondary senses).

(d) Between the secondary sense markers ♦ and ◊ in **DLR** there exists a dependency established within the first DH of Fig. 3: senses marked with ♦ are more general than those marked by ◊. The same is true for the corresponding sense markers // and ◊ in **DMLRL**. It is not established (until now) a clear dependency relation between the semantic granularities of the senses delimited by the markers || and | in **DAR** thesaurus-dictionary (Curteanu et al., 2012: Fig. 4, p.43). To these secondary sense markers one may assign equal priority scores, with equal scores attributed to their literal enumeration refinements, under them being situated *all* the *atomic sense definitions* in **DAR**. Thus one can't establish dependency relations between, *e.g.*, a literal in the enumeration refining the sense marker "||", and the sense marker "|"; the reverse, *i.e.* changing each other the "||" and "|" markers in the previous statement, does hold too (Fig. 2).

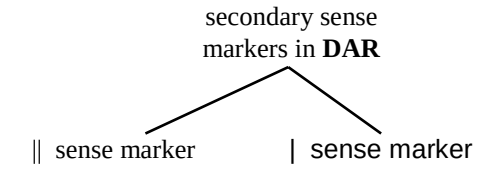


Figure 2: Non-dependency relation between || and | secondary sense markers in **DAR**

(e) Finally, the atomic sense definitions receive the smallest priority scores (represented with the greatest even natural number, compared to the other sense scores), since their lexical-semantic granularity is the smallest. For instance, the atomic senses of the *RegDef*, *BoldDef*, or *ItalDef* definitions (Curteanu et al., 2012) may *all* receive the priority score 16 (or 18), whether there are no established dependency relations among them, while to the literal enumeration under such (a block of atomic) sense

definitions should be assigned the priority score 17 (or 19). Under such an enumeration, no other lexical-semantic refinement is permitted.

The *total ordering procedure* for the representation of *sense marker classes* in DHs, especially including their literal enumeration refinement, can be replaced by any other numerical or literal encoding of the sense priority scores within dictionary entries, provided that it can be preserved the total ordering of the sense definitions, entailed by their lexical-semantic granularities and delimited by the corresponding sense markers.

5. Least Upper Bounds of Optimized Dependency Hypergraphs

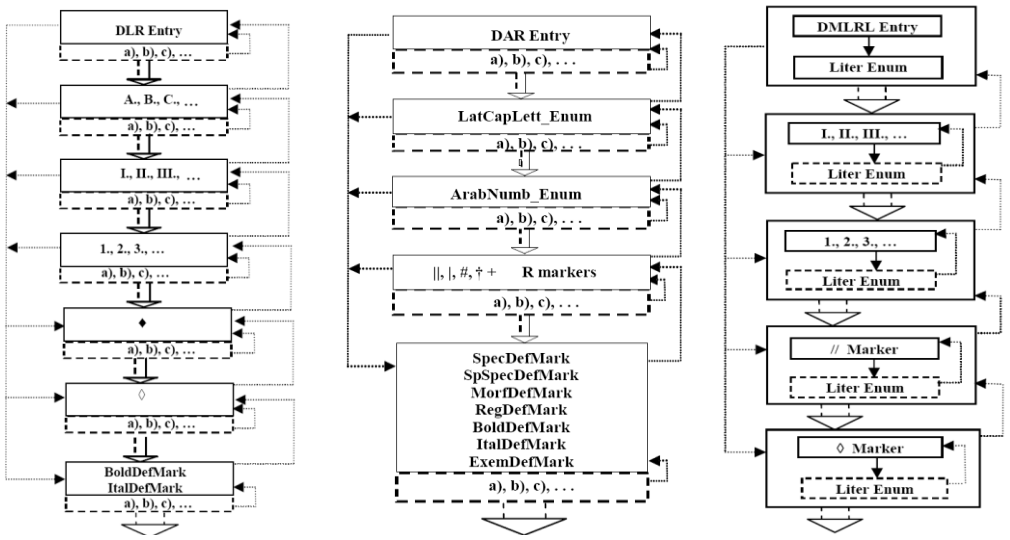


Figure 3: Linearly recursive (optimized) DHs for **DLR**, **DAR**, and **DMLRL** – version 2013

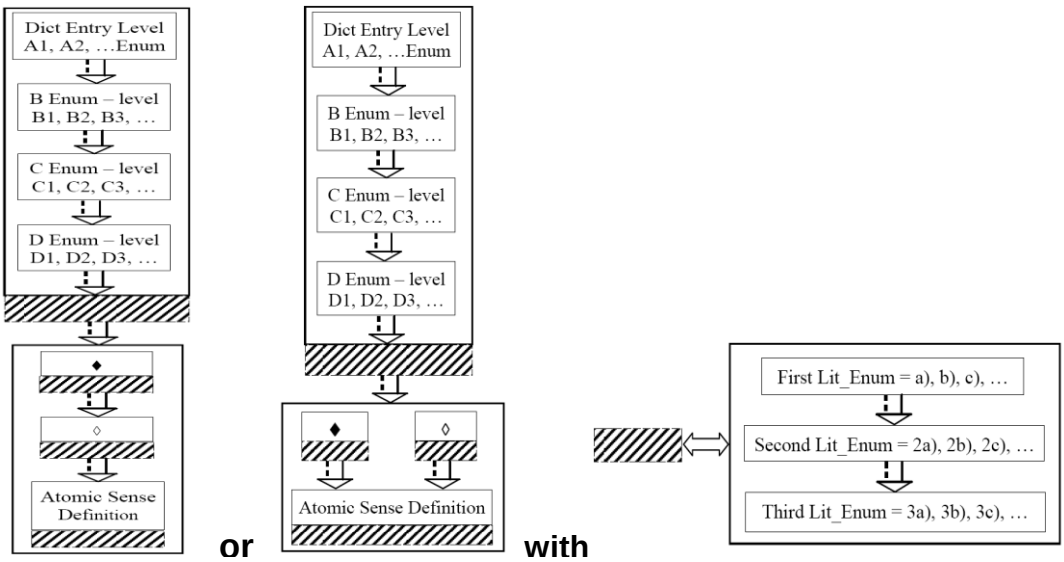


Figure 4: The two DHs as the LUB (unification procedure) outcome of the three DHs in Fig. 3

In (Curteanu & Moruz, 2012b), constructing the procedural DTDs on the three main parsing levels (SCD configurations) was outlined as a result of building LUBs of the par-grammars derived from DHs, which at their turn were designed on each SCD configuration of the considered six thesaurus dictionaries. The elegant and efficient solution to the problem of optimizing DHs, based on the total ordering of sense marker classes, remarkably including the literal enumeration(s), opened a much simpler approach to the procedural DTD computing as the (only) *par-grammar* of $DH_{LUB} = \underset{i}{LUB}(DH_i)$, such as the two DHs in Fig. 4, obtained through matching-unification algorithms, as the LUB-outcome of the three optimized DHs in the Fig. 3 above.

6. Parametrized Grammars for Linear-Cyclicity Dependency Hypergraphs

We propose the following *par-grammar* assigned to *the first* DH in Fig. 4, which is one of the two LUB DHs obtained from the three optimized DHs in Fig 3, on the SCD-Config2 parsing level for **DLR**, **DAR**, and **DMLRL**. The grammar rules are grouped in packages according to the direction of generation: *descending rules* go towards less general senses (*e.g.* from **A.** to **B.** enumeration), *ascending rules* return to super-ordinated senses (*e.g.* from **C.** to **B.** enumeration), expressing the *Enumeration Closing Condition* (ECC) in (Curteanu et al., 2012), while *splitting rules* are calls to the enumeration partitioning.

The grammar rule attributes are *parent* and *item*. The *parent* of a node is the sense from which that node is generated, and the *item* of an element is its position in the list of its sister elements. In order to jump over sense levels, as most dictionaries do (*e.g.* from **A.** to **C.** enumeration class), we have used a *dummy node* for each skipped level, as the grammar is built such that it cannot generate a lower sense level without its super-ordinated level (this is a correctness restriction). The dummy nodes derivate to the empty string and are not itemized (the *item* attribute is never incremented for them).

Table 1: Par-grammar for the first DH in Fig. 4, as the LUB(DH_i) outcome of the optimized DHs in Fig. 3

```
//primary_sense → entry | LatCapLetA | LatCapLetB |
                                LatCapLetC

entry → newPrg e LatCapLetA;
    parent(LatCapLetA) = e; item(LatCapLetA) = 0
entry → e
entry → e LatSmallLet;
        parent(LatSmallLet) = e; item(LatSmallLet) = 0
... … … … … ... … … … …... … … … …... … … … … …
---------------------------------------------------------- 2 --
LatCapLetC → LatCapLetC_Mrk FilledDiamond;
    parent(LatCapLetC_Mrk) = parent(LatCapLetC);
        item(LatCapLetC_Mrk) = item(LatCapLetC) + 1;
    parent(FilledDiamond) = LatCapLetC_Mrk;
        item(FilledDiamond) = 0
LatCapLetC → LatCapLetC_Dummy FilledDiamond;
    parent(LatCapLetC_Dummy) = parent(LatCapLetC);
        item(LatCapLetC_Dummy) = item(LatCapLetC);
    parent(FilledDiamond) = LatCapLetC_Mrk;
        item(FilledDiamond) = 0
LatCapLetC → LatCapLetC_Mrk;
    parent(LatCapLetC_Mrk) = parent(LatCapLetC);
item(LatCapLetC_Mrk) = item(LatCapLetC) + 1;
==splitting==
LatCapLetC → LatCapLetC_Mrk LatSmallLet;
    parent(LatCapLetC_Mrk) = parent(LatCapLetC);
        item(LatCapLetC_Mrk) = item(LatCapLetC) + 1;
    parent(LatSmallLet) = LatCapLetC_Mrk;
        item(LatSmallLet) = 0
==ascending==
LatCapLetC → LatCapLetB;
    parent(LatCapLetA) = parent(parent(LatCapLetC));
        item(LatCapLetA) = item(parent(LatCapLetC))
---------------------------------------------------------- 4 –
==enumeration==
==descending==
LatSmaLet → LatSmaLet_Mrk LatSmaLet2;
    parent(LatSmaLet_Mrk) = parent(LatSmaLet);
    item(LatSmaLet_Mrk) = item(LatSmaLet) + 1;



    parent(LatSmaLet2) = LatSmaLet_Mrk;
     item(LatSmaLet2) = 0;
LatSmaLet2 → LatSmaLet2_Mrk LatSmaLet3;
```

```
//attributes are similar to those in the previous rule
LatSmaLet3 → LatSmaLet3_Mrk FilledDiamond;
    parent(LatSmaLet3_Mrk) = parent(LatSmaLet3);
//attributes are similar to those in the previous rule
==ascending==
FilledDiamond → LatSmaLet3, if parent(FilledDiamond) =
            LatSmaLet3;
    parent(LatSmaLet3) = parent(parent(FilledDiamond));
        item(LatSmaLet3) = item(parent(FilledDiamond))
LatSmaLet3 → LatSmaLet2, if parent(LatSmaLet3) =
            LatSmaLet2;
    parent(LatSmaLet2) = parent(parent(LatSmaLet3));
        item(LatSmaLet2) = item(parent(LatSmaLet3))
LatSmaLet2 → LatSmaLet, if parent(LatSmaLet2) =


LatSmaLet;
    parent(LatSmaLet) = parent(parent(LatSmaLet2));
item(LatSmaLet) = item(parent(LatSmaLet2))
LatSmaLet → LatCapLetC, if parent(LatSmaLet) =
        LatCapLetC;
    parent(LatCapLetC) = parent(parent(LatSmaLet));
        item(LatCapLetC) = item(parent(LatSmaLet))
LatSmaLet → LatCapLetB, if parent(LatSmaLet) =
        LatCapLetB;
    parent(LatCapLetB) = parent(parent(LatSmaLet));
        item(LatCapLetB) = item(parent(LatSmaLet))
LatSmaLet → LatCapLetA, if parent(LatSmaLet) =
        LatCapLetA;
    parent(LatCapLetA) = parent(parent(LatSmaLet));
        item(LatCapLetA) = item(parent(LatSmaLet))
LatSmaLet → "", if parent(LatSmaLet) = entry;
------------------------------------------------------------ 5 –
//secondary_sense → FilledDiamond | EmptyDiamond |
                    BoldMrk | ItalMrk
==descending==
FilledDiamond → ♦ EmptyDiamond;

    parent(♦) = parent(FilledDiamond);
        item(♦) = item(FilledDiamond) + 1;
    parent(EmptyDiamond) = ♦;
        item(EmptyDiamond) = 0
FilledDiamond → FilledDiamond_Dummy
        EmptyDiamond;
    parent(FilledDiamond_Dummy) = parent(FilledDiamond);
        item(FilledDiamond_Dummy) =
            item(FilledDiamond);
    parent(EmptyDiamond) = FilledDiamond_Dummy;
        item(EmptyDiamond) = 0
FilledDiamond → ♦;
    parent(♦) = parent(FilledDiamond);
        item(♦) = item(FilledDiamond) + 1;
==ascending==

FilledDiamond → LatCapLetC;
    parent(LatCapLetC) = parent(parent(FilledDiamond));
        item(LatCapLetC) = item(parent(FilledDiamond))
FilledDiamond → LatSmallLet, if parent(FilledDiamond) =
            LatSmallLet;
    parent(LatSmallLet) = parent(parent(FilledDiamond));
        item(LatSmallLet) = item(parent(FilledDiamond))
==splitting==
FilledDiamond → ♦ LatSmaLetFD1;
    parent(♦) = parent(FilledDiamond);
        item(♦) = item(FilledDiamond) + 1;
    parent(LatSmaLetFD1) = ♦; item(LatSmaLetFD1) = 0
LatSmaLetFD1 → LatSmaLetFD1_Mrk LatSmaLetFD2;
    parent(LatSmaLetFD1_Mrk) = parent(LatSmaLetFD1);
        item(LatSmaLetFD1_Mrk) =
            item(LatSmaLetFD1) + 1;
    parent(LatSmaLetFD2) = LatSmaLetFD1_Mrk;
        item(LatSmaLetFD2) = 0
LatSmaLetFD2 → LatSmaLetFD2_Mrk LatSmaLetFD3;
    parent(LatSmaLetFD2_Mrk) = parent(LatSmaLetFD2);
        item(LatSmaLetFD2_Mrk) =
            item(LatSmaLetFD2) + 1;
    parent(LatSmaLetFD3) = LatSmaLetFD1_Mrk;
        item(LatSmaLetFD3) = 0
LatSmaLetFD3 → LatSmaLetFD3_Mrk LatSmaLetFD3
    parent(LatSmaLetFD3_Mrk) = parent(LatSmaLetFD3);
        item(LatSmaLetFD3_Mrk) =
            item(LatSmaLetFD3) + 1;
    parent(LatSmaLetFD3) = parent(LatSmaLetFD3_Mrk)
        item(LatSmaLetFD3) = item(LatSmaLetFD3_Mrk)
==ECC==
LatSmaLetFD1 → FilledDiamond
LatSmaLetFD2 → LatSmaLetFD1
LatSmaLetFD3 → LatSmaLetFD2
```

Table 2:Schematic grammars for **DLR** entry parsing on the lexicographic-segment and sense-tree levels

```
    // lexicographic segment parsing in DLR / DAR
entry → entryMarker entryRootSense entryBody
        entryTail
entryBody → S
S → Seg | Seg S
Seg → Mrk Root_sense Body_sense Tail_sense
Mrk → "" | depTreeNode_SCD1
Root_sense → "" | text | subSegMrk
Body_sense → "" | senseBodyDLR | frBodyDAR |
    roBodyDAR | senseBodyDAR | nestDAR |
    MorphologicalPart
Tail_sense → ""
    // sense tree parsing in DLR
senseBodyDLR → sense
sense → senseMarker definition sense_list
sense_list → sense sense_list | ""definition → defItem
definition | defItem
defItem → MorfDef | spSpecDef | specDef | regDef |
        defExemList
defExemList → defExemPair defExemList |
        defExemPair
defExemPair → quote sigle
regDef → regDefPart regDef | regDefPart
regDefPart → regDefPartComponent regDefPart |
        regDefPartComponent
regDefPartComponent → gloss | reference | synonym |
        sigle | specDef | spSpecDef
specDef → ( specDefPart specDefRec) | ( specDefPart )
specDefRec → specDefPart specDefRec | specDefPart
specDefPart → specDefKeyword | freeText
```

The effective construction of a new, *procedural* DTD and of a corresponding SCD-based *general parser* for large dictionaries is the result of the following steps:

(S1): For each new dictionary, in the process of its lexicographic modeling (Curteanu et al., 2010, 2012), the DHs for the three main parsing levels (*viz.* SCD*k* configurations, $k = 1 \div 3$) have to be well-defined, including their calls between the three parsing levels to be structurally embedded. Whether necessary, the essential process of their *optimization* has to be applied, *i.e.* their recursiveness linearization by eliminating the non-embedded call cycles between sense marker classes and literal enumeration. **(S2):** On each SCD*k* configuration, $k = 1 \div 3$, for the dependency hypergraphs SCD*k*-DH$_i$ ($i = 1 \div n$) defined for *n* distinct dictionaries, the LUB$_i$(SCD*k*-DH$_i$) = SCD*k*-DH, $k = 1 \div 3$, has to be defined. The optimization procedure in Section 4 assures the process soundness. *E.g.*, the three SCD*2*-DH$_i$ ($i = 1 \div 3$, for DLR, DAR, DMLRL) have been defined in Fig. 3, their SCD*2*-DHs = LUB$_i$(SCD*2*-DH$_i$), *i* = DLR, DAR, DMLRL, being displayed in Fig. 4. **(S3):** For each SCD*k*-DH, $k = 1 \div 3$, its par-grammars represent the procedural DTD$_k$ of the $i = 1 \div n$ considered dictionaries. Their representational DTD is the unified package of the three par-grammars on the SCD*k* configurations, $k = 1 \div 3$. *E.g.*, the par-grammar in Table 1 is associated to SCD*2*-DH = LUB$_i$(SCD*2*-DH$_i$), *i* = DLR, DAR, DMLRL. **(S4):** Several par-grammars have to be integrated within each par-grammars$_k$ associated to the SCD*k* parsing level ($k = 1 \div 3$). **(S4a):** Par-grammars for constructing the dependency trees of the lexicographic structures on each SCD*k* level (*e.g.*, the schematic grammars for segment and sense tree parsing in Table 2 for SCD*2*). **(S4b):** Backus-Naur grammars and their LUB outcome(s), for the atomic sense definitions of the involved dictionaries. **(S4c):** The procedurally connected and / or LUB-computed par-grammars for *all* the above considered formal grammars should constitute the new procedural DTD, resulted incrementally for the *n* dictionaries at hand.

The procedural DTD and its SCD-based associated parser for very large dictionaries still deserve substantial efforts and innovative solutions in order to be accomplished.

7. Conclusion and Continuation

The DH optimization involves the following remarks, driving to the solution of our problem: **(a)** The literal enumeration **a).**, **b).**, **c).**, … under **I.**, **II.**, **III.**, … primary sense

markers is **not the same** as **a).**, **b).**, **c).**, … under **1.**, **2.**, **3.**, … since the lexical-semantic granularity of the former literal enumeration is strictly larger than that of the latter. **(b)** The same fact holds, with even more substance and practice, for ♦ super-ordinating ◊ secondary sense markers. **(c)** For a sound parsing of dictionary entries, the solution to DH optimization problem entails a **sense marker-depending renaming** of the **literal enumerations**, **totally ordering** these sense splitting processes in DHs. In (Curteanu & Moruz, 2012b), a par-grammar has been proposed to represent the DLR DH, the first (optimized) DH in Fig. 3. In the presence of non-optimized DHs, computing their par-grammars, and then their *least upper bound* (LUB) par-grammar, is an intricate process. Solving the problem of DH optimization changes radically the solution to obtaining the general DTD and dictionary parser (Section 6): instead of computing the LUB of par-grammars from non-optimized DHs, we apply the optimization procedure to the DHs of the involved dictionaries, compute their LUB DH(s), and write its (or their) corresponding par-grammar(s).

The project of a new, procedural, DTD and of a general SCD-based parser for the largest thesaurus-dictionaries is a huge challenge because it would make possible a direct comparison among the sense marker classes utilized in the most computerized languages, among the adequacy of the lexicographic sense markers and the lexical-semantics granularity of the lexicographic units they delimit within various large dictionaries. It brings the effective means for a standardization of these such complex constructions and their automatic (and efficient) parsing. As further developments of the standardized thesauri one can mention the design of an optimal and cross-linguistic compatible *network of Romanian electronic dictionaries*, similar to a very good project of dictionary network, *viz.* the German Woerterbuch-Netz, with possible links to well-known foreign dictionaries.

References

Curteanu, N., Trandabăţ, D., Moruz, A. M. (2008). Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing, *Proceedings of CogAlex Workshop*, Manchester, 55-63, http://aclweb.org/anthology/W/W08/W08-1908.pdf

Curteanu, N., Moruz, A., Trandabăţ D. (2010). An Optimal and Portable Parsing Method for Romanian, French, and German Large Dictionaries. *CogAlex II – The Second Workshop on Cognitive Aspects of the Lexicon*, COLING-2010, Beijing, China, 38-47, http://www.aclweb.org/anthology-new/W/W10/W10-3407.pdf

Curteanu, N., Cojocaru, S., Burcă, E. (2012). Parsing the Dictionary of Modern Literary Russian Language with the Method of SCD Configurations. The Lexicographic Modeling. *Computer Science Journal of Moldova*, Academy of Sciences of Moldova, Vol. 20, No.1(58), 42-81, http://www.math.md/files/csjm/v20-n1/v20-n1-(pp42-82).pdf

Curteanu, N., Moruz, A. (2012a). Toward the Soundness of Sense Structure Definitions in Thesaurus-Dictionaries. Parsing Problems and Solutions. *Computer Science Journal of Moldova*, Academy of Sciences of Moldova, Vol. 20, No.3 (60), 275-303, http://www.math.md/files/csjm/v20-n3/v20-n3-(pp275-303).pdf

Curteanu, N., Moruz, A. (2012b). A Procedural DTD Project for Dictionary Entry Parsing Described with Parameterized Grammars. *CogALex*-3 Proceedings, *The*

Third Workshop on Cognitive Aspects of the Lexicon, COLING-2012, Bombay, India, 127-136, http://aclweb.org/anthology-new/W/W12/W12-5110.pdf .

Erjavec, T., Evans, R., Ide, N., Kilgariff A. (2001). From Machine Readable Dictionaries to Lexical Databases: the CONCEDE Experience. *Research Report* on TEI-CONCEDE LDB Project, Univ. of Ljubljana, Slovenia. Consortium for Central European Dictionary Encoding – INCO-COPERNICUS project no. PL96-1152.

Hauser, R., Storrer, A. (1993). Dictionary Entry Parsing Using the LexParse System. *Lexikographica* 9, 174-219.

Lemnitzer, L., Kunze, C. (2005). Dictionary Entry Parsing, ESSLLI.

Neff, M., Boguraev, B. (1989). Dictionaries, Dictionary Grammars and Dictionary Entry Parsing, *Proc. of the 27th annual meeting on Association for Computational Linguistics*, Vancouver, British Columbia, Canada, 91-101.

Tufiş, D., Rotariu, G., Barbu, A. M. (1999). TEI-Encoding of a Core Explanatory Dictionary of Romanian. *Proceedings of the 5th Comp. Lexicography COMPLEX 1999*, Pecs, Hungary, (F. Kiefer, G. Kiss, and J. Pajzs eds.), 219-228.

XCES TEI Standard, Variant P5 (2007). http://www.tei-c.org/Guidelines/P5/

ROMANIAN ETYMOLOGICAL CHAINS – A PRELIMINARY ANALYSIS

RALUCA MOISEANU[1], DAN CRISTEA[2]

[1] *Alexandru Ioan Cuza University of Iasi, Computer Science Faculty, Computational Linguistic Department*

[2] *Romanian Academy, Institute for Theoretical Computer Science;*

{raluca.moiseanu, dcristea}@info.uaic.ro

Abstract

In this paper the origin of describe the preliminary steps towards a recursive reconstruction of Romanian words together with the positioning of their loans within a time frame, as reflected in the European Linguistic Thesauri. A pilot application accepts as input a Romanian word and accesses online linguistic resources, such as eDTLR – *The Thesaurus Dictionary of the Romanian Language in electronic form*, displaying etymological information. The etymology of a word is subsequently searched in foreign sources (for the time being only French and Italian online dictionaries), in order to compute its etymological trajectory. Import years, where available, are used to place on the time axes the approximate time of imports. The research intends to highlight a methodological framework on which a future real scale investigation could be anchored.

Keywords: etymon, online dictionaries, database, parser

1. Introduction

This project has been triggered by the need of having a dynamic and complex structured database able to provide the etymological information of any Romanian word (except the ones with unknown etymology). In our attempt to recreate the etymological chain of a word we shall, first of all, provide an insight of what etymology as a science is, as well as the main features of the Romanian etymology. Once the theoretical background is established we shall move on to the linguistic resources and technologies used to support the generation of etymological chains.

An etymological chain is a string of one or more etymons along with their origin language and entry year. As data structure, etymological chains are graphs (Alt, 2006) that have a root word in the studied languages (Romanian, in our case) and one or more descendants from source languages (Central and Eastern languages, in our case, with whom Romanian languages has had contact throughout the years).

The paper describes the beta version of the application used to automatically extract the information from online Italian and French dictionaries, version that has been tested on a number of 2000 XML files from the eDTLR – the Romanian Thesaurus Dictionary in electronic form (Cristea et al., 2007), corresponding to the same number of dictionary entries.

2. Etymology as a science

Derived from the Greek *etymon* meaning "true sense" and the suffix, *logia*, denoting "the study of", etymology as a science studies the origin of words. Etymology considers words as having either an internal origin (therefore, in the target language, by applying transformation rules specific to the lexicon or the grammar of that language, through affixation, compounding and conversion) or an external origin (through borrows/loans from one or more languages). Regardless the acceptance channels, the etymology has to decipher the phonetic and morphological transformations from the original word to the actual word.

The Linguistic calque is to be situated at the border between the internal and external generation of words as the new words are formed within the source language by imitating an external structure.
An etymon can come from two or more languages either during the same period of time or throughout different periods of time. This is called multiple etymology. Most of the Romanian words have multiple etymology, Latin being referred to as an indirect source.
Romanian is a Romance language, belonging to the Italic branch of the Indo – European language family, having much in common with languages such as French, Italian, Spanish and Portuguese.
However the closest to Romanian are the other Eastern Romance dialects, spoken south of Danube: Aromanian/Macedo-Romanian, Megleno-Romanian and Istro-Romanian dialects. An alternative name for Romanian used by linguistics to disambiguate with the other Eastern Romance languages is Daco-Romanian, referring to the area where it is spoken (which corresponds roughly to the onetime Roman province of Dacia).
Marius Sala et al (1988) considered 2581 words as being representative for the Romanian vocabulary. The etymological structure of this vocabulary is shown below:

- Romance elements 71.66%, out of which:
  - 30.33 % Latin
  - 22.12 % French
  - 15.26 % Classical Latin
  - 3.95 % Italian
- Internally formed 3.91 % (most from Latin etymons)
- Slavic 14.17 %, out of which:
  - 9.18 % Old Slavic
  - 2.6 % Bulgarian
  - 1.12 % Russian
  - 0.85 % Serbian-Croatian
  - 0.23 % Ukrainian
  - 0.19 % Polish
- German 2.47 %
- Neo-Greek 1.7 %
- Thracian – Dacian, a sub-layer, 0.96 %
- Hungarian 1.43 %
- Turkish 0.73 %

- English 0.07 % (and growing)
- Onomatopoeias 0.19 %
- Unknown origin 2.71 %

The data listed above has been used to establish the first two Latin languages of focus for this preliminary study.

3. Data collection

The collection of resources (online dictionaries) and simulation, trials of manually generated etymological chains represented the starting point of the project. The manually gathered etymological chains were also used as validators for the application (Burhui, 2013) that has been put together for the automatic generation of etymological chains.

The quest for online resources has proved itself rather sinuous as many of the online etymology dictionaries or online dictionaries did not display etymological data. For the purpose of this paper we have narrowed down the area of research to only Italian and French, which sum up (Sala, 1988) 26.07% of the representative vocabulary of the Romanian language.

Once we have identified the two online sources, for Italian – http://www.sapere.it, and for French – http://www.cnrtl.fr/etymologie, that seemed to best fit our purposes , we have extracted from these dictionaries a list of notations used to to mark the etymon (such as: *fr.*, *fr.ant.*, *it.*, *ital.*,*lat.*, *lat.class.*, *lat.vulg.*, *lat.mediev.* etc.). This list has been included as an external resource onto the program.

What follows below is a list of examples of etymological chains, manually extracted from the two online dictionaries (Italian and French). In these examples, the details that we would want the application to return upon interrogation are also indicated: POS, gender, entry year, and source language.

➢ ***bastard*** *s.m. din it. bastardo;*

IT. bastardo sec XV dal fr. ant. batard;

FR. batard 1150 l'orig. de bastard est obsc.;

ro.bastard ← **it.bastardo** ← **fr.batard** ← **unknown etymon;**

➢ ***ciment*** *s.n. sec XIX din it. cimento, fr. Ciment* ;

IT. Cimento s.m. dal lat. Caementum;

FR. Ciment s.m. 1165-70 du lat. class. caementum;

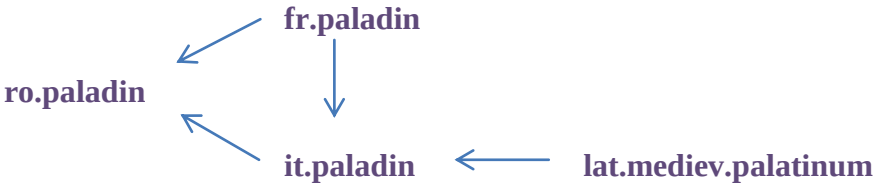
➢ ***cortina*** *s.f. din it. Cortina;*

IT. Cortina n.f. dal lat. tardo Cortina;

➢ ***paladin*** *s.m. din fr. paladin, it. paladino;*

FR. paladin s.m. 1552 empr. a l'ital. paladino;

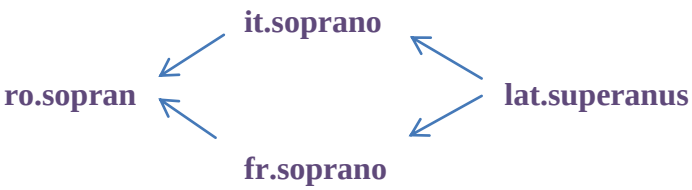
IT. paladino n.m. dal lat. mediev. palatinum;



➢ ***sopran*** *s.m. din fr., it. soprano;*

FR. soprano 1768 du lat. vulg. superanus;

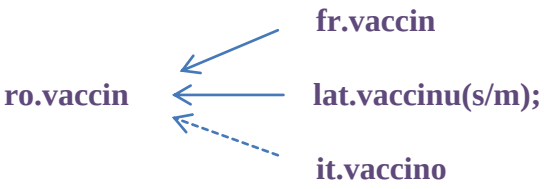
IT. soprano dal lat. vulg. superanus;



➢ ***vaccin*** *s.n. 1827 din fr. vaccin, lat. vaccinus, cf. it. vaccine;*

FR. vaccin 1801 du lat. vaccinu(s);

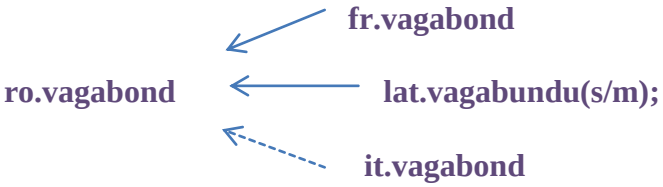
IT. vaccino dal lat. vaccinu(m);



➢ ***vagabond*** *s.m. 1795 din fr. vagabond, lat. vagabundus, cf. it. vagabond;*

FR. vagabond 1382 du lat. vagabundu(s);

IT. vagabondo dal lat. vagabundu(m).



As shown in the above examples we have manually extracted the entry year, where available and also listed the etymons with unknown origins.
Most of the above examples have double etymology, the etymon being both Italian and French, both pointing to Latin as being an indirect origin for the Romanian words.

From this early stage four types of etymological chains can already be seen:

- type1: root ← orig1 ← orig2 ← orig*
- type2: root ← orig1, orig1 ↔ orig3, orig2 ← orig3
- type3: root ← orig1, orig1 → orig2, orig2 → orig3
- type4: root ← orig1, orig1 ← orig2, orig2 ⇠ orig3

4. The application and a comparison with other approaches

A beta version of the application (Burhui, 2013) allows a user to input an entry Romanian word, out of which it generates one or more linear etymological chains. At this stage the application searches the entry in the Romanian lexicographic thesaurus (eDTLR) and, once found, it extracts the etymological information. If the etymological sources indicate a French or an Italian origin, it directs the search onto the corresponding French (http://www.cnrtl.fr/etymologie) or Italian (http://www.sapere.it) online dictionaries, parses the etymological information and displays it. The year of the import is filled in as the year of the first citation.

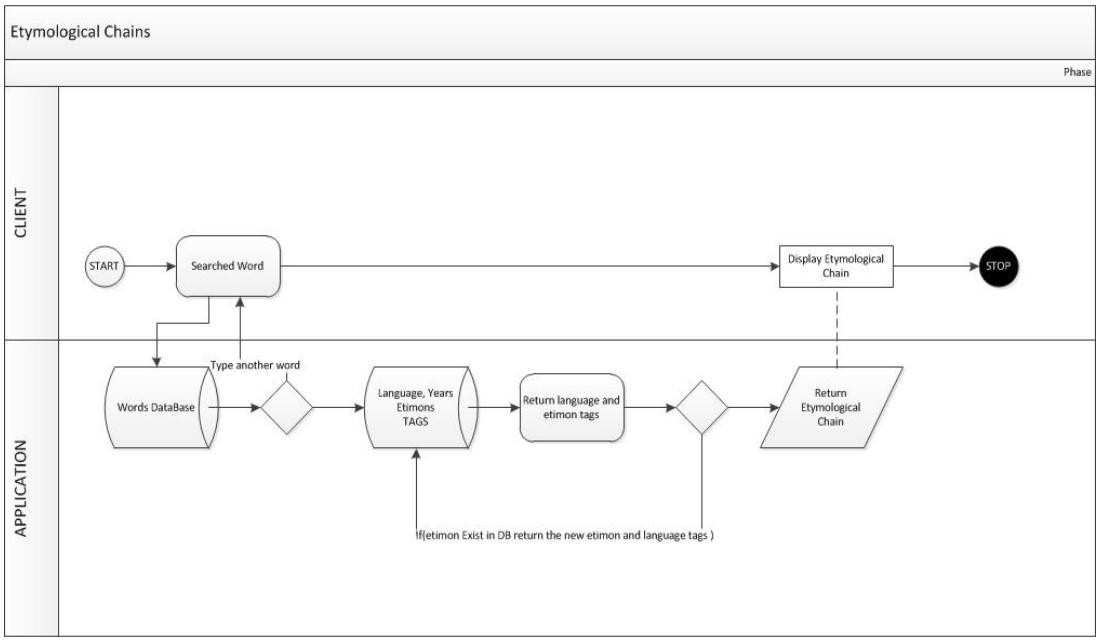


Figure 1: The general architecture of the system

A high level overview of the application design is shown in **Figure 1**. A graphical interface able to display the four schemes put in evidence in the previous section remains to be implemented.
Susan Alt (2006) describes the etymons as being words, located in time and space, which stand in a particular diachronic relation to other words, and etymological links as being the etymological relations between linguistic units. In her attempt to define a model of etymological structures she uses the TLFI (http://www.tlfi.fr) as the primary linguistic material to recover data.

The nodes of her graph are lexical entries in diverse lexicographic sources. In linear chains, the first entry is the anchoring word, the second one represents its direct etymon, the third one – the etymon of the first etymon, a.s.o. In case of compound words, her graphs diverge towards two entries, each one continued with their corresponding sources.
Alt pays a particular attention to the type of links between the entries (such as loan word relations or compound word relations). This type of information will be inserted also in our graphs once parsers would become refined enough to be able to distinguish this type of information in the source dictionaries.

5. Conclusions

Our preliminary manual investigations, as well as the first experiments done with the tool have brought to light a high number of entries with unknown or uncertain etymons, which can easily turn into the subject of some statistics drawn based on this project. Moreover, the attachment of the import dates makes it possible to detect some incorrectly dated etymons (which, as mentioned are extracted in our primary source from the date of the first mention of the imported word). Among the peculiar etymological chains that we have obtained during our manual trials we have stumbled across entries for which the first or second etymon entry year is subsequent to the one of the target language. What we believe to be incorrectly dated etymons would have to be validated against a collection of online dictionaries, rather than just one dictionary.

The main difficulty that we have faced so far is the lack of online resources that would contain the etymon and entry year as well, or the lack of online resources altogether (Bulgarian, Slavic, etc.). Among the resources that we have found so far (English, German, Spanish), differences in notation of the etymon in each dictionary makes the parsing challenging. However, in the future we aim to increase the number of online dictionaries accessed, the most wanted for studying the origins of Romanian being the German, Bulgarian, Russian, Turkish, Greek, English, Polish, Ukrainian, Hungarian and Latin dictionaries.

The language barrier is not to be neglected as well. The Greek, Turkish, Russian and Slavic dictionaries pose a real issue upon retrieving the required information.

Although the first steps have been done, the project is far from being completed and also raising more questions than solutions.

We believe that the research, only an inception of which is described in this paper, would rather convincingly motivate the birth of an international consortium that would look into the development of this project at European scale. Let's note that similar initiatives have been suggested already for other languages: (Alt, 1996) for French or the Etymology explorer (http://roots.robestone.com) for English. Agreeing on some common conventions of notation of etymological chains, sharing lexicographic resources, parsing technologies for dictionaries and the software that builds the etymological graphs itself, could result in a reconstruction of interchangeable etymological graphs that would configure more and more dense parts of a map of linguistic influences. Their correlation with historical events could bring into light new insights over cultural interferences, could correct errors and reveal unknown linguistic and historical facts.

Acknowledgments: We are grateful to Gabriela Haja, from the "A.Philippide" Institute of the Romanian Academy for coining the idea of etymological chains, and to Andrei Scutelnicu and Alin Placinta – Salaru, from the Computational Linguistics at the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iași, and Anca Bibiri, from the Department of Interdisciplinary Studies of the same University, for contributing to the elaboration of software and for acquiring information about dictionaries.

References

Burhui A. (2013). Reconstruirea lanțurilor etimologice pentru limba română (The reconstruction of etymological chains for Romanian language). Dissertation thesis in Computational Linguistics, "Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science;

Cristea D., Răschip M., Forăscu C., Haja G., Florescu C., Aldea B., Dănilă E. (2007). The Digital Form of the Thesaurus Dictionary of the Romanian Language. In *Proceedings of SPeD-2007 (Speech Technology and Human − Computer Dialogue)*, Iași.

Hristea T. (1984). Structura etimologică a lexicului românesc modern, în: Theodor Hristea (coord.), Mioara Avram, Grigore Brâncuș, Ghorghe Bulgăr, Goergeta Ciompac, Ion Diaconescu, Rodica Bogza – Irimie, Flora Suteu, Sinteze de limba română, Bucharest;

Moroianu C. (2005). Dublete etimologice (Etymological doublets), Bucharest.

Marius Sala (coord), Mihaela Bîrlădeanu, Maria Iliescu, Liliana Macarie, Ioana Nichita, Mariana Ploae-Hanganu, Maria Theban, Ioana Vintilă-Rădulescu (1988) Vocabularul reprezentativ al limbilor romanice (The representative vocabulary of Romance languages). Editura Ştiinţifică şi Enciclopedică, Bucharest.

Susan A. (2006). Data Structures for Etymology: Towards an Etymological Lexical Network, Bulag;

Dictionaries:

eDTLR – The Thesaurus Dictionary of Romanian Language in electronic form

http://www.cnrtl.fr/etymologie;

http://www.sapere.it;

http://dexonline.ro

VIRTUAL CIVIC IDENTITY

DANIELA GÎFU[1], DAN STOICA[2], DAN CRISTEA[1,3]

[1] *"Alexandru Ioan Cuza" University, Faculty of Computer Science, Iaşi – România*
[2] *"Alexandru Ioan Cuza" University, Faculty of Letters, Iaşi – România*
[3] *Institute for Theoretical Computer Science, Romanian Academy - Iaşi branch, România*

{daniela.gifu, dcristea}@info.uaic.ro, dstoica_ro@yahoo.com

Abstract

The paper presents a study on a typology of civic identities of public contributors of online articles on forums and their possibilities of automatic identification. We analyse the dialogic means and exploration of automatic extraction of features from forum utterances. The research suggests new perspectives for defining types of online commentators of public discourses addressing domains such as politics, arts, education, etc. In the investigation we apply some pragmalinguistics approaches on communication, mainly taken from polyphony and enunciation areas. The classification of user profiles make use of criteria that take into consideration: common topics, expression of sentiments, style features, lexical n-grams, morphosyntactic analytics and pragmatic features. Our purpose was to lay the basis for a thorough classification of categories of publics and to suggest ways of their automatic identification, in the benefit of editors of media institutions, specialists in public communication, intelligence agencies, political structures, etc.

Keywords: civic identity, pragmalinguistics, semantic classes, journals forums, editors.

1. Introduction

Nowadays, a part of the reality has moved in the cyberspace. And the same happened with the bar or side-way chats, traditional in older times. Almost every public page we visit is cast into a stream of ongoing discussions, comments, gossips, thus becoming a property jointly owned by its composer and any person who may want to back react. The civic identity seems to be manifested on the Internet without constraints.

This study attempts to identify a model of identification of the *civic identity* of an individual, as revealed through online channels, by evidencing decision features whose values can be extracted automatically. The investigation focuses on a corpus of online journals' forums, from where the commentators' profiles are being extracted. Profiling the *civic identity* of readers of articles should exploit their inputs, therefore now seen in the position of writers of forums' short comments. The process puts at its basis a

panoply of pragmatic markers, extracted by linguistic methods at the following levels: lexical (tracking patterns of specific vocabulary), syntactic (grammar errors, punctuation, enumerations, repetitions, use of emoticons, etc.), semantic (frequent use of some semantic classes), discourse (rhetorical markers). Using these features, the resulted portrait should be characterised along the following dimensions: the capacity to stay within article's topic, the capacity to express opinions on another forumist's comment, tendency of presenting themselves rather then following the forum's debate, degrees of assuming their respective identity as individuals, preoccupation for really participating in the debate opened by the article (vs. just the desire to assert vague, general ideas).

When the media product is on the Internet, the actors of the cyberspace who decide to interact with it have tremendously numerous possibilities to shadow or even hide their identities, and, of course, their communicative intentions. As such, the attempt to determine the *civic identity* of people hiding their identities as individuals seems impossible. Up to date, there are no consistent instruments or studies on the different nature of forums' users, and the statistics are used just to group up reactions to the journalistic material. More than this, the lack of studies on the true nature of the forums' writers makes it impossible to apply advanced statistical calculi in order to rank positions or attitudes. A smart argument put out in well-formed phrases could reveal a civic activist, but also a good PR from a political party, trying to influence the readers of the forum; an upset man who wants to let it out on any subject could reveal a shy person accepting to express himself from behind the protection of the anonymity. Basic criteria like age, gender, education level are insufficient for determining the *civic identity* of people under study. The markers used by specialists in pragmalinguistic analysis could reveal in one's discourse a lot more on the personality of the writer than the writer would accept to unveil. Based upon this kind of findings, a typology of forums' users from the point of view of their respective civic identity is possible. This approach shows the importance of a natural language processing system capable to extract basic linguistic features from large amounts of online texts and to organize them as a collection of pragmatic knowledge aiming to inventory the profile of online commentators. The outcome of the study could provide tools for public speakers to be used for improving their future discourses. This is why the effort to mentally represent the interlocutor – and if not the actual interlocutor, the general profile s/he belongs to – is important in improving one's ability to communicate. There are many ways one could enhance his/her capacity of well representing the others before or during an online interaction. One of them is to analyze their public discourse, in order to extract information to be used in orienting your own discourse, making it efficient. A good apprehension of media products' respective publics, for example, could serve to improve their editorial politics and so be of better use for the communities they serve.

Section 2 presents the state of the art. Section 3, after a short description of the corpus analyzed, during the two hot months of the presidential crisis (July – August 2012), presents the methodology applied in identifying lexical-semantic and pragmatic features of the civic identity online. Finally, Section 4 presents some conclusions and directions for the future work.

2. State of the art

Our study combines automatic user profiling techniques (opinion mining, authorship classification) with pragmatic and linguistic studies of computer-mediated communications. In this moment, many systems collect various information about millions of people on the Web. Some of the current systems rely on the information manually provided by users. In others, information is obtained often from users' actions. In this case, user profiling requires inferring acquired information, both observable and unobservable data, such as, users' behaviour (Schiaffino and Amandi, 2009), (Zukerman & Albrecht, 2001). His/her behaviour and profile can be obtained from this information using different techniques like machine learning and statistical methods. Thus we have a wide range of techniques that were used to create user profiles, such as Bayesian networks (Nurmi, 2006), (Withby *et. al.*, 2005), (Weiwei *et. al.*, 2007), (Mui *et. al.*, 2001), (Garcia *et. al.*, 2007), fuzzy models (Grishchenko, 2004), (Sabater *et. al.*, 2002), (Manchala, 1998), association rules (Adomavicius & Tuzhilin, 2001), mechanisms of text classification (Trandabăţ *et. al.*, 2012), (Gîfu & Cristea, 2012), and more.

Discourse/text output of users (posts, comments, forum messages) is used to infer elements about authors' identity (gender, sex, age, level of education and much more). In these text productions a user expresses his/her opinions about a given topic and interacts with other users. Content analysis is used in several applications to identify conflicts (Denis *et. al.*, 2012), or to detect various opinions (Grivel et Bousquet, 2011). The challenge is to involve theories of pragmalinguistics, mainly from the works on polyphony and enunciation (Ducrot et. Anscombre, 1989, Plantin, 2005 and also Tuţescu, 2005, Kerbrat-Orecchioni, 1999, Maingueneau, 2000). Language is no longer seen as a means to represent the world (referential function of the language), but as a means of argumentation in linguistic interactions among human beings. Enunciation is making a choice from the infinite offers of a given language: a choice of words, a choice of the order in which the words are uttered, a choice in the tone, the intensity of the voice and so on and so forth. Making those choices reveal a social profile of the enunciator will be our aim, and this is what we will try to track down, in order to set up patterns. We will search for patterns of linguistic behaviour that reveal patterns of social profiles. Trying to situate our research, we shall mention that the French revue HERMÈS published along the years papers on communication and the Internet, on social relations and the online communication, or on civic exchange in the cyberspace (Loh, 2009, Akiyoshi, 2009, Cardon, 2007, Oliveri, 2011), and also that (Holt, 2004) might be a model of how to use particularities of language use to determine the kind of citizen the speaker is. Email discussion messages are often expressed in a familiar register, with slang, abbreviations, and profanity and their composers frequently seem to delight in disregarding traditional rules such as those governing syntax, conventional logic, evidence and idea development, is the idea expressed by Holt in his *Dialogue on the Internet*. (Mortensen, 2003) discusses the use of language productions to understand the mind of a player. (Stoica, 2001) comments on the degrees of liberty authors have when writing for traditional, printed scientific journals and when they write for the web.

Pragmatic and rhetoric studies identify several relevant features for characterizing specific genres focusing on the expected audience (scientific articles vs. popular science articles (Hyland, 2009). Some research projects collect new media communication documents (Lin, 2007) (Stark and Dürscheid, 2011) to study their features for classification purposes.

3. A case study

The methods, the techniques and the tools in the development phase of the project create the premises for a thorough investigation of categorisation of online civic identities, drawn from statistics on large amount of textual data. The approach has a high degree of generality that makes it applicable to other types of investigations, provided they rely on text analytics.

3.1. The corpus

For the elaboration of preliminary conclusions on the configuration process of the online „civic identity", we collected, stored and processed 11,100 relevant texts/day/newspaper (summing up 146,000 words)[1], published during July-August 2012 (July 01-06, 2012 – a week before President's suspension; July 07-11, 2012 – a week after President's suspension; August 11-16, 2013 – a week before President's return at the Cotroceni Palace) by three important Romanian online newspapers having similar profiles[2] (*Evenimentul zilei*, *Gândul*, *Jurnalul Naţional*) but usually displaying totally disjoint opinions and journalistic styles on any political topic. We talk about the hot political period when the President was suspended.

3.2. Methodology

In the following, we briefly describe the steps of our analysis:

- by attentive reading, we identified 10 typologies of commentators, that can be called: the-decent, the-porn-aggressive, the-incitator, the-linkable, the-affected, the-author-attacker, the-supporter, the-intellectual, the-rational, the-irrational (see. Table 1).

- after manually processing the whole corpus, it resulted that 6 that out of the 10 profiles were rather accidental (too few data): the-decent, the-porn-aggressive, the-linkable, the-author-attacker, the-supporter, and the-intellectual). As the average of their occurrences was under 5%, we eliminated these texts. Only the remaining 4 profiles are quantitatively analysed below.

- we established a number of features (belonging to the lexical, syntactic, semantic and discourse levels of analysis) that are, more or less, subject to automatic extraction: declared ID (hide, partial expose, expose, invented, etc.); making use of emoticons, familiarity in dialog, jokes, punctuation, etc.; the semantic classes of being rational emotional (with their sub-classes), and swear; comments that follow the topic, that have

[1] We are aware that the actual dimension of our corpus is still insufficient to obtain an accurate categorization of the clasiffication criteria, but in this study we are merely interested to investigate a research methodology than to arive to precise conclusions over types of civic identities, as revealed by text analytics.

[2] These are national dailies of general information, tabloids with a circulation of tens of thousands of copies per edition, each. The newspapers were monitored on their websites: *Evenimentul zilei* – www.evz.ro, *Gândul* – www.gandul.info, *Jurnalul naţional* – www.jurnalul.ro.

no correlation with the topic, that are connected to other comments, that are aggressive, etc.; number of appearances of the ID / article and the number of appearances ID in other online publication(s);

- all comments belonging to the same type, irrespective of their actual identity, have been put in the same folder, as belonging to the same type;

Table 1**:** Profile's typology after manually annotations

Abbreviations		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Style		caps lock, politeness formulas, quotes	porn formulas, k instead of c, sh, tz	familiar formulas, swears	links	emotional tone	familiar and aggressive formulas	too more punctuation marks	historical arguments	more info, official names	abbreviation, repeated punctuation marks
Rational		x					x		x	x	x
Positive				x							
Negative		x	x	x		x	x	x			
The nature of comment	related to topic	x							x	x	x
	off-topic		x	x		x					
	related to others						x	x			x
	incite commentators		x	x							
No. ID/articles		123	230	47721	230	19982	378	402	134	86650	32850
Profile's commentator		Decent	Porn-aggressive	Incitator	Linkable	Affected	Author-attacker	Supporter	Intellectual	Rational	Irrational

- consequently, we adorned all texts with values on the established features, either manually or automatically. For instance, the semantic level has been automatically annotated with values for each of the semantic classes residing under the general classes: emotional, rational and swear, in total, 12 semantic classes;
- these data are discussed below as possible input for training a classifier to recognise the civic identities (portrait types).

3.3.Lexical-semantic features

After eliminating six of the manually annotated profiles, as identified initially, together with their comments, the remaining corpus was processed with the DAT[3] tool (initially intended to analyse political discourses). Out of the 33 semantic classes in DAT, arranged hierarchically – see two examples of XML class definitions in (1) –, we selected only those noticed to have dominant tonalities: rational, with 5 subclasses (uncertain, inhibition, intuition, certain, and determine), emotional with 2 subclasses (positive and negative), each of them having other 3 subclasses (positive with moderation, firmness and spectacular, and negative with anxiety, anger and sadness), and swear.

```
<class name="negative" id="8">                    (1)
<class name="anxiety" id="8" parent="9">
```

[3] DAT (Discourse Analysis Tool) has some similarities with LIWC (Linguistic Inquire and Word Count), used during the American presidential elections in 2008 (Pennebaker, 2001). The Romanian lexicon resourcing DAT contains a collection of over 9,500 entries (roots and lemmas).

The placement of classes in hierarchies makes that, when an occurrence belonging to a lower level class is detected in the input file, all counters in the hierarchy, from that class to the root, be incremented.

For instance, in *Evenimentul zilei*, we can see the results outputted by DAT (Fig. 1), when analysing the streams of textual data for each semantic class. So, we analysed 4 profiles of online commentators (abbreviated with "C"), that we have considered to be predominated in cyberspace as follows:

- the first type of commentator, C5, predominate the self-confidence (the class certain), he is, rather, the type of dynamic blogger (the class emotional). In general, he comments in line with the subject, being convinced about his ideas (the class firmness);
- the second type, C10, is unsure (the class uncertain). He comments in line with the subject, because he looks for a way to get himself into the dialog;
- the third type, C3, has an insulting language (the classes swear, negative, anger). He prefers to shock the audience, in general he is out of subject or binds onto other commentators;

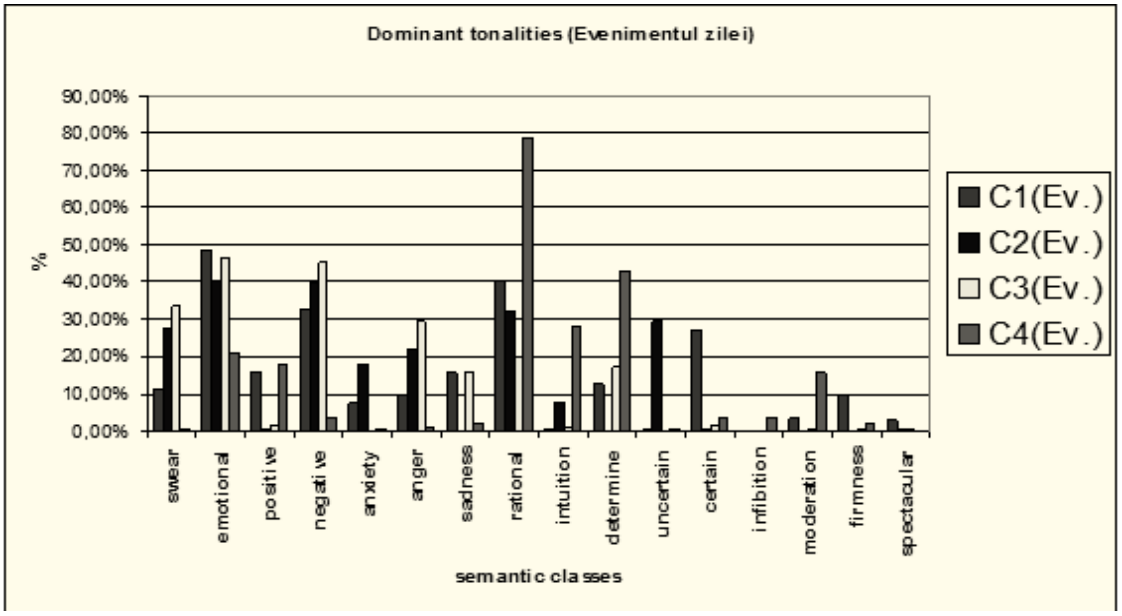


Figure 1: Analysis of user's profiles in *Evenimentul zilei* journal

- the last type, C9, adopts a rational discourse (the class rational), with sustainable arguments (the class determine), and, often, he has a moderate tone (the class moderate) about the political topics.

3.4. A comparative lexical semantic analysis between two profile's online journals

We present below a chart with two streams of data, collected during the presidential crisis, representing comments between the two profile's online journals, *Gândul* and *Evenimentul zilei*. Our experience shows that an absolute difference value below the threshold of 0,75% should be considered as irrelevant and, therefore, ignored in the interpretation. Apart from simply computing frequencies, the system can also perform comparative studies. The assessments made are comprehensive over the selected classes because they represent averages on collections of texts, not just a single text.

To exemplify, one type of graphics considered for the interpretation was the one-to-one difference, as given by Formula (2), included in the DAT Mathematical Functions Library:

$$Diff_{x,y}^{1-1} = average(x) - average(y) \qquad (2)$$

where *x* and *y* are two streams; *average(x)* and *average(y)* are the average frequencies of *x* and *y* over the whole stream, and the difference is computed for each selected class. So, the graphical representation in Figure 2, where the commentator C1 of *Gândul* is compared against the commentator C1 of *Evenimentul zilei*, should be interpreted as follows:

- the first profile, C1, is much better argued than the second one (the classes rational, firmness), predominating self-confidence (the class certain), and uttered in an affective tone (the classes emotional, negative);
- the second profile, C1, is more emotionally implicated in comments, manifesting upset, even anxious (the classes anxiety and anger). He prefers to comment with sustainable arguments (the class determine), but, often with a precaution tonality (the class moderate) because he has no intention to start a dispute with the others.

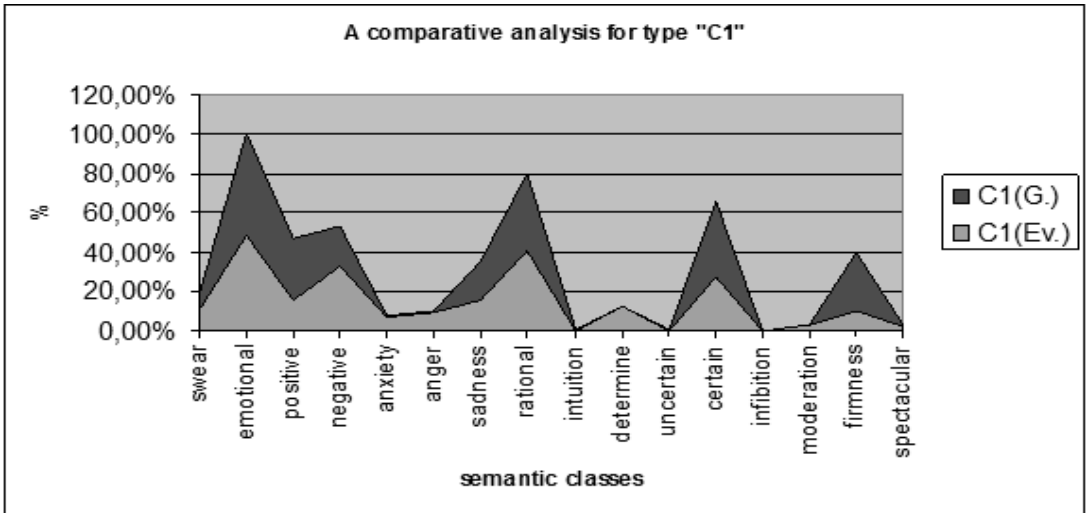


Figure 2 A comparative analysis between the users profiles in the journals *Gândul* and *Evenimentul zilei*

3.5. The pragmatic perspective

The pragmatic analysis should be based on the knowledge of the civic intentions of the commentator in connection with the meanings of the article or of the other comments. Only a good knowledge of the civic aspirations of the receptors and knowing that the editor knows himself this spectrum of civic aspirations, could make a human analyst succeed in interpreting the whole range of subtleties of a comment. It is clear that pragmatics makes a good deal of the forums interpretation process. It is nevertheless true that an experienced human analyst would succeed to acquire these facets of the pragmatic context of a comment even having little direct knowledge on them. It is like in an act of reverse engineering in which the analyst is able to infer the civic behaviour of the speaker or of the receptor from the text itself.

A closer look on a pragmatic analysis of online comments reveals the following aspects: interpretation of the text in terms of psychological distance between the

partners, opponents, etc.; defining the transmitter's attitude before and after the instance of communication; determining the receptor's attitude (i.e. being pro, against or undecided); pursuing echoes of the article in the audience (immediately), or in time (offline comments), etc.; discovering the writer's intentions by evidencing the semantic roles of different sentence constituents (reiterations, expressions, etc.).

4. Conclusions

The discourse is a place where the personality is disclosed, but not at a level of certainty that could lead to establishing incontestable patterns. Furthermore, on Internet, possibilities for manipulating information are endless. Manipulation by people who design web sites or participate in discussion groups can give the clues whether the information on a site is reliable. However, by using statistical tools and pragmatic methods we will challenge these risks on safer ground than before.
Some features are mentioned earlier only for a theoretical reason, as their effective recuperation in the text by the technology is still out of the present day possibilities. For instance, the intentions of the online commentator, a feature falling into the pragmatic perspective, are not yet technologically feasible. An author of a text is conscious that he wakes a reaction onto the reader's mind, so her/his message has an intentional component (we talk mainly about conscious intentions, as they can be reflected in the author's and/or the editor's convictions about how the reader could be influenced). However, the automatic detection of the authors' intentions, apart from the line of research triggered by the Attentional State Theory (Grosz & Sidner, 1986) and Rhetorical Structure Theory (Mann & Thompson, 1988), are still far from being conclusive.

This research opens a new direction for the study of online journals' commentators in areas such as: politics, culture, education, etc. The DAT tool becomes a necessary instrument of editorial policy and public relations departments. The study presented shows how one could shape profiles of commentators on forums of online publications (which are in a permanent dynamism). As the cyberspace is the perfect environment for hiding one's identity, some risks occur from this.

An analysis on the lines presented in this study could prove helpful to different categories of beneficiaries, mainly media editors and PR specialists. They could use the results of such analyses to better plan their policy, to adapt to different categories of public they might not even imagine be part of the general public (as they call it). Public segmentation is a continuous activity for PR specialists, and it has to be performed by using adequate criteria for each topic they want to develop in a discourse.

This kind of research could and will be continued further on: as society changes, media techniques change, the relation between media and their reader's changes all the time, and last but not least, the civic identity changes, but the need to know whom you can count on remains of paramount importance.

Acknowledgments: In performing this research, the first author was supported by the POSDRU/89/1.5/S/63663 grant.

References

Adomavicius, G., Tuzhilin, A. (2001). Using Data Mining Methods to Build Customer Profiles, *IEEE Computer* 34:2.

Akiyoshi, M. (2009). Les Japonais en ligne: le prisme des générations et des classes sociales, in HERMÈS (55).

Cardon, D. (2007). Le style délibératif de la «blogosphère citoyenne», in HERMÈS (47).

Denis, Al., Quignard, M., Freard, D., Detienne, F., Baker, M. and Barcellini, F. (2012). Détection de conflits dans les communautés épistémiques en ligne? Grenoble.

Ducrot, O. et. Anscombre, J.-C. (1989). Logique, structure, énonciation. Lectures sur le langage, Minuit.

Garcia, P., Amandi, A., Schiaffino, S., Campo, M. (2007). Evaluating Bayesian Networks' Precision for Detecting Students' Learning Styles. *Computers and Education* 49:3.

Gîfu, D., Cristea, D. (2012). Multi-dimensional analysis of political language. Future Information Technology, Application, and Service: FutureTech2012 (volume 1) Springer, Netherlands (James J. , Jong Hyuk Park, Victor Leung, Taeshik Shon, Cho-Li Wang Eds.)

Grishchenko, V. (2004). A fuzzy model for context-dependent reputation, at the Trust, Security and Reputation, *Workshop at ISWC*, Hiroshima, Japan.

Grivel, L., Bousquet, O. (2011). A discourse analysis methodology based on semantic principles - an application to brands, journalists and consumers discourses, *Journal of Intelligence Studies in Business* 1.

Grosz, B.J., Sidner, C.L. (1986). Attentional State Theory, *Journal of Computational Linguistics*, 12:3, 175-204, The MIT Press Cambridge, MA, USA.

Hyland, K. (2009). Academic Discourse: English In A Global Context (Continuum Discourse).

Holt, R. (2004). Dialogue on the Internet. Language, Civic Identity, and Computer-Mediated Communication, Westport, Conn., PRAEGER.

Lin, J. (2007). Automatic Author Profiling of Online Chat Logs, M.S. Thesis, Naval Postgra duate School, Monterey.

Loh, C. (2009). Une ancienne députée de Hong Kong sur la Toile: le site «Civic Exchange» (entretien avec Éric Sautedé), in HERMÈS, (55).

Kerbrat-Orecchioni, C. (1999). L'énonciation : De la subjectivité dans le langage. 4-ème édition. Paris: Armand Colin.

Maingueneau, D. (2000). Analyser les textes de communication. Paris: Nathan.

Manchala, D.W. (1998). Trust metrics, models and protocols for electronic commerce transactions, *Proceedings of the 18th International Conference on Distributed Computing Systems*.

Mann, W.C., Thompson, S.A. (1988). Rhetorical Structure Theory. Toward a functional theory of text organization, *Text - Interdisciplinary Journal for the Study of Discourse*. 8: 3, 243–281.

Mortensen, T. E. (2003). Pleasures of the player. Flow and control in online games, Volda University College.

Mui, L., Mohtashemi, M., Ang, C., Szolovits, P., Halberstadt, A. (2001). Ratings in distributed systems: a Bayesian approach, *Proceedings of the Workshop on Information Technologies and Systems* (WITS).

Nurmi, P. (2006). A Bayesian framework for online reputation systems, *Proceedings of the Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services*.

Pennebaker, J. W., Francis, Martha E., Booth, R. J. (2001). Linquistic Inquiry and Word Count – LIWC2001, Mahwah, NJ, Erlbaum Publishers.

Plantin, C. (2005). L'Argumentation, PUF, Que sais-je?

Pragmatics, (2006)/(2011). Metaphysics Research Lab, CSLI, Stanford University.

Oliveri, N. (2011). La cyberdépendance: un objet pour les sciences de l'information et de la communication, in HERMÈS (59).

Sabater, J., Sierra, C. (2002). Social ReGreT, a reputation model based, on social relations, SIGecom Exchanges 3.1.

Schiaffino, S., Amandi, A. (2009). Intelligent user profiling, *Artificial intelligence*, Lecture Notes In Computer Science, Vol. 5640. Springer-Verlag, Berlin, Heidelberg (Max Bramer ed.).

Stark, A., Dürscheid C., (2011). SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland, *Crispin Thurlow/Kristine Mroczek (Hrsg.): Digital Discourse. Language in the New Media*. Oxford: Oxford University Press.

Stoica, D. (2001). Modalités de la communication scientifique, în NOESIS. Travaux du Comité Roumain d'Histoire et de Philosophie des Sciences, vol. XXVI. Bucureşti, Editura Academiei Române.

Trandabăţ, D. Irimia, E., Barbu Mititelu, V., Cristea, D., Tufiş, D. (2012). The Romanian Language In The Digital Age. META-NET White Paper Series, Springer.

Tuţescu. M. (1998). L'argumentation. Introduction a l'étude du discours, Bucureşti, Ed. Universităţii.

Zukerman, I. and Albrecht, D. (2001). Predictive Statistical Models for User Modeling. User Modeling and User-Adapted Interaction, 11(1-2).

Weiwei, Y., Donghai, G., Sungyoung, L., Young-Koo, L., Heejo, L. (2007). Bayesian Memory-Based Reputation System, in Proceedings of the 3rd international conference on Mobile multimedia communications.

Withby, A., Josang, A., Indulska, J. (2005). Filtering Out Unfair Ratings in Bayesian Reputation Systems, Journal of Management Research.

CHAPTER 3
SPEECH PROCESSING

ROMANIAN CORPUS FOR SPEECH-TO-TEXT ALIGNMENT

ANCA – DIANA BIBIRI[1], DAN CRISTEA[2,3], LAURA PISTOL[2,3],
LIVIU – ANDREI SCUTELNICU[2,3], ADRIAN TURCULEȚ[1]

[1] *"Al. I. Cuza" University, Department of Interdisciplinary Research in Social-Human Sciences, Iaşi – Romania*
[2] *"Al. I. Cuza" University, Faculty of Computer Science, Iaşi – Romania*
[3] *Institute of Computer Science, Romanian Academy, Iaşi – Romania*

anca.bibiri@gmail.com, {dcristea, laura.pistol, liviu.scutelnicu}@info.uaic.ro, aturcu@uaic.ro

Abstract

In this paper we present the methodology employed in the creation of an aligned speech-to-text Romanian Corpus. The corpus uses recordings from the AMPER-ROM and AMPRom projects as well as ad-hoc recordings of continuous speech. The protocol for speech recording and labelling, as well as the manual annotation procedure, are described. The corpus is intended to be used for training a speech segmentation module and an automatic speech-to-text aligner module.

Keywords: Corpus, speech-to-text, alignment, PRAAT

1. Introduction

Since the early days of intonation research, automatic transcription of the intonation in speech corpora has been on the wish list of many researchers in phonetics, linguistics, and discourse analysis. For several decades, linguistics has gathered a great amount of audio material to study the aspect of spoken language. Unfortunately, some of the recordings have different dialectal signals/marks, for example, background noise, different phonetic intonation, differences in time of intonation and voice changing, etc.

Alignment of the phonemes and text is the first stage of data processing necessary to provide useable training data for many phoneme-to-text conversion systems, including the most successful symbolic rule-based systems and most neural network systems (Bullinaria, 2011). A common requirement in speech technology is to align two different symbolic representations of the same linguistic message, for instance, phonemes with letters (Damper et al., 2005). As dictionaries become even bigger, manual alignment becomes less and less tenable, yet automatic alignment is a hard problem for a language like Romanian.

In this paper we describe a methodology for building an aligned speech-to-text corpus for Romanian. The investigation has as goal to set the principles of acquiring a significant corpus of signal-text aligned recordings, to be used for training a speech segmenter and a speech-to-text aligner module. By exploiting already existent continuous speech tracks, doubled by their textual transcriptions, an automatic aligner could be used to fabric a large corpus of speeches aligned to their textual transcription, creating thus the prerequisite for training a speech recognition system for Romanian. Other applications of speech-to-text alignment systems are in fields, such as multimedia indexing, training of large vocabularies for speech recognition, health-related research, etc.

2. Corpora

2.1. AMPER-ROM[ANIA]

L'Atlas Multimédia Prosodique de l'Espace Roman (AMPER) is a last generation atlas which combines principles of geolinguistics with techniques of instrumental phonetics and those of informatics. The atlas is conceived as an interactive database bringing together data collection and acoustic analysis concerning prosodic features of linguistic varieties specific to the Romance languages.

The Romanian Multimedia Prosodic Atlas (AMPRom) is the first prosodic atlas which aims to present the main intonation patterns of the Romanian language varieties identified both at the level of the diatopic variants of the standard language and at the level of the dialect variants.

During the prosodic dialectal investigations, two questionnaires are used: AMPER-ROM[ÂNIA] and AMPRom. The first questionnaire consists of a series of statements (45 sentences) established by morpho-syntactic and phonetic criteria and are formed of: *declarative- affirmative* and *declarative-negative* sentences and total *interogative-affirmative* and *interogative-negative* sentences, having the syntactic structure SVO (subject – verb – object). The S and O receive, in turns, adjective and/or prepositional determinants; the nouns and adjectives that are used in the utterances are trisyllabic oxitones (the last syllable of the word is stressed), paroxitones (the penultimate syllable of the word is stressed) and proparoxitones (the antepenultimate syllable of the word is stressed). Since in the Romanian language the negation usually receives the stress of the phrase, the negative-declarative and interrogative-negative sentences were also introduced in the questionnaire.

The occurrences of the words are at the right and at the left of the verb for capturing all the prosodic indices (S – subject, V – verb, O – object, Adj – adjective – with the mention that the subject is interchangeable with the object):

[S + V + O / S + Adj/ + V + O / S + V + O + Adj / S + S + V + O / S + V + O + S]

AMPER-ROM questionnaire (sequence) (Each sentence is labeled with a unic code in order to identify the sentence when the acoustic analysis is made: *bwt, dwk, fwt, gwt, kwt, pwt, swk, twg, twk, zwk.*):

twk *Nevasta vede un căpitan.*/ *The wife sees a captain.*
kwt *Un căpitan vede nevasta.*/ *A captain sees the wife.*

dwk *Nevasta tinerea vede un căpitan.*/ *The young wife sees a captain.*
gwt *Un căpitan elegant vede nevasta.*/ *An elegant captain sees the wife.*

swk *Nevasta frumoasă vede un căpitan.*/ *The beautiful wife sees a captain.*
pwt *Pasărea vede nevasta.*/ *The bird sees the wife.*

zwk *Nevasta harnică vede un căpitan.*/ *The hardworking wife sees a captain.*
bwt *Pasărea papagal vede nevasta.*/ *The parrot bird sees the wife.*

twg *Nevasta vede un căpitan elegant.*/ *The wife sees an elegant captain.*
fwt *Pasărea frumoasă vede nevasta.*/ *The beautiful bird sees the wife.*

There are in AMPER-ROM questionnaire sentences with broad focus, as in the following examples. The labels of the sentences represent: *twkae1* – the declarative affirmative sentence with the focus on the first element – subject; *twkie2* – the interrogative affirmative sentence were the object is stressed; *twknev* – the declarative negative sentence with focus on the verb.

twkae1 ***Nevasta*** *vede un căpitan.*/ ***The wife*** *sees a captain.*
twkie2 *Nevasta vede* ***un căpitan****?*/ *The wife sees* ***a captain****?*
twknev *Nevasta* ***nu vede*** *un căpitan.*/ *The wife* ***does not see*** *a captain.*

2.2. AMPRom

In order to capture a larger number of Romanian intonation patterns in their territorial distribution, a second questionnaire includes other statements, simpler (with not so many formal constraints) to facilitate the contact with the subjects and to prepare them for the fixed questionnaire. This includes about 100 sentences and has two variants: short version (compulsory, with 84 sentences) and extended version (optional, having 111 sentences), the latter is applied only at some points of inquiry.
Types of syntactic structures that make up the AMPRom questionnaire:

- VO structures (with inclusive subject): 1a: *L-ai văzut pe Ion?/ Have [you] seen John?* 3a: *Ai văzut fetele?/ Have [you] seen the girls?*
- Structures pursuing the relation between the word order and prosody: (1) 1b: *Pe Ion l-ai văzut?/ John was that you have seen?* 3b: *Fetele le-ai văzut?/ Girls were that you have seen?*
- VS/SV Structures: 25a: *Vine Ion./ There comes John.* 25b: *Ion vine./John is coming.* 28a: *Cine vine?/ Who is coming?* 28b: *Ion vine./John is coming.*
- Structures with double negation elements both in the question and in the answer: (26): *Nu vine nime(ni) la noi?/ There comes There comes nobody/none to us?* (30): *N-a venit nime(ni) la noi./Nobody/none came to us.*
-

- Structures in which modulators are used (adverbs of manner and semi-adverbs – *sure, precisely, certainly, immediately, surely, maybe, whether, really* or even modal verbs – *I think, it might*): 20b: *Chiar vine Ion?/ Really, is John coming?* 21a: *Sigur/Precis (că) vine/ Sure/precisely he is coming.* 23c: *Cred că vine./ I think he is coming.*
- Structures containing different types of questions: partial, alternative, confirmation: 56a: *Cât e ceasul?/ What time is it?* 41: *Vii ori nu vii?/ Are you coming or not?* 55b: *Pleci mâine la Iaşi, nu-i aşa?/ You are going tomorrow to Iaşi, aren't you?*
- Structures containing vocative addressing and calling: 40: *Ion (Ioane), dă-mi un măr (te rog)!* / *Ion (John), give me an apple (please)!* 35a: *Ana!/ Ann!*, 35b: *Maria!/ Mary!*,
- Structures that require an intonation of continuity (in suspension): 49: – *Apucă-te/Ia şi-nvaţă, că de nu.../ Start/ Put yourself at work/to learn, or else...*
- Exclamatory structures: 84: *Ce batic frumos ai!/ What a beautiful scarf [you] have!*
- Structures on intercalation prosody: 74a: *Tata mi-a zis: Du-te repede şi cheam-o pe soră-ta! / My father said, 'Go quickly and call your sister'!* 74b: *Du-te repede şi cheam-o pe soră-ta! mi-a zis tata. / Go quickly and call your sister! my father said.*
- Structures containing enumerations: 66: *Am fost la piaţă/târg şi am cumpărat: roșii, ceapă, morcov și ardei./ I was at the market/fair and bought tomatoes, onions, carrots and peppers.*
- Structures containing a sequence of short sentences: 79: *De dimineață m-am trezit, am pregătit micul dejun și apoi am plecat la serviciu./ This morning I woke up, I made breakfast and then I went to work.*
- Sentences with the same structure (V) for the affirmative, interrogative and imperative mood: 80: *Aşteaptă./[He/she]waits.* 81: *Aşteaptă?/Does [he/she] wait?* 82: *Aşteaptă!/Wait!/ Aşteaptă-mă!/Wait for me!*
- Structures with a focus on different constituents 4a: Pe **Vasile** l-ai văzut ?/Was **Basil** that you saw? 4b L-ai văzut pe **Vasile**?/ Did you see **Basil**?; 58: Bei **vin**?/Are you drinking **wine**?
- Structures with a successive focus on constituents 64: ***Mănânci*** *peşte?/ Are you* ***eating*** *fish?* 65a: *Mănânci* ***peşte****?/ Are you eating* ***fish****?*
- Affective structures: 56f: *E/îi amiază?* / *Is it/ It's noon? It's already noon?* 59: ***Bei vin?/Are you drinking wine?***
- The extended form of the questionnaire contains other type of syntactic structures:
- Structures pursuing the prosody of idioms and phrases: 89 a, b, c…: *da de unde!/what? no way!; nu mai spune!/ yah, do not say!; ce folos?/ so, what?; nici vorba/pomeneala!/no way!/not at all!; cum/unde să facă ea aşa ceva?/what/how did she do that?; da mai ştii?/that could be?;ei şi?/so, what?.*
- Structures containing greetings and politeness: 91: *Bună ziua! Good afternoon!*; 97: *Poftim/There you go!/ Na!/Here! – Mulţumesc/mulţam!/ Thank you/Thanks!*
- Structures that use adverbs and adverbial phrases to strengthen the assertion and negation: 104: *Da, sigur/ firește/ negreșit!/ Yes, sure/ surely/ no doubt!* 105: *Nicidecum!/No way! Niciodată/Never! Nici în ruptul capului!/On no account!*

- Imprecations: 107 a, b, c…: *Arde-l-ar focu' să-l ardă!/ May he burn in hell! Lua-l-ar naiba/dracu să-l ia!/ The hell/the devil with him! Fir-ar/fi-o-ar a dracului!/ Damn it/Damn with it! – Du-te dracului/la dracu/la satana!/ Go to the devil/to Satan!*

The statements are recorded at least three times and are obtained through indirect questions and by verbal and non-verbal implications (facial expressions, gestures) to the context, and/or forming some speech situations during the continuous dialogue between the investigator and the informant.

In rural areas, two indigenous subjects are used, representative for the local speech, with elementary education, middle-aged, who speak natural under the conditions of the investigation. In urban areas the surveys are twofold: besides the informants belonging to low and/or middle class with influences of the local dialect, there are used subjects with higher education, speaking a cultivated language.

2.3. The IIT corpus

The IIT continuous speech corpus consists of recordings, summing up 45 minutes of continuous speech, uttered in an office environment and following a standard voice recording procedure, by three female speakers who currently speak Romanian standard language, aged between 33 and 50, having no pathological disorder and originated from the geographical area of North-Western Romania (the Iași district). The recordings were single channelled with a sampling frequency of 22050 Hz and 16 bit resolution. The sentences chosen for recordings are paragraphs from "Amintiri din copilărie" (*Childhood Memories*), by the classical Romanian writer Ion Creangă and dialogues from sketches by the Romanian writer and dramatist Ion Luca Caragiale. The choice towards this piece of classical belletrist work was imposed by the necessity for the corpus to be copyright-free.

The size of the IIT database is shown in the following table:

Table 1: Size of the database (Only for the writer Ion Creangă)

sentences	341
vocabulary size	2000
words (occurrences)	6505
words per sentence	19.07

3. Notation of sounds, phonemes, graphemes

In the following, by *sound* we mean a segment of a speech track, as it is heard by a human or is recorded by a machine. A sound, in general, is characterised by steady physical parameters (amplitude, frequency) and corresponds to a letter in an alphabetic transcription. There is a huge variance of sounds corresponding to the same letter, depending on the articulatory and the co-articulations conditions of the sounds and to other factors, such as the context of communication and the speaker (sex, age, tonality, momentary physical and psychological state).

A *phoneme* is the conceptualization of a sound. The Romanian language has 31 phonemes. As such, one cannot say that phonemes are recorded. Only sounds can be recorded, but out of them, phonemes are deciphered (interpreted) and, accordingly, noted. In the real world, a phoneme does not exist, but we can say "this sound records the phoneme *a*". The phonemes are noted in the International Phonetic Alphabet – IPA (see below).

The *speech-to-text* alignment conventions are based on the mappings between the two planes of expression of language: the concrete plane (of the substance of the language), populated with sounds, and the abstract plane (of the form, the linguistic plane), where phonemes coexist. These two planes are both doubled by two levels of expression: phonic and graphic (as suggested by Figure 1).

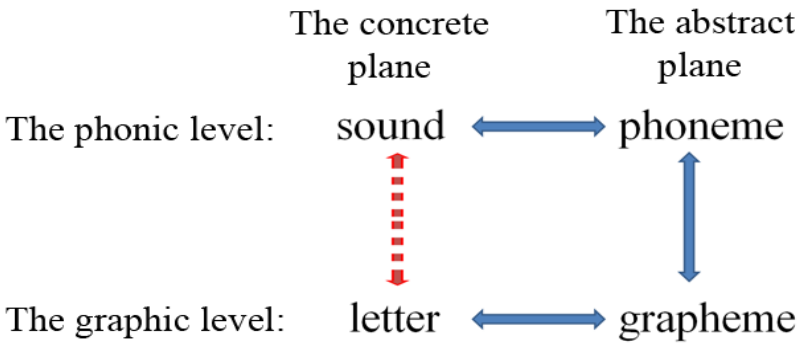


Figure 1: The speech to text correspondence

Although a phonemic language (sounds as they are transcribed), Romanian has some particularities:

- the sounds z and s in *dezbat* vs. *desfac*, or *răzbate* vs. *răsplati* have as a variant /S/: /deSbat/, /rəSbate/; *îmbrac* and *învăţ* have as a variant /N/: /ɨNbrak/, /ɨNvətz/; as a result of neutralization of the opposed /z/ and /s/, respectively, between /n/ and /m/ in such examples it is noticed the occurrences of the archiphonemes /S/ and /N/.

- the use of morphematic principle in order to maintain the formal identity of the words, especially when one speaks about the alternation of the diphthongs *oa* and *ua*, respectively *ea* si *ia*: *oa ~ o*: *oameni – om, toată – tot , oală – ol; ua ~ u: băcăuan – Bacău, flăcăuaş – flăcău*, and in the case of some neologisms: acuarelă, scuar; *ea ~ e*: *teamă – tem, cheamă – chem, ceas – cesuleț, ea – ele*; *ia ~ ie*: *iarnă – ierni, piatră – pietre* or in the situation when there is no alternative: *chiar, ghiaur*;

- the morphematic principle is rarely used to differentiate morphemes: *aceea(şi)* vs. *aceiaşi, ea* vs. *ia*;

- it is maintained (totally or partially) the etymological spelling: *eu, el, ei, ele, eram*; *absent, lied, watt, subţire, fotbal; alură, bleu*; the most typical case is that of loans from English: *computer, laptop, site, whisky, weekend*.

Romanian spelling includes graphemes created using diacritical marks (because of the lack of specific letters in the Latin alphabet: ă, â, î, ş, ţ) as well as polyvalent, compound graphemes having different contextual values.

There are polyvalent vocalic graphemes <e>, <i>, <o>, <u>, noting both the vowels /e/, /i/, /o/, /u/, and the corresponding semivowels /e̯/, /i̯/, /o̯/, /u̯/; also the sequence of a vowel + a dependent semivowel: <e> = [i̯e] in *eu, eram, vie*; <i> = [i̯i] in *cais, fiinţă, oişte* or [ii̯] in *academia, ia* ‘bluză’; <o> = [u̯o] in *fior*, [ou̯] in *merituos*; <u> = [u̯u] in *aur* or [uu̯] in *luând*. In some cases, according to the morphematic principle, the graphemes <e>, <o> also note semivowels <i>, <u>: *aceea, ea, oameni, vioară*.

The consonantic graphemes <c>, <g>, <k>, <n>, <x> have double values depending on the context where they occur: /k/ and /tʃ/, /g/ and /ɟ/, /k/ and /c/, /n/ and /N/, /ks/ and /gz/ in *car* and *cer*, *gară* and *ger*, *kaliu* and *kaki*, *nas* and *învăţ*, *aks* and *exemplu*.

There are graphemes compound of two or three letters: <ce>, <ci>, <ge>, <gi>, <ch>, <gh>, <che>, <chi>, <ghe>, <ghi>: *ceas, arici, geam, ungi, chem, ghem, cheamă, gheaţă, ochi, unchi, unghi*.

The description of the phonemic system of the Romanian language has several interpretations, with different numbers of phonemes, depending on the authors’ theoretical and methodological assumptions. The Romanian linguist E. Petrovici (1956) proposes in his phonemic theory the largest number of phonemes: 5/7 vowels and 72 consonants, and E. Vasiliu (1965) – the smallest number of phonemes: 7 vowels, 20 consonants, and one special phoneme called ’syllabic juncture’.

For our corpus we propose a simple phonemic system, which best corresponds to Romanian writing, in accordance with the Latin alphabet.

This phonemic system (Turculeț, 1999) is made up of 7 vowels: /e/, /i/, /a/, /ə/, /ɨ/, /o/, /u/, 4 semivowels /e̯/, /i̯/, /o̯/, /u̯/ and 20 consonants ([c], [ɟ] are considered allophones of the phonemes /k, g/) – see Table 2.

Table 2: Symbols for consonants

Place→ **↓Manner**	**Bilabial**	**Labio-dental**	**Dental-alveolar**	**Alveolar**	**Alveolo-palatal**	**Velar**	**Glottal**
Plosive	/p/ /b/		/t/ /d/			/k/ /g/	
Nasal plosive	/m/		/n/				
Fricative		/f/ /v/	/s/ /z/		/ʃ/ /ʒ/		/h/
Affricate			/ʦ/		/ʧ/ /ʤ/		
Lateral				/l/			
Trill				/r/			

The reduced vowel i̯, asyllabic and voiceless, specific to the Romanian language called' final, asyllabic, post-consonant i' such as in [lupĭ], [potzĭ] (it occurs rarely within a compound word, at the morpheme limit [orĭkɨnd], [kɨtzĭva]) as a variant of semivowel /i̯/. Thus, the phonetic label is [ĭ] and the phonematic one as /i̯/ (it occurs after a consonant in the final position and between two consonants in medial position).

The back rounded vowels [ö] and [ü] originated in some French and German loans can be considered as situated at the Romanian phonetic and phonemic periphery: <alură> [alürə], <bleu> [blö], <röntgen/roentgen> [röntɟen] or [röntgən]. They are realised usually as the diphthongs [i̯u], respectively [e̯o].
Regarding the correspondence between phonemes and graphemes we propose some simple solutions according to the combinations of Romanian letters used in writing. They concern the evaluation of compound graphemes (see *supra*). The compound graphemes from the following examples <*ceas*> [tʃas], <*arici*> [aritʃ], <*geam*> [dʒam], <*ungi*> [undʒ] are reduced to simple graphemes <c>, <g>, followed by the 'latent' phonemes /e̯/ and /i̯/ (possible solution proposed in generative phonology) with the phonemic transcription / tʃe̯as/, / aritʃi̯/, /dʒe̯am/, / undʒi̯/, and the trigraphs <*che*>, <*chi*>, <*ghe*>, <*ghi*> followed by vowels <*a*>, <*o*>, <*u*> or in final position are reduced at <*ch*>, <*gh*> : <*cheamă*> [camə], <*chiar*> [car], <*chior*> [cor], <*chiul*> [cul], <*gheaţă*> [ɟatzə], <*ghiozdan*> [ɟozdan], <*ochi*> [oc], <*unghi*> [uɲɟ], with the phonemic transcription /ce̯amə/, /ci̯ar/, /ci̯or/, /ci̯ul/, /ɟi̯atzə/, /ɟi̯ozdan/, /oci̯ /, /unɟi̯/.
The compound graphemes are, in fact, the digraphs <*ch*> and <*gh*> as in <*chem*> [cem] /cem/, <*ghem*> [ɟem] /ɟem/, <*cheamă*> [camə] /ce̯amə/, <*gheaţă*> [ɟatzə] /ɟe̯atzə/, <*ochi*> [oc] /oci̯/, <*unchi*> [uɲc] /unci̯/, <*unghi*> [uɲɟ] /unci̯/.

Figure 2 shows an example of a speech-to-text alignment: partial interrogative sentence uttered by a subject from Bucharest (Cristina Dăbuleanu, 49 years old, computer programmer): *Cum te cheamă?* (*What is your name?*).

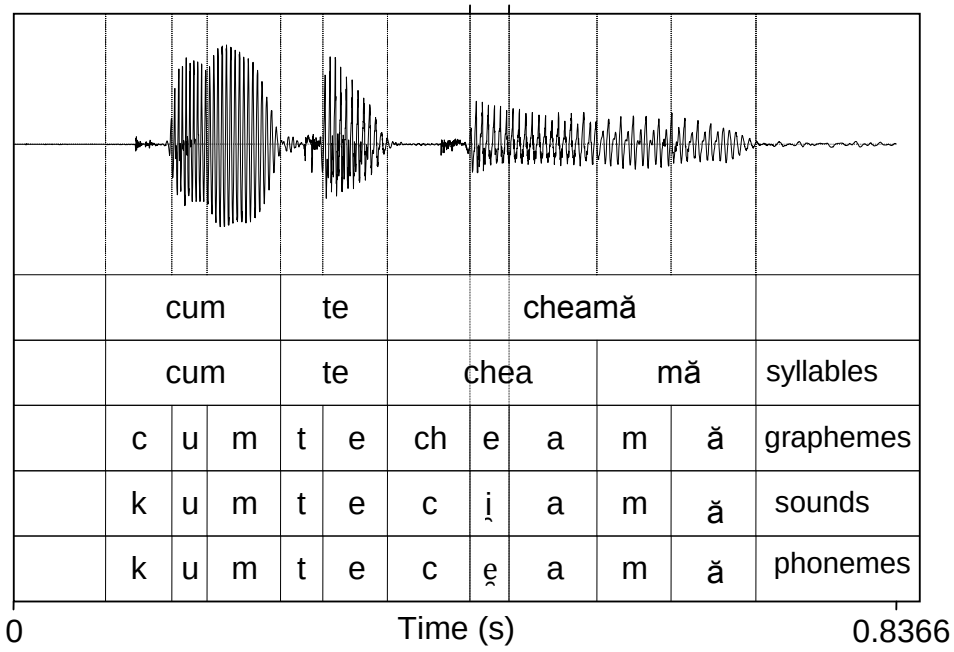


Figure 2: Praat screen in the speech-to-text alignment of the utterance *Cum te cheamă?*

For some loans (most of them from English), there are applied the rules for writing and speaking of foreign language, will be marked with a special sign. The letters/graphemes and the sounds/phonemes will be maintained as they are in the foreign language: <*laptop*> [læptop], <*site* > [sajt].

4. Speech-to-text alignment

The purpose of the manual speech-to-text alignment is to determine with precision the boundaries of sounds belonging to the phonic layer and to align them with letters from the grapheme layer. The task is done by one of the co-authors, having an extensive experience in reading spectrograms and labelling phonemes. By using the graphical interface and listening the audible track in Praat, she identifies the acoustic changes in order to determine the phoneme boundaries. The annotation levels are: utterance, word, syllable, phoneme and grapheme. Table 3 shows the notations used with Praat in the alignment process.

Table 3: 9 tracks revealed by Praat, shown at different moments of time; apart from duration, the first 3 tracks (sound, syllable and word) represent manual annotation, while the other 5 are automatically recorded

Time	Duration	Sound	Syllable	Word	Intensity(dB)	F0 (Hz)	F1	F2	F3 (Hz)
0	0.09	n	ne	nevasta	74	0	390	1768	3096
0.09	0.05	e/\ef	ne	nevasta	73	205	463	1835	3088
0.14	0.06	v	\'1vas	nevasta	70	0	567	1484	2220
0.2	0.12	a::	\'1vas	nevasta	78	237	807	1523	3187
0.32	0.06	s	\'1vas	nevasta	61	0	806	1728	3208
0.38	0.06	t	ta	nevasta	75	0	621	1484	2965
0.44	0.06	a	ta	nevasta	79	228	875	1529	3092
0.5	0.08	v	\'1ve	vede-un	63	0	648	1375	2780
0.58	0.06	e	\'1ve	vede-un	77	230	497	2252	2927
0.64	0.07	d	di-un	vede-un	70	0	371	2291	2958
0.71	0.08	i\nvu\~^	di-un	vede-un	70	236	399	1087	2776
0.79	0.04	\ng	di-un	vede-un	71	0	452	1010	2593
0.83	0.06	k	c\sw	c\swpitan	63	0	257	1634	2912
0.89	0.04	\sw	c\sw	c\swpitan	76	223	527	1528	2827
0.93	0.09	p	pi	c\swpitan	53	0	464	1903	3146
1.02	0.04	i	pi	c\swpitan	69	212	389	2504	3353
1.06	0.1	t	\'1tan	c\swpitan	53	0	301	2541	3395
1.16	0.09	a\~^\~v:	\'1tan	c\swpitan	66	146	942	1746	3229
1.25	0.06	n	\'1tan	c\swpitan	62	0	1039	3137	3568

PRAAT is a flexible tool for the analysis of acoustic speech signals. It offers a wide range of standard and non-standard procedures, including spectrographic analysis, articulatory synthesis and neural network. Speech segmentation is the process of identification of boundaries between words, syllables and phonemes. Performed manually, this process attaches a label to each segment. For example, after we have finished segmenting the words and labelled them, follows the segmentation of the syllables of the structure and, finally, those of the compound sounds. The steps in the analysis of a speech waveform are as follows:

- The script reads sound files (.wav format – *Waveform Audio File Format*) from a user-specified folder;
- Then create a TextGrid (which consists of a number of tiers – an interval tier is a connected sequence of labelled intervals, with boundaries in between);
- Selecting both .wav and Text Grid files it opens a window spectrogram in which the annotation is made manually: 3 tiers are open in order to annotate words, syllables and phonemes;
- Once the speech signal is segmented and labelled, by pushing the run button a text file is generated in output, including different parameters: the fundamental frequencies (F0, in the three points of a vowel – F1, F2 and F3), the duration and the intensity of the acoustic signal.

For the speech-to-text alignment of the corpus, the supra-segmental features of the utterance are also taken in consideration: the stress, the intonation and the break indices (as indicated by punctuation marks). A more appropriate rendering is that used in ToBI[1] – a framework for developing community-wide conventions for transcribing the

[1] Tones and Break Indices: http://www.cs.indiana.edu/~port/teach/306/tobi.summary.html

intonation and prosodic structures of spoken utterances in a language variety. A ToBI framework system for a language variety is grounded on the intonation system and the relationship between intonation and the prosodic structures of the language.

5. Conclusions

In this paper we presented a methodology of manual annotation of an aligned speech-to-text corpus for Romanian, and the phonetic peculiarities of this language. The intention is to use this corpus to train a speech segmentation and aligner program (let's call it a SEG-ALI module) that would be able to detect the boundaries of sounds in correlation with a text track where the textual transcription is noted. Different parameters of the speech signal, some of them having been suggested in this paper by presenting the processing capabilities of the Praat system, will be exploited by a learning system that will finally train the SEG-ALI module. A top-down strategy will, most probably, be employed for this purpose, by searching first the pauses in the sound track and aligning them with the boundaries between sentences and words and using more high level features to detect phonemes boundaries in between pauses of the continuous speech.

Once such a SEG-ALI module is obtained, it could be used to segment and align automatically a very large corpus of parallel tracks containing human produced continuous speech and their textual transcription. In the long run, the intention is to acquire a large corpus of aligned speech-to-text records that will be used in training a speech recognition system for Romanian. Knowing the high costs encumbered by manual segmentation of the voice track and its alignment against the text track, our hope is to arrive at a very good performance of the SEG-ALI module that would permit the automatic acquisition of a very large corpus in a short time and with reduced costs.
We do not neglect also the possibility to use a boot-strapping strategy in acquiring a high quality aligned corpus: use the manually annotated corpus as a core corpus on which a beta version (v0) of a SEG-ALI module is first trained. Use then this SEG-ALI-v0 to segment&align a larger corpus, and then involve specialised humans to correct it. This activity is supposed to take less time than building it from scratch and also cost less. Once finished, use this larger corpus to retrain the SEG-ALI module to a new and enhanced version – v1, and so on.

References

AMPER – Atlas Multimédia Prosodiques de l'Espace Roman, http://w3.u-grenoble3.fr/dialecto/AMPER/amper.html
AMPROM – http://amprom.uaic.ro/

Handbook of the International Phonetic Alphabet. A Guide to the use of the International Phonetic Alphabet (1999). Cambridge University Press.
Boersma, P., Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.42, retrieved 8 February 2013 from http://www.praat.org/

Bullinaria, John A. (2011). Text to Phoneme Alignment and Mapping for Speech Technology: A Neural Networks Approach, IJCNN, IEEE, 625-632.

Damper, R. I., Marchand, Y., Marsters, J.-D. S., Bazin, A. I. (2005). Aligning text and phonemes for speech technology applications using an EM-like algorithm. *International Journal of Speech Technology*, no. 8, 149-162.

Hosom, J.-P. (2009). Speaker-independent phoneme alignment using transition-dependent states, *Speech Communication*, no. 51, 352-368.

Petrovici, E. (1956). Sistemul fonematic al limbii române (*The phonemic system of the Romanian language*), in Studii și Cercetări Lingvistice, VII :1-2, 7-21.

Turculeţ, A. (1999). Introducere în fonetica generală şi românească (*Introduction to general and Romanian phonetics*), Demiurg Editorial House, Iaşi.

Vasiliu, E. (1965). Fonologia limbii române (*The phonology of the Romanian language*), Editura Ştiinţifică, Bucharest.

DATA-DRIVEN METHODS FOR PHONETIC TRANSCRIPTION OF OUT-OF-VOCABULARY (OOV) WORDS

TIBERIU BOROȘ[1], RADU ION[1], DAN ȘTEFĂNESCU[2]

[1]*Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest, Romania*
[2] *University of Memphis, USA*

{tibi,radu}@racai.ro, {dstfnscu}@memphis.edu

Abstract

Letter to Phoneme conversion (L2P) is a crucial problem in any modern text-to-speech (TTS) synthesis system. The L2P conversion is routinely done with the help of a lexicon. An inherent problem of this approach is that regardless of the size of the lexicon, there will always be out of vocabulary (OOV) words, for which a method for automatic phonetic transcription is required. In this paper we present our L2P system which uses a set of 4 methods for obtaining phonetic transcriptions for OOV words. We compare our results with current existing state of the art methods showing that our system is up to par.

Keywords: letter-to-phoneme conversion, text-to-speech synthesis, out of vocabulary words.

1. Introduction

Predicting pronunciation for OOV words is a major challenge for any TTS system. While sometimes this can be a simpler task for certain languages where there is a clear relationship between letters and their phonetic transcription (e.g. Romanian, which has a preponderantly phonemic orthography), for others, such as English, it may pose considerable difficulties. Consequently, phonetic transcription is a key component in every TTS system, but this is not the only appliance of it. Other tasks, like spelling correction, can be addressed using phonetic transcription by means of phonetic similarity and perceptive search.

L2P conversion usually means detecting a set of language-dependent rules that will map letters to phonemes. These rules may be written by linguists or automatically inferred from a given list of word/phonetic transcription pairs. Phonetic transcription is the next step, where possible L2P rules are applied to the OOV word's written form and the best phonetic transcription is selected according to an optimum criterion.

Various scientific studies have focused on automatically extracting L2P conversion rules from available *hand-made* transcriptions (Black et al., 1998; Jiampojamarn et al., 2008; Paget et al. 1998). At this point, we should note that phonetic transcription of OOV words does not address other phonetic transcription problems such as homograph

disambiguation. OOV words are not included in the training lexicon and it is impossible to infer that these words have multiple pronunciations depending on their senses.

In this paper we describe Bermuda, a system for automatic phonetic transcription of words, starting from the alignment provided by *GIZA++* (Och and Ney, 2003). Bermuda combines a set of four different data-driven methods which will be detailed later in this paper. Our entire training process is automatic: there is no need for manual intervention in finding alignments between words and phonetic transcriptions. We intend to extend this system to include other state of the art methods for automatic phonetic transcription.

2. Related work

Phonetic transcription is an area of active research, which produced a multitude of solutions, mostly based on machine learning (ML) methods. Basically, their objective is to generate sequences of phonemes (phonetic transcription) from sequences of letters (words).
Divay and Vitale (1997) presented a L2P method that used a large number of context-sensitive and context-free rules with a minimum number of ordering constraints for phonetic transcription of words. Another approach was to use part-of-speech (POS) tagging methods (Hidden Markov Models) and to treat each individual letter independently as if it were a word in a sentence that required POS tagging (Taylor, 2005). This method did not yield high accuracy results. Later, it was shown that better results could be obtained by pairing letter substrings with phoneme substrings (Bisani & Ney, 2002; Marchand & Damper, 2000; Jiampojamarn et al., 2008), instead of treating each letter individually. The reason resides in the fact that phonetic transcriptions are context dependent -- the next phoneme in line dependents on the *current and previous letters* -- and reportedly, also on the *next letters* (Demberg, 2007). Multinomial classifiers have also been used to predict phonetic transcription based on features extracted from letters and groups of letters inside words (Black et al., 1998; Jiampojamarn et al., 2008; Paget et al. 1998).

All ML methods require training data but obtaining such a corpus is not straightforward. Lexicons usually contain words with associated phonetic transcriptions. However, the relationship between letters and phonemes is not always a one-to-one relationship. For example, not all words have the same number of letters as the number of phonemes in their phonetic transcription (e.g. feared: F IH R D) and, even if the number of phonemes is equal to the number of letters, this does not necessarily imply that only one-to-one alignments exist between them (e.g. experience: IH K S P IH R IY AH N S; the letter x spawns two phonemes 'K'+ 'S' and the ending 'e' is silent). This relationship is captured by what is known as L2P alignments. The *Expectation-Maximization (EM)* algorithm and its variants have been used to find one-to-one or many-to-many alignments between letters and phonemes in (Black et al., 1998; Jiampojamarn et al., 2008; Paget et al. 1998). Given the fact that certain pairs of letters and phonemes are much more frequent than others, EM can be employed in order to automatically detect the most probable alignments given a list of pairs of words and their transcriptions as training data.

2.1. System overview

Bermuda's architecture is organized in two layers. The first layer uses two methods for obtaining phonetic transcriptions of words: the first method implements the Dictionary Lookup or Probability Smoothing (DLOPS) algorithm (see section 3.1) and the second method (Phonetic Transcription Classifier - PTC) is based on a MaxEnt classifier (see section 3.2). The second layer is designed to automatically correct systematic failures in the first layer methods. As shown in section 7, chaining the second layer correction method (ERC) to the output of the first layer methods gives an increase in accuracy ranging from 1 to 7%.

DLOPS is a data-driven algorithm used for generating phonetic transcriptions of OOV words by optimally adjoining maximal spans of phonetic transcriptions found in a transcription dictionary, corresponding to adjacent parts of the input word.
The MaxEnt classifier uses features constructed from contextual letters, groups of letters and previously predicted phonemes in order to predict the phonetic transcription of an input word.
Starting from the alignment between letters and phonemes, Bermuda trains the first layer methods, building models for DLOPS and PTC (sections 3.1 and 3.2). Next, the first layer\'s methods are used to predict phonetic transcriptions of words inside the training lexicon. ERC uses features similar to PTC features, supplemented by features extracted from the predicted phonetic transcriptions which, at this step, have become available. Systematic errors in the phonetic transcriptions obtained using the first layer methods are corrected using ERC

2.2. Letter to phoneme alignment

According to Jiampojamarn et al. (2008) the L2P task is characterized by a hidden structure that connects the input set (letters) to the output set (phonemes). Pairing (aligning) the two sets is not a straightforward problem (in section 1 we presented the example of the word experience). Bermuda uses the services of GIZA++ (Och and Ney, 2003) in order to find alignments between the input word segmented at letter level and its composing phonemes. GIZA++ is a free toolkit for generating word alignments in a parallel corpus. In is usually used to create training data for machine translation (MT) systems but, as Rama et al. (2009) showed, it can also be used to pre-process training data for L2P conversion systems.
For each training lexicon we run GIZA++ for a primary letter to phoneme alignment with default parameters (10 iterations of IBM-1, HMM, IBM-3 and IBM-4 models). The available dictionary is split into two files: the first file contains one word per line with its letters separated by spaces, so that GIZA++ will treat them as words in the source language. The second file contains phonetic symbols that "*translate*" the corresponding word on line number n, also separated by spaces (regarded as words in the target language).

3. First layer methods

This section focuses on the first layer methods. We introduce the Dictionary Lookup or Probability Smoothing (DLOPS) algorithm (section 3.1) and we explain how we used the Maximum Entropy classifier to predict the pronunciation of OOV words (section 3.2).

3.1. The DLOPS algorithm

DLOPS is a recursive, *divide and conquer* algorithm. Although its name starts with *Dictionary Lookup*, this does not mean that it tries to retrieve whole words from a dictionary. Instead, it attempts to get the phonetic transcription of a group of letters, either by doing a table lookup or approximating the transcription from smaller contained units. Its primary goal is to predict pronunciation for OOV words, without getting into the problem of disambiguating between heteronyms: words having the same spelling and different pronunciations. This would require additional contextual, semantic or etymologic information about a word and such information is not available for stand-alone OOV words.

The pseudo code for our method is:

Input:

- w[] – vector containing letters of the word
- n – size of vector (number of letters)
- table – hash table containing groups of letters and phonetic transcriptions with probabilities

Output:

- t[] – vector of phonetic transcriptions and their scores

```
1. DLOPS( w[] ) {
2.   if ( exists(table[w]) ) then
3.      return transcriptions from table[w];
4.   else
5.      idx←findPivot(w);
6.      return MergeResults( DLOPS(w[1…idx]), DLOPS(w[idx…n]));
7.   endif
8. }
```

The algorithm performs a dictionary lookup (**line 2**) and if there is a corresponding set of phonemes for the given letter sequence, all possible phonetic transcriptions with their associated probabilities (**line 3**) are returned. If the lookup procedure fails the algorithm seeks an optimal split position in the letter sequence (**line 5**). Once this location is obtained, the phonetic transcription of the given letter sequence is approximated using phonetic transcriptions of the two overlapped substrings (see the next paragraph). Given that the two substrings overlap on the character located at the juncture point, we expect the candidate phonetic transcriptions for the two substrings to also overlap.

The score S of a transcription candidate, composed of two adjoined phoneme sequences S_1 and S_2, is computed using the original transcription probabilities (P_1 and P_2 given letter sequences w[1...idx], w[idx...n]: $P_1=P(S_1|w[1...idx])$; $P_2=P(S_2|w[idx...n])$) of these phoneme sequences and a fusion probability. The fusion probability is a smoothing function applied over a 5 symbols phoneme sequence that is composed of the last two symbols of S1 before the fusion index and the next 3 symbols (equation 2);

$$S = P_1 P_2 \prod_{j=t-k}^{t+k} F_j \qquad (2)$$

P_1, P_2 - emission probability of phoneme sequences S_1 and S_2
t - the fusion index
F_j - N-gram interpolation model for position j using a smoothing function.
K - half of the length of the fusion window (k=2)

FindPivot is a function that maximizes the transcription probability of the first ranking transcription candidate for one or both letter substrings.
We have experimented different functions for estimating the pivot location (line 5 of the pseudo code, *findPivot*; see section 4 for results). For the first test (FP_1 version of *findPivot*) we calculated the position by splitting the word in half. For the second test (FP_2), we tried to detect an index position that would yield the highest score for the first ranking candidate in the transcription probabilities table for the letter sequence found either to the left or to the right. For the third test (FP_3), we looked for an index position that would maximize the transcription score for the first ranking transcription candidate (the highest score after merging overlapped results) for both left and right letter sequences, if they are contained in the transcription table. In case this was not possible, the FP3 version of the *findPivot* function falls back to the FP_2 version.

A runtime example is illustrated in figure 3. The chosen word for L2P conversion is "*absenteeism*" and we explain the execution of our method using the FP_1 pivot function (splits the letter sequence in the middle). This example has an equal execution depth for each node. This does not apply to all cases and some strings generate unequal execution depth for the nodes.

The algorithm has to cope with a couple exceptions. In case there are no transcription candidates that overlap we use the Cartesian product of the all transcription candidates. If the input sequence has the length of 2 and there are no transcription candidates for this letter sequence we split the input string into non-overlapping sequences ("ab" → "a"+"b"). For the algorithm to always return results, the database must contain transcription candidates for every letter in the alphabet of the target language.

Results for each FP function were considered for CMUDICT.06D (English) (CMU, 2011), BRULEX (French) (Content et al., 1990), CELEX (German) (Baayen & Gulikers, 1995) to show how each FP function influence the results. The tests were performed using the 10 fold method.

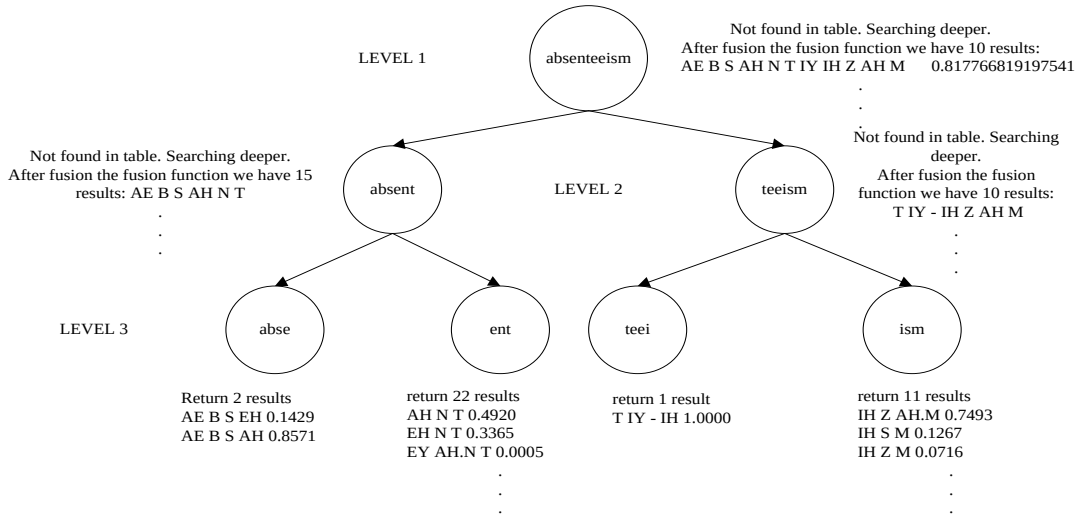


Figure 1: Execution of DLOPS for the word "absenteeism"

We calculated the Mean Reciprocal Rank (MRR) for each FP function we used (Table 2). As shown FP3 function gets the best results, so we used its output when chaining the second layer ERC method. The results obtained using the FP_3 function are very close to those obtained using CART (Black et al., 1998).

$$P = \frac{Number\ of\ correct\ first\ ranking\ suggestions}{Total\ number\ of\ words} \quad (3)$$

$$MRR = \frac{1}{Total\ number\ of\ words} \sum_i \frac{1}{rank_i} \quad (4)$$

where $rank_i$ is the rank of the correct transcription.

Table 1: Experimental results with FP functions

Dictionary	FP_1		FP_2		FP_3	
	P	**MRR**	**P**	**MRR**	**P**	**MRR**
CMUDICT	55.16	72.55	53.21	76.22	57.22	74.54
BRULEX	80.41	93.99	79.63	93.79	80.64	93.28
CELEX-GERMAN	81.41	93.99	80.63	93.79	82.94	93.28

DLOPS training

For DLOPS we extract n-grams up to order 5 from the phonetic transcription symbols by moving a context window and counting occurrences of symbol sequences. Next, we build a model consisting of a set of letters and their possible phonetic transcriptions with corresponding probabilities. We compute transcription probabilities for 1, 2, 3 and 4 letters (Equation 5).

$$P_{i,k} = \frac{C(LS_i, PT_{i,k})}{C(LS_i)} \quad (5)$$

$C(LS_i)$ - number of occurrences of letter sequence i

$C(LS_i, PT_{i,k})$ - number of occurrences of letter sequence i with phonetic transcription k

$P_{i,k}$ - emission probability of phonetic transcription k given the letter sequence i

3.2. Phonetic Transcription with Maximum Entropy

The second method used for the task of PT is based on a Maximum Entropy (MaxEnt) classifier. MaxEnt classifiers have been used in the NLP field to solve problems such as detecting sentence boundaries (Reynar and Ratnaparkhi, 1997; Agarwal, 2005), POS tagging (Ratnaparkhi, 1996), text classification (Nigam et al., 1999), etc.

The guiding principle of MaxEnt is constructing a statistical prediction model from training data without extrapolating for unseen data. The model assumes a uniform distribution for the data, maximizing the entropy. This principle is thoroughly described in Berger et al. (1996).

We apply this powerful methodology to phonetic transcription by employing a publically available MaxEnt classifier: SharpEntropy[1] . In order to do this, we need to reframe the phonetic transcription problem as a label prediction process applied to each letter inside the word. Each letter is now described by an object characterized by a set of *n* features (corresponding to a point inside the ***n***-dimensional feature space).

We experimented with features extracted from a limited context window divided into lexical features (based on letters of the word) and phonetic features (based on previously predicted labels). After testing different feature sets, we chose the one yielding the best results (see Table 4 for an example). For a given letter L, we have the following features:

- l_{-i} L l_{+i} , for i=$\overline{1,2}$: features 1 and 2 in Table 4,
- l_{-i} L , for i=$\overline{1,3}$: features 3 to 5,
- L l_{+i} , for i=$\overline{1,3}$: features 6 to 8,
- p_{-1} : feature 9,

where l_{-i} is the previous i-th letter and l_{+i} is the next i-th letter; p_{-1} is the previous predicted phoneme.

We have tested some other features based on word length or the position of the letter in the word or weather the letter is a vowel or not, etc., but the use of such features did not improve the model.

[1] http://www.codeproject.com/Articles/11090/Maximum-Entropy-Modeling-Using-SharpEntropy

Table 2: The features corresponding to every letter of the word abolish

Letters of *abolish*	Features	Label
abolish	**1**:#ab **2**:##abo **5**:#a **6**:ab **7**:abo **8**:abol **9**:#	AH
a***b***olish	**1**:abo **2**:#abol **4**:#ab **5**:ab **6**:bo **7**:bol **8**:boli **9**:AH	B
ab***o***lish	**1**:bol **2**:aboli **3**:#abo **4**:abo **5**:bo **6**:ol **7**:oli **8**:olis **9**:B	AA
abo***l***ish	**1**:oli **2**:bolis **3**:abol **4**:bol **5**:ol **6**:li **7**:lis **8**:lish **9**:AA	L
abol***i***sh	**1**:lis **2**:olish **3**:boli **4**:oli **5**:li **6**:is **7**:ish **8**:ish# **9**:L	IS
aboli***s***h	**1**:ish **2**:lish# **3**:olis **4**:lis **5**:is **6**:sh **7**:sh# **9**:IS	SH
abolis***h***	**1**:sh# **2**:ish## **3**:lish **4**:ish **5**:sh **6**:h# **9**:SH	-

There are cases when certain features are excluded (see table above). For example, the l_{-3} *L* feature (the 4-gram ending with the given letter) is never used for the first letter of a word mainly because the information it encodes is already contained by the l_{-1} *L* feature. Moreover this feature would be completely indiscriminative in these cases because its value would be identical for all the beginning *L*s.

4. Second layer methods

Systematic errors in predictions for both the DLOPS method and the MaxEnt PT method were noticed, so a second layer method trained to correct these errors was added. This task is also performed by a MaxEnt classifier. Here, we use different features than the ones used in the first MaxEnt classifier. The already predicted labels for all the letters in a word are used to add additional (phonetic based) features. The system is then trained to re-label all the letters inside the word based on the initial prediction and the correct label (according to the training data). This is done in order to assure cohesion at the phonetic level, correcting certain predictions that would be unpronounceable. Thus, when doing error correction we use the following features for a given letter L, having the phonetic transcription P:

- l_{-i} L l_{+i} , for i=$\overline{1,2}$: features 1 and 2 in Table 5,
- l_{-i} L , for i=$\overline{1,3}$: features 3 to 5,
- L l_{+j} , for i=$\overline{1,3}$: features 6 to 8,
- p_{-i} P p_{+i} , for i=$\overline{1,2}$: features 9 and 10,
- p_{-i} P , for i=$\overline{1,3}$: features 11 to 13,
- P p_{+j} , for i=$\overline{1,3}$: features 14 to 16.

where l_{-i} is the previous i-th letter and l_{+i} is the next i-th letter; p_{-i} is the previous i-th predicted phoneme and p_{+i} is the next i-th predicted phoneme.

Since there are two first layer methods, an error correction classifier must be trained for each of these methods. Thus, we end up with two error correction models, each trained to correct the systematic errors of each prediction method. This means there are actually

four ways to perform phonetic transcription: DLOPS, PTC, DLOPS + ERC and PTC + ERC.

5. Comparison to other methods

In this section we compare our method to other approaches in L2P conversion. Only the results obtained using the same dictionaries and similar evaluation methods were taken into consideration. We express the performance of the algorithm in terms of word accuracy rate and we use the first ranking result as the transcription candidate when we calculate the scores (the DLOPS method produces more transcription suggestions with an associated confidence score). We do not use n-best score functions or letter error rates because they do not correctly assess whether this phonetic transcription tool can be used in text-to speech synthesis, where only the first ranking candidate is used for the phonetic transcription of a word.

Table 3 contains the results obtained by our method compared to various other methods (the best results are marked with BOLD): CART Decision Tree System (Black et al., 1998), 1-1 Align, M-M align, HMM: one-one alignments, many-many alignments, HMM with local prediction (Jiampojamarn et al., 2007), Constraint Satisfaction Inference(CSIF) (Bosch & Canisius, 2006), minimum error rate training, A* search decoder (MeR+A*) (Rama et al., 2009), averaged perceptron (Perceptron) and Margin Infused Relaxed Algorithm (MIRA) (Jiampojamarn et al., 2008). The results of the CART method for Romanian are extracted from Stan et al. (2011) (the training corpus is identical to the one we used).

We conducted tests on all dictionaries (except the Romanian one) using the datasets on the Pronalsyl Website. Each dictionary was 10-folded (divided into 10 sets) and the final score was computed as the average of the 10 scores obtained by testing against each set while training on the other 9.

We have to acknowledge that the score obtained for the CMUdict lexicon is lower than expected. This can be explained by the fact that it contains many non-English words which are hard to predict because they do not follow the same phonetic transcription rules as the English ones, a fact also noted in Black et al. (1998) and Jiampojamarn et al. (2007). Such words are practically isolated examples and so, the evidence for inferring phonetic transcription rules for them is practically non-existent. On one hand, if they are to be found in the test set, then there will be no similar examples to learn from in the training data. On the other hand, if they are to be found in the training set, they will merely be a source of noise for the model. Consequently it is practically impossible to predict their pronunciation. Bermuda's score for this lexicon is 68%, being 3% below MIRA's performance and 2% below that of the Perceptron's. On the Netlalk lexicon, Bermuda's PTC+ERC method outranks all other methods by 2%.

The only method outside Bermuda which was applied on the Romanian lexicon is CART (Stan et al., 2011). For this lexicon, Bermuda's best score is 6% higher than CART's accuracy of 87%. Again, on the CELEX and BRULEX lexicons, the PTC+ERC method places third after MIRA and Perceptron.

$$SCORE = \frac{1}{10}\sum_i S_i \qquad (5),$$

where S_i is the score obtained by testing on test set i and training on the other 9.

Table 3 - Experimental results

		CMU dict	**UK BEEP**	**Net Talk**	**BRU LEX**	**CEL EX**	**CEL EX**	**Romanian**
Bermuda	DLOPS FP3	57.00	64.07	53.14	79.17	79.27	78.11	85.74
	PTC	67.22	72.41	68.55	90.99	90.17	90.49	93.29
	DLOPS FP3 +ERC	63.60	67.96	59.70	85.79	86.99	84.89	91.81
	PTC +ERC	68.29	**73.56**	**69.19**	91.68	92.25	91.05	**93.34**
Other methods	Perceptron	71.03	-	64.87	93.89	95.13	92.84	-
	MIRA	**71.99**	-	67.82	**94.51**	**95.32**	**93.61**	-
	CART	57.80	-	-	-	-	89.38	87
	1-1 Align	60.30	-	-	87.00	-	86.60	-
	1-1+CSIF	62.90	-	-	86.50	-	87.50	-
	1-1 HMM	62.10	-	-	88.20	-	87.60	-
	M-M Align	65.10	-	-	90.60	-	91.10	-
	M-M+HMM	65.60	-	-	90.90	-	91.40	-
	MeR+A*	63.81	-	-	86.71	-	90.63	-

Looking on the performance figures, one might consider MIRA a better L2P system. However, the reader should bear in mind the fact that we did NOT perform any pre-processing on the training sets. The letters and phonemes where automatically aligned with GIZA++ and NO supplementary intervention was conducted on the alignments. The other systems have pre-processing steps which include removing heteronyms, words that have no more than four letters or functional words. We did not include such steps in the first set of tests, firstly because we aimed at developing a purely data-driven phonetic transcription system and secondly, because they are very hard to be identically reproduced for an accurate comparison. Still, the above mentioned pre-processing steps can be performed, but we are not recommending this practice since it might lead to unreliable results. For example, removing the words that have less than four letters will considerably reduce the necessary evidence for predicting the phonetic transcription for small words. We also encourage leaving heteronyms inside the training lexicon, because the classifiers will learn consistent rules from their letter-phoneme pairs. Instead of removing words based on letter counts and duplicate entries we recommend filtering out all non-English words as we did in our next experiment.

Using the support of the WordNet (WN) (Miller, 1995) lexical ontology we cleaned up CMUDictionary of all the non-compliant English words. For each entry inside CMUDict we used the WN interface to check if there was a corresponding entry inside the WN. All unknown entries were removed from CMUDict. The purpose for conducting such a trial is obvious: if one wants to train a method for automatic L2S conversion for OOV words on a given language, there is no need to create difficulties by introducing foreign words that do not employ the same phonetic transcription rules as the target language.

After removing foreign words and abbreviations the number of entries in CMU was reduced to 46K words. Some examples of removed entries are:

- **Italian**: braggiotti, castelli, castelluccio
- **German**: aachen, abbenhaus, schlender, schlenker
- **Polish**: zawistowski

Table 4 shows results obtained by our methods on the cleaned CMUDict lexicon, using the same 10-fold validation methodology. There is a clear improvement over the previously obtained results on the unmodified lexicon.

Table 4 – Experiments on the filtered CMUDict

Method	**Result**
DLOPS FP3	60.44
DLOPS FP3 + ERC	69.35
PTC	72.33
PTC + ERC	75.45

6. Conclusions

We thoroughly described the Bermuda system, which implements two methods for data-driven phonetic transcription and a method for error correction. This system can be used within a TTS system for L2P conversion on OOV words, but also for problems like perceptive search and spelling correction. The required training data is freely available on the Internet (downloadable from the Pronalsyl Challenge website – see section 5) for the languages we have tested (English, French, Dutch and German) and it can also be generated from existing resources, if these contain phonetic transcriptions. Also, our tests showed that a cleaner version of CMUDict (using WN) will significantly increase the accuracy of the results (from 68 to 77 percent).

Furthermore, in the case of CMUDict, we have conducted paired t-tests between the runs without error correction and the runs with error correction. Specifically, we have compared DLOPS FP3 with DLOPS FP3+ERC and PTC with PTC+ERC and we can report semnificative increases of the mean word accuracies when using error correction at a significance level \alpha of much less than 0.0001.

It is important to state that comparing to the other existing systems, Bermuda is freely available on-line at *RACAI TOOLS Website*[2]. We offer a downloadable version but also an on-line test version of Bermuda that has be trained using the internal Romanian lexicon, CMUdict and UK BEEP. We also offer International Phonetic Alphabet (IPA) transcriptions for both Romanian and English. In the immediate future, we will add French and German for the online version and we will also include a page for of our perceptive search tool (based on Google APIs).

Moreover, the phonetic transcription module is already used within the Bermuda Voice Synthesizer system, also publically available at RACAI Romanian TTS demo page[3].

The next version of Bermuda will add other current state-of-the art methods for phonetic transcription to its inventory of already implemented techniques. We are also working on a voting system that will increase the accuracy of the tool.

While tweaking some parameters of the DLOPS algorithm (fusion probability function, cut-off factor for the letter-phoneme pairs etc.) we noticed that it can achieve better results on some lexicons. We conclude that always using the same values for these parameters is not effective and that, in order to obtain better results, they should be tuned for each lexicon. Consequently, in the near future we plan to include a minimum error rate training (MERT) option for these parameters. Furthermore, we plan to address the reverse problem of phoneme to grapheme (P2G) conversion.

[2] http://nlptools.racai.ro/nlptools/index.php?page=phontrans
[3] http://nlptools.racai.ro/nlptools/index.php?page=tts

References

Baayen,R., Piepenbrock,R., and Gulikers, L. (1995).The CELEX lexical database. *Linguistic Data Consortium*, University of Pennsylvania, Philadelphia.

Bisani,M., and Ney, H. (2002). Investigations on joint-multigram models for grapheme-tophoneme conversion. *Proceedings of the 7th International Conference on Spoken Language Processing*, 105–108

Black, A., Lenzo, K. and Pagel, V. (1998). Issues in building general letter to sound rules, *ESCA Speech Synthesis Work-shop*, Jenolan Caves.

Bosch, A., and Canisius, S. (2006). Improved morpho phonological sequence processing with constraint satisfaction inference. *Proceedings of the Eighth Meeting of the ACL-SIGPHON at HLT-NAACL*, 41–49.

CMU (2011). Carnegie Mellon Pronuncing Dictionary. *http://www.speech.cs.cmu.edu/cgi-bin/cmudict.*

Content, A., Mousty, P., and Radeau, M. (1990). Une base de données lexicales informatisée pour le français écrit et parlém. *L'Année Psychologique,* 90:551–566.

Demberg, V. (2007). Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion. *Proceedings of ACL-2007*.

Divay, M. and Vitale, A. J. (1997). Algorithms for grapheme-phoneme translation for English and French: Applications, *Computational Linguistics*, 23(4):495-524.

Jiampojamarn, S., Cherry, C. and Kondrak, G. (2008). Joint processing and discriminative training for letter-to-phoneme conversion (2008). *Proceedings of ACL-2008: Human Language Technology Conference*, Columbus, Ohio, 905–913.

Laurent, A., Deleglise, P., and Meignier, S. (2009). Grapheme to phoneme conversion using an SMT system. *Interspeech*.

Marchand, Y. and Damper, R.I. (2000). A multistrategy approach to improving pronunciation by analogy. *Computational Linguistics*, 26(2):195–219.

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

Pagel, V., Lenzo, K. and Black, A. (1998). Letter to sound rules for accented lexicon compression. *International Conference on Spoken Language Processing*, Sydney, Australia.

Rama, T., Singh, A. K., Kolachina, S. (2009). Modeling Letter-to-Phoneme Conversion as a Phrase Based Statistical Machine Translation Problem with Minimum Error Rate Training. *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP*, 124–127, Suntec, Singapore.

Stan, A., Yamagishi, J., King, S., Aylett, M. (2011). The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, 53: 3, 442-450.

Taylor, P. (2005). Hidden Markov Models for grapheme to phoneme conversion. *Proceedings of the 9th European Conference on Speech Communication and Technology*.

USING FUNCTION WORDS FOR GUIDING THE PREDICTION OF THE ROMANIAN INTONATION

VASILE APOPEI, DOINA JITCĂ, OTILIA PĂDURARU

Institute of Computer Science of the Romanian Academy

Iaşi Branch, Romania

jdoina@iit.tuiasi.ro

Abstract

This paper presents a method for determining the syntactic markers of a given text, starting from a set of function words. The syntactic markers usually consist of sequences of function words combined or not with content words. We selected a set of key words and searched for different lexical contexts of each key word into a large Romanian text corpus. The context including a key word was structured into a set of morpho-lexical descriptions (sequences). The prosodic aspects of the morpho-lexical contexts were analyzed starting from the utterances of the corresponding texts. Each context will be assigned a set of prosodic markers which can be further processed by the prosodic prediction module of a Romanian Text-to-Speech (TtS) system. The syntactic markers are useful for guiding a prediction module for the Romanian intonation to correctly generate the prosodic phrasing of the input text and the melodic contours of each phrase.

Keywords: function words, morpho-lexical contents, prosodic marks

1. Introduction

The goal of this paper is to build an inventory of syntactic markers, starting from a Romanian text corpus and from the detection of the function words and their morpho-lexical contexts within a given text. A prosodic module has to assign to these markers prosodic markers (focus events, boundary events, etc.) in order to generate an adequate F0 contour for an input text.

This preliminary step is useful for improving the prosodic predictor of a Romanian TtS system, developed starting from a previously described functional model of Romanian intonation (Jitcă and Apopei 2007, 2009). The functional model-based predictor assigns a functional label to each prosodic word (accentual units). Consequently, a phrase is described by a sequence of function labels. The task of the prosodic predictor is to assign a melodic contour to the input text, starting from the prosodic markers deduced from the analysis of the syntactic markers. Each contour has a particular functional description.

Not all focus events of the neutral sentences can be predicted based on lexical analysis, because the occurrence of a focus event has sometimes prosodic reason (Kratzer & Selkirk 2007). According to the authors of this study, the distribution of the major phrase stresses (major focuses) in all-new sentences is determined by the principles underlying the syntax-phonology interface, whereas the distribution of the minor phrase stresses (the rest of the focuses) is apparently a matter for the phonology per se, and is determined by the principles of the prosodic structure organization.

In most cases, the syntactic markers can be correlated with the predicate prosodic units (accentual units - AUs, intermediate phrases - ips, intonational phrases - IPs) of an F0 contour. Setting correctly the positions of the predicate units between phonological phrases entails a correct identification of the beginning of the next phrase and of the position of its focused prosodic unit, by taking into account different types of functional prosodic unit structures. Our approach consists in generating rules that allow the identification of various syntactic markers, without the need to resort to the syntactic structure of the text.

The role of the predicate prosodic units within the prosodic hierarchy of a F0 contour is detailed in chapter 2. Several examples of markers, accompanied by different lexical contexts, are presented in chapters 3 and 4.

2. A functional perspective on the prosodic events

Apart from the text semantics, intonation has its own meaning, resulting from its own grammar and several functional categories of prosodic constituents (Selkirk 1995; Schwarzschild 1999). In these papers, the authors have analyzed a text from a functionally semantic perspective to obtain an Information Structure (IS) in terms of 'Focus' and 'Given' marks. At the prosody level, 'Focus' is associated with the 'Stress' prosodic category, while 'Given' is assigned the 'Destress' category. However, the analysis of the intonational contours reveals 'Focus' constituents without a pitch accent and pitch accents without a focus.

For this reason, other authors have not agreed with deriving expression of 'Focus' and 'Given' constituents directly from their marks (Fery & Vieri 2006). They have suggested that the prosodic events results entirely from the interaction between the constraints governing the prosodic organization of the clause (the prosodic reasons) and the general constraints governing the prosodic expression ('Stress'-'Focus' and 'Destress'-'Given') of the discourse status. In their opinion, the relation between the discourse structure and prosody relies on the ranking of several constraints. Three of them (two resulting from the 'Stress-Focus' association and one from the 'Destress-Given' association) relate the accent or its absence to the discourse structure. The others govern the position of the prosodic prominences resulting from the phrasal stress, head alignment to the intonational phrase (IP) boundaries and head alignment to the phonological phrase boundaries. Consequently, accent assignment to the heads results from the prosodic constraints, while deviation from this default is imposed, when necessary, by the discourse constraints.

In our intonational model, the prosodic constrains are expressed by functional label sequences (Jitcă & Apopei 2009), assigned to different melodic contour types applied at the IP and intermediate phrase (ip) levels. These functional label sequences are further translated into F0 pattern sequences, each pattern having its own prominence. The functional labels of a sequence correspond to the prosodic constituents (usually prosodic word) of a phrase (IP/ip). The main functions are the followings:

- PUSH and POP – correspond to the delimitative units of a phrase. In descending contours, the PUSH accentual units mark the beginning of a phrase, while the POP accentual units mark its end. In neutral intonation, a PUSH unit is more prominent than a POP one.
- LINK - corresponds to a prosodic unit endowed with a predication function at the prosodic level. It links the initial AU/AU group to the final AU/AU group within an intermediate or intonational phrase. At the morphological level, the 'Link' unit may correspond to a verbal constituent or not, while the predicate units frequently correspond to adverbial and prepositional constituents or to nouns derived from verbs.
- FOCUS (F) – corresponds to a prominent prosodic unit with a target tone reaching the maximal pitch level in an affirmative statement.

A prosodic constituent of a phrase may cumulate two functions. For example, in neutral intonational contours, when the target tone of a PUSH unit reaches the top level of the tonal space, a PUSH+FOCUS unit is generated. In contrastive focus intonational contours, when a POP unit has a target tone reaching the top level of the tonal space, a POP+FOCUS unit is generated. Therefore, the functional analysis of an intonational contour has led to different melodic contours, described by sequences of functional accentual units: *PUSH – LINK+FOCUS - POP*, (*PUSH + FOCUS) – POP, PUSH – (POP+FOCUS)*, etc.

The aim of our research we have examined how certain functional prosodic units can be correlated with certain morpho-lexical constituents to generate rules for prosody prediction of an input text. For this study we have limited the analysis to a set of function words and to their context extracted by searching them into different text corpuses.

3. Using the function words by the prosodic prediction module

In what follows, we shall analyze the intonational contours of a short Romanian text, to illustrate how the function words predict the prosodic events of the intonational contour corresponding to the text.

Example: *Dificultatea inerentă în cadrul acestor competiţii apare deoarece activitatea evaluatorilor nu este una absolut cuantificabilă*. (The inherent difficulty within these competitions emerges because the assessors' work is not absolutely quantifiable.)
The lexical analysis of the text led to the following lexical cues: *în cadrul* (within), *acestor* (these). *deoarece* (because), *nu* (not), and *absolut (absolutely)*.

In this paper, the term 'lexical event' will refer to the occurrence of a function word. A morpho-syntactic event will refer to the occurrence of a sequence of words with a particular morphological function sequence. A model of building up rules to predict prosodic markers is presented in Table 1. The symbols in this table have the following meaning:

- NG = nominal group;
- V = verb;
- N = noun;
- VAux = auxiliary verb;
- P_mark = predicate marker;
- F_mark = focus marker;
- VA_mark = auxiliary verb marker;
- B*i*_mark = break index *i*, *i* = 2, 3 or 4. In Table 1, only B2_mark and B4_mark are present.

The first four rows in Table 1 present rules based on lexical events. Here, 'în cadrul' (within), 'deoarece' (because), 'nu' (not), and 'absolut' *(absolutely)* are function words. For example, the rule in the third row has the following meaning: *If* a word sequence composed of a nominal group containing more than two words, followed by the word 'nu' (not), followed by a verbal group is detected, *then* the nominal group receives (←) a break index 3 mark and 'nu' (not) receives a focus marker.

The last five rows present rules based on morpho-syntactic events. An example of a morpho-syntactic event is the occurrence of the verb 'apare' (emerges) after a nominal group containing more than two words (row 5 in Table 1). In this case, the nominal group will end in a B4_mark.

Table 1: A model of building up rules to predict prosodic markers

Event	**Event type**	**Rule: *if* 'sequence' is detected *then* 'prosodic markers' are set**	
		Sequence	**Prosodic markers**
în +NG	lexical	*în cadrul*+ **{NG}**	*în cadrul* ← P_mark
deoarece	lexical	{V}+*deoarece*+{NG}	V ← F_mark; *deoarece* ← P_mark
nu	lexical	{NG>2 words}+*nu* + {VG}	{NG>2 words} ← B3_mark *nu* ← F_mark
absolute	comparison degrees	*absolut*+ {adjective}	*absolut* ← P_mark.
{NG>2 words}+V+deoarece	morpho-syntactic	{NG>2 words}+{VG}	{NG>2 words} ← B4_mark
dificultatea inerentă	morpho-syntactic	{N}+{adjective}	B2_mark
acestor competiţii	morpho-syntactic	Determinat+N	B2_mark
este	morpho-syntactic	{VAux}	VA_mark
point	lexical	word+.	{word+.} ← B4_mark

After detecting all lexical and morpho-lexical events, the predictor maps the input text at the prosodic level. As a result, the words will be translated into accentual units. In the selected example, each word is assigned an accentual unit, except for the clitic 'în', which forms an accentual unit together with 'cadrul'.

For each event, the prediction module is endowed with one or more predefined rules, used to check its morpho-syntactic context. The existence of at least two rules means that the contexts have different mapping at the prosodic level. When processing a rule, the predictor uses one or more prosodic marks which characterize the mapping of the input text at the prosodic level. There are three types of prosodic markers:

- markers for FOCUS prosodic units. Such a marker corresponds to a word in the context of the rule being processed;
- markers for LINK prosodic units. Such a marker also corresponds to a word in the context of the rule being processed;
- markers for the end boundary of an intonational phrase. They are similar to the Break Indices of the ToBI annotation system:
  - B4_mark is used for an IP boundary, corresponding to a Break Index 4;
  - B3_mark is used for an ip boundary, corresponding to a Break Index 3.
  - B2_mark is used for a phonological group boundary, corresponding to a Break Index 2.

Using the prosodic markers deduced after processing the rules in Table 1, the prediction module has generated the following phrasing for the selected text:

{[(***Dificultatea*** *inerentă)* *în cadrul* *(acestor competiţii)]} {[**apare** deoarece (activitatea evaluatorilor)][**nu** este una absolut cuantificabilă]}*.*

Here, the IPs are demarcated by '{ }', the ip's by '[ ]', and the minor phrases by '( )'.
The words selected by the predictor for predicate (link) intonation appear in underline, while the focused words appear in bold.
The prosody prediction module (Jitcă & Apopei 2011) has been designed to use these prosodic markers during the phrasing process and also when selecting a melodic contour for each phrase from the utterance tree hierarchy.

4. Building the set of lexical events starting from function words

In this section, we shall briefly present the results of an analysis on the occurrence of a set of function words and on their accompanying contexts in a selected Romanian text corpus. This corpus represents the text of George Orwell's "1984" novel and has 108.000 words. The set of function words was chosen taking into account the discourse markers analyzed by Teodorescu (2005), the statistical structure of words in a literary Romanian corpus analyzed by Vlad et al. (2011) and our own remarks concerning the implications of certain functional words on prosody.

The search for the selected function words and their accompanying contexts within the text corpus was performed with our own Visual C++ program. The program output the number of occurrences of each word in the input list and a set of lists containing the contexts of the searched words.

The selected set of function words contains the following words: acest (this), *adică* (that is, i.e., I mean), *aşadar* (therefore), *astfel* (thus), *atunci* (then, at that time, in this case/situation, in these circumstances; *atunci când* = when), *când* (when), *care* (which, who), *căci* (because), *cărei* (whose), *către* (to, toward), *chiar* (even), *câte* (how many), *cum* (how), *dacă* (if), *deci* (therefore), *deoarece* (since, because, as), *deşi* (even if, even though), *doar* (only, just), *după* (after, following), *încă* (yet, still, even; *încă o dată* = one more time, again), *încât* (that; *astfel încât* = so that), *în* (in), *însă* (but), *însăşi* (herself, itself), *întotdeauna* (always), *întrucâtva* (somewhat), *la* (to, at), *nici* (neither, nor, even), *nu* (no, not), *numai* (only), *parcă* (I think/thought; *de parcă* = as if), *pentru* (for), *tocmai* (just, precisely, exactly, very, right), *tot* (all, everything), *totuşi* (however, but).

Fig. 1 depicts the number of occurrences of the selected function words within the text corpus used in the present study. This figure shows: a large number of statements containing the preposition (*în, la*), the relative pronoun *care* and negation words (*nici*, *nu*); a high rate of occurrence of the indefinite pronoun/adjective *tot*, in all its forms (*tot*, *toată*, *toţi*, *toate*); a relatively high rate of occurrence of particular prepositions (*pentru*, *după*) and adverbs (*parcă*, *când*, *atunci*).

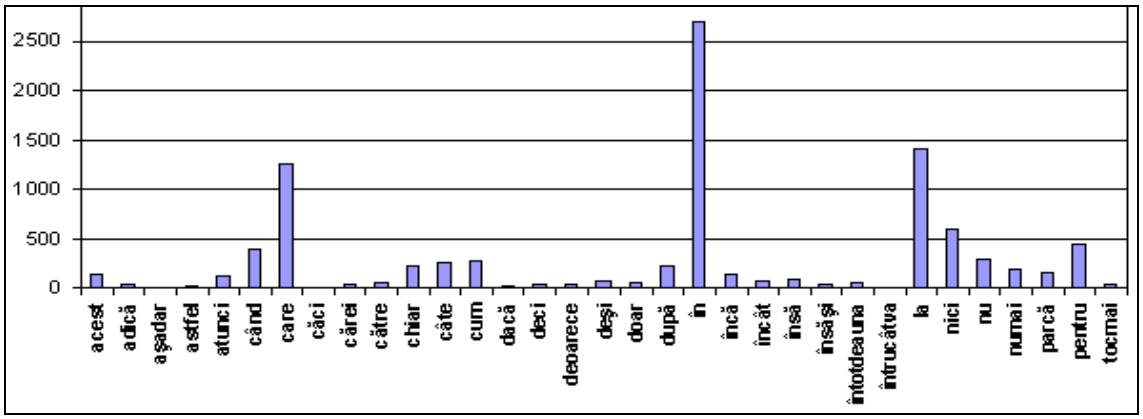


Figure 1: Number of occurrences of the analyzed words within the selected text corpus

The function words with the highest rates of occurrence were selected for a subsequent analysis of their lexical contexts.

The lists containing the contexts of the selected function words were built up by truncating the text in the vicinity of the function words as follows:

- the left context begins immediately after the last punctuation mark preceding the analyzed word;
- the right context begins after the analyzed word and ends either when reaching a dot, colon or semicolon or after the first eight words.

Table 2 presents a fragment of the context list corresponding to the word ‘atunci’.

Table 2: Some lexical contexts of the word ‘atunci’

Word	**Context**
atunci	Şi chiar şi **atunci** pătrunzi numai printr-un labirint de reţele de sârmă (…and even **then**, you can enter there only through a maze of wire networks.)
atunci	**atunci** i-a trecut pentru prima oară prin minte idea (…**then**, this idea crossed his mind for the first time.)
atunci	Ce i s-a părut curios **atunci** a fost faptul că în vis vorbele acestea (…what seemed curious to him **at that time** was the fact that these words, in his dream…)
atunci	că le vedea **atunci** în ochii mari ai mamei şi ai surorii (…that he saw them **at that time** in his mother’s and sister’s open eyes…)
atunci	Winston şi-a dat seama **atunci** că bătrânului tocmai i se întâmplase cine ştie-ce (Winston realized **then** that something terrible happened to the old man.)
atunci	dacă toate documentele povestesc aceeaşi gogoriţă, **atunci** minciuna se transferă în istorie şi devine adevăr. (…if all documents tell about the same bugbear, **then** the lie passes into history and becomes truth.)
atunci	Dacă nu poate, **atunci** măcar să-l deformeze sau să-l mânjească. (If not, **then** at least to deform or mar it…)
atunci	chiar şi **atunci** ar fi putut suporta să trăiască lângă ea, (…even **then**, he could bear to live with her.)
atunci	Lui Winston i-a sărit **atunci** inima din loc. (Winston’s heart jumped into his throat **then**.)

The contexts accompanying a function word are elicited by the computer program, then are further processed during two stages. During the first stage, the program detects all contiguous sequences (contexts) of the function words. In the second stage, these sequences are manually extended with new words, to build up meaningful phrases, ready for a subsequent utterance.

Table 3 presents the list of contiguous contexts including the function word ‘atunci’. These contexts will be analyzed from a prosodic point of view in order to assign a set of rules containing prosodic markers to the corresponding function word.

The utterances of the sentences associated with the selected contexts allow us to build up rules for setting adequate prosodic markers which will further be used by prediction module for phrasing and melodic contour selection.

Table 3: The contiguous contexts of function words containing the word ‘atunci’

Left context	**Key word**	**Right context**
,	atunci	
[de] abia		
ca [şi]		
[ceva] care		
că		
chiar [şi]		
[iar] dacă		
dar [şi]		însă
de parcă		înseamnă
deci		când
decât		numai
deşi [pe]		aia
doar		acel; acea
numai		abia
[de] pe		încoace; încolo
pentru că		parcă nici
[niciodată; ceva care; în care] până		
încă		
fiindcă		
şi [tot]		
imediat; exact; precis; tocmai;		
VG	atunci	cum
		că
		încă

Table 4 presents the set of meaningful phrases built up after the second stage of processing the contexts of the word ‘atunci’. Some of them have been uttered and are ready for prosodic analysis at the IP/ip level, using the functions presented in section 2.

Table 4: Excerpt of meaningful sentences selected for further utterance and prosodic analysis

Word	**Phrases including morpho-lexical contexts**
atunci	Este imposibil să intri acolo altfel decât cu treburi oficiale, **şi chiar şi atunci** pătrunzi numai printr-un labirint de reţele de sârmă ghimpată. (It is impossible to enter there, except on official business, **and even then**, you can enter only through a maze of wire networks.)
atunci	Lucrul rămâne valabil **şi atunci când** acelaşi eveniment trebuie modificat de mai multe ori în cursul aceluiaşi an. (This thing remains valid **even when** the same event must be changed several times during the same year.)
atunci	Parcă vede şi acum preţioasa bucăţică de ciocolată **care pe atunci încă** se mai măsura în uncii. (Even now, he sees in his mind's eye the precious piece of chocolate **which** was **still** measured in ounces **at that time**.)
atunci	Julia nu pune la îndoială tezele partidului **decât atunci când** îi afectează propria ei viaţă într-un fel sau altul (Julia does not question the party's theses **except when** these affect her own life in one way or another.)
atunci	Winston şi-a dat seama **atunci că** bătrânului tocmai i se întâmplase cine ştie-ce lucru cumplit. (Winston realized **then** that something terrible happened to the old man.)
atunci	**Dacă** toate documentele povestesc aceeaşi gogoriţă, **atunci** minciuna se transferă în istorie şi devine adevăr. (**If** all documents tell about the same bugbear, **then** the lie passes into history and becomes truth.)
atunci	Idealul celor de jos, **atunci când** se întâmplă ca aceştia să aibă vreun scop în viaţă, este desfiinţarea tuturor diferenţelor între oameni copleşiţi de greutăţile vieţii. (Lower class people's ideal, **when** it happens that they have a purpose in life, is the abolition of all differences between people overwhelmed by the hardships of life.)
atunci	**Din moment ce** toate aceste bunuri nu mai constituiau proprietate privată, **atunci însemnă** că formau proprietate publică. (**Since** all these goods have no longer been private property, **it means** that they turned into public property,
atunci	Vede armata eurasiană năvălind peste frontiera **până atunci** neatinsă şi scurgându-se spre sudul Africii. (He sees the Eurasian army rushing across the **hitherto** untouched border, and running toward South Africa.)

5. Conclusions

In this paper, we have proposed a method of using function words and their morpho-lexical contexts by the prosody prediction module to generate prosodic markers. These markers will be further used during the phrasing process and when selecting a melodic contour for each phrase.
The analysis of the rate of occurrence of the function words presented in section 4 has allowed us to find the most frequently encountered contexts. The sentences in the selected text corpus including these contexts have been elicited for further utterance. The prosodic analysis of the parallel text-speech corpus has led to finding prosodic markers which will be assigned to morpho-lexical contexts.

Acknowledgments: This study has been conducted within the research program of the Institute of Computer Science of the Romanian Academy.

References

Féry C., Vieri S. L. (2006). Focus projection and prosodic prominence in nested foci. Language 82, 131–150.

Jitcă D., Apopei V. (2007). Corpus de voce pentru limba română adnotat cu etichete funcţionale la nivelul unităţilor de accentuare, *Lucrările atelierului "Resurse lingvistice şi instrumente pentru prelucrarea limbii române"*, Iaşi, 31-39.

Jitcă D., Apopei V., Jitcă M. (2009). The F0 contour Modelling as Functional Accentual Unit Sequences, International Journal of Speech Technology, 12:(2-3), 75-82.

Jitcă D., Apopei V. (2011). An Intonation Prediction Module for Romanian TTS System, as a Prosodic Tree Generator, SPED-2011, *IEEE Conference Publications Program, IEEE Xplore Digital Library*.

Katzer A., Selkirk E. (2007). Phase theory and prosodic spellout: The case of verbs The Linguistic Review 24, Special issue on Prosodic Phrasing, (Sonia Frota and Pilar Prieto eds.), 93-135.

Schwarzschild R. (1999). GIVENness, AVOIDF and Other Constraints on the Placement of Accent, *Natural Language Semantics* 7, 141-177.

Selkirk E. (1995). Sentence Prosody: Intonation, Stress and Phrasing. Handbook of Phonological Theory, (John Goldsmith ed.), Cambridge, MA: Blackwell, 550-569.

Teodorescu H.N. (2005). A proposed Theory in Prosody Generation and Perception: The Multidimensional Contextual Integration Principle of Prosody, Trend in Speech Technology, Editura Academiei Române, 109-118.

Vlad A., Mitrea A., Ciuca S., Luca A. (2011). A study on the statistical structure of words and of word digrams in a literary romanian corpus, SPED-2011, *IEEE Conference Publications Program, IEEE Xplore Digital Library*.

MAXIMUM ENTROPY BASED MACHINE TRANSLITERATION. APPLICATIONS AND RESULTS

ADRIAN ZAFIU[1], TIBERIU BOROŞ[2]

[1]*University of Pitesti, Electronics, Communications and Computers Department, Piteşti, Romania*

[2]*Research Institute for Artificial Intelligence "Mihai Drăgănescu, Romanian Academy, Bucharest, Romania*

adrian.zafiu@upit.ro, tibi@racai.ro

ABSTRACT

Transliteration has been previously used in the field of Natural Language Processing (NLP) with emphasis for machine translation (MT) between languages that are either incompatible at the phonetic level or employ very different alphabet systems. In this article we propose a new statistical method for transliteration and we discuss the possibility of using transliteration for two new tasks, besides MT. The first task refers to multilingual search based on the phonetic similarity between words (what we call perception-based search) and the second task is linked to text-to-speech (TTS) synthesis in the multilingual environment. The method that we propose for transliteration is similar to direct-orthographic-mapping in the sense that it does not require any intermediate phonetic level. Our experiments currently focus on the following languages: English, Bulgarian, Romanian and French. For the above mentioned languages, we seek to answer two questions: "can transliteration be achieved based on limited lexical context classification?" and "what other applications besides MT can benefit from transliteration?".

Keywords: machine translation, maximum entropy optimization, transliteration, text-to-speech synthesis

1. Introduction

Machine translation (MT) systems are often faced with the task of handling words that do not have a (known) corresponding translation (e.g. proper nouns, some technical terms, etc.). When the two languages share similar orthographic inventories it is a common practice to leave such words as they appear in the original text. Such a resolution is not possible when the two languages are highly incompatible at the orthographic and phonetic levels (for example, the English sounds 'L' and 'R' collapse into a single sound in Japanese). A solution to this task is to convert the original words, using a set of mappings from one orthographic system to another, in such a way that the resulting word would have a similar phonetic representation. The transliteration is the process used by the component responsible for this type of operation. By definition, **transliteration** means *converting letter by letter from one writing system to another* and **transcription** is the process of *phonetically mapping words* between languages.

However, most transliteration systems work by mapping the letters of a word to similarly sounding letters in the target language. So, to be consistent with other research papers we will use the term transliteration to refer to the task of mapping the letters of a word in the source language to letters of a "pseudo-word" in the target language so that the two words have similar pronunciations.

In the past, several methods for transliterating between two languages were introduced, mainly focused on automatic transliteration between English, Chinese, Japanese, Korean and Arab.

In (Knight & Graehl, 1997), finite state transducers were used to transliterate between Japanese and English. Their method was later adapted in (Stalls and Knight, 1998) for bidirectional transliteration between English and Arab. Similar methods for transliteration were presented in (Jung et al., 2000), (Meng et al., 2001), (Virga & Khudanpur, 2003).

In their work, (Haizhou et al., 2004) classify the above mentioned methods as phonetic approaches to transliteration. They propose a new technique that focuses on direct orthographic mapping (DOM). Their method is also referred as n-gram based transliteration.

In this paper we seek to answer two questions: "can transliteration be achieved based on limited lexical context classification?" and "what other applications besides MT can benefit from transliteration".

To answer the first question we proposed a data-driven method for transliteration based on a MaxEnt classifier (see section 3). The proposed method performs transliteration at orthographic level without using an intermediate phonetic level and it only requires a lexicon composed of original words in the source language with their corresponding transliterations in the target language. (Haizhou et al., 2004) introduce a comparison between transliterations obtained with their method versus an ID3 algorithm for limited context classification. However, it was clearly demonstrated that this algorithm (ID3) is outranked by other classifiers when applied to letter-to-sound (LTS) conversion (Black et al., 1998), (Jiampojamarn et al., 2008), (Pagel et al. 1998), (Bisani & Ney, 2002), (Marchand & Damper, 2000), (Demberg, 2007). Given the similarities that arise between LTS and transliteration it is likely that other classifiers could perform better than ID3.

For the later question we set out to see if transliteration can improve TTS synthesis (first application) (section 4.1) and we propose a multilingual phonetic perception based search technique (second application) that can highly improve user experience with search engines (section 4.2), travel assistants and navigation systems.

2. Building the training lexicons

This research was initially focused on improving the performance of a TTS system, when handling out-of-vocabulary (OOV) words. For objective reasons we focused our attention on transliteration between English, Bulgarian, French and Romanian. When we run our TTS experiments, we noticed the problem with some OOV words belonging to the foreign word class. Most of these words originated from English and French. We added Bulgarian to our list, because it uses a different alphabet from the others. Our work was focused on minimizing the impediments posed by foreign OOV words in Romanian TTS synthesis and, in our case we had to handle words coming from the above mentioned languages. To our knowledge there are no freely available transliteration lexicons between any of these languages. For this reasons we set out to create our own corpora, which will be made publically available for research.

The general method for building transliteration lexicons as presented in (Knight & Graehl, 1997) is:

1. Choose a set of representative words for the source language and obtain their phonetic transcriptions manually or automatically using rules specific to the source language.
2. Adjust the phonetic transcriptions using hand-written rules that map from the phonetic inventory of the source language to the phonetic inventory of the target language
3. Manually or automatically map back to orthography using rules specific for the target language.

The first transliteration lexicon we created was an English to Romanian corpus. We chose the CMUDict as a starting point in our development and we proceeded using the phonetic transcriptions provided inside. However, the CMUDict contains a lot of foreign words adapted to English such as: Italian: braggiotti, castelli, castelluccio; German: aachen, abbenhaus, schlender, schlenker; Polish: zawistowski.

Because we aimed (in this case) at learning transliteration rules only from English native words to Romanian we filtered out all foreign words and proper names, leaving 40,606 entries in the CMUDict. The remaining data was converted to their Romanian transliterations using a set of hand-written rules with post-validation (see table 1 for examples).

Table 1: English to Romanian transliteration examples

En Phoneme	**Example word**	**English phonetic transcription**	**Romanian transliteration**
AA	odd	**AA** D	**a**d
AE	at	**AE** T	**e**t
AH	hut	HH **AH** T	h**a**t
AO	ought	**AO** T	**o**t
AW	cow	K **AW**	c**au**
AY	hide	HH **AY** D	h**ai**d
B	be	**B** IY	**B**i

For building our second lexicon (Bulgarian to Romanian), we compiled a list of 54,189 word-forms from various corpora. These words were also converted to their Romanian

orthographic representations using hand-written rules (table 2) (without requiring phonetic transcriptions, due to the preponderantly phonetic orthographies of both languages).

Table 2: Bulgarian to Romanian transliteration examples

BG orthography	**Word**	**RO orthography**
б	банско	b
д	видин	d
ш	свищов	ș
я	смолян	ia
х	хасково	h

The French to Romanian transliteration corpus was created similarly as the English to Romanian lexicon (see table 3 for some example rules). In this case, we used the Brulex pronunciation lexicon as a starting point.

Table 3: Examples of French to Romanian transliteration corpus

FR phoneme	**Romanian orthography**
y	u
f	f
m	m
`	o
1	e

3. Automatic transliteration using a limited lexical context classifier

The task of transliteration can be formulated as finding a set of rules, which applied to an input sequence of orthographic symbols/characters specific to the source language, generates a set of symbols/characters specific to the target language. The goal is to maximize the value of a similarity function between the two phonetic representations of the original and the processed words.

There are two types of methods used for transliteration:

- **Type 1**: Phonetic based methods require three sets of rules to be applied for orthography-to-sound conversion (for the source language), phonetic adaptation (between source and target languages) and sound-to-orthography conversion (for the target language);

- **Type 2**: Direct orthographic methods do not require such knowledge, as the idea is to infer rules for direct conversion at orthographic level.

All sub-tasks of first type methods are prone to errors when applied separately, and the overall error rate is higher than the errors of the second type methods, hence our choice to base our research on direct orthographic mapping.

Transliteration does not have a one-to-one (bijective) correspondence between the source and target sequences at orthographic level. One orthographic symbol can spawn two or more orthographic symbols in the target language or can even have a void (NULL character) mapping. This means that before proceeding with the training process, transliteration requires alignments between letters of the source and target words. These alignments can be obtained using the Expectation Maximization (EM) algorithm (Hartley, 1958), (Dempster et al., 1977).

We based our method on a limited lexical context Maximum Entropy (MaxEnt) classifier. MaxEnt classifiers have been previously used in natural language processing (NLP) for part-of-speech (POS) tagging (Ratnaparkhi, 1996), sentence splitting (Reynar and Ratnaparkhi, 1997), (Agarwal, 2005) etc. Maximum Entropy builds a model that maximizes entropy by assuming a uniform distribution for unseen data (Berger et al., 1996).

Treating transliteration as a classification task means that for each orthographic symbol in the input sequence the system has to predict a label using features extracted from a limited lexical context window. The label represents an orthographic symbol, group of symbols or an empty sequence (NULL character) in the target language. The sum of predicted labels represents the transliteration of the input sequence to the target language. Using s to denote the current orthographic symbol, s_i to denote the orthographic symbol at distance i from the current symbol and l_{-1} to represent the previously assigned label (for the previous symbol) we have the following features:

- s_{-1},s – current symbol plus previous symbol;
- s_{-2} ,s_{-1},s – current symbol plus the previous two symbols;
- s,s_{+1} – current symbol plus the following symbol;
- s,s_{+1},s_{+2} – current symbol plus the following two symbols;
- s_{-1},s,s_{+1} – current symbol plus the previous and following symbols (the identity feature);
- l_{-1} – previously assigned label (for output cohesion).

This set of features was chosen using a trial an error process. The current symbol alone was not informative enough compared to the combination of the current symbol plus the previous and following symbol, thus we named this composite feature the **identity feature**. Increasing the current context window length did not significantly improve the prediction accuracy of the model and in some cases lead to overtraining. Ignoring the

previously assigned label had a large negative impact on the accuracy, and adding more labels to the history did not yield statistically relevant higher accuracy rates.

3.1. Evaluation

We evaluated our transliteration accuracy using 10-fold validation methodology. Each training lexicon was divided in 10 equal subsets and we measured our systems accuracy by averaging the prediction accuracy on each subset while training on the other nine. We present the results in terms of Word Accuracy Rates (WAR) as the number of fully correct transliterated word versus the total number of words (see table 4).

Table 4 : Current transliteration results in terms of word accuracy rates

		Target			
		English	**Romanian**	**Bulgarian**	**French**
Source	**English**	-	78.15%	77.12%	N/A
	Romanian	43.18%	-	97.34%	92.08%
	Bulgarian	N/A	97.21%	-	N/A
	French	N/A	56.45%	N/A	-

4. Application of transliteration

Recent focus in improving accessibility and general access to information through constant advances in the field of human-computer interaction has led to the wide spread of spoken language processing technologies applied in computers and micro-devices, with an increased interest for speech synthesis, speech recognition and improved text accessibility.

In this section, we focus on the mainstream task of TTS synthesis and how transliteration can help to improve the quality of speech synthesis from arbitrary texts. We then introduce an application for transliteration, which improves user experience with search engines, GPS systems and any other type of travel assistants. To our knowledge, the later introduced application has not been suggested by other authors. In this section, we focus on the mainstream task of TTS synthesis and how transliteration can help improve the quality of speech synthesis from arbitrary texts.

4.1.Transliteration and text-to-speech synthesis

We have to start by pointing that text-to-speech synthesis has to be able to synthetize voice from any arbitrary or unrestricted text. This involves a series of pre-processing steps such as: converting numbers, dates, formulas to their spoken form; phonetically transcribing words; syllabification; lexical stress prediction etc. Such tasks can be

normally attained using large lexicons of already processed words but there are always exceptions, in the form of out-of-vocabulary (OOV) words, which have to be treated automatically. Over the years, a number of machine-learning (ML) methods have been proposed to solve each of the above stated subtasks with the main focus on OOV words that are also native language words. It is expected that there are also foreign words that fall into the class of OOV. Such situations are fatal for TTS synthesis, as leaving such words unchanged and applying the same rules for phonetic transcription or syllabification as those specific to the native language of the TTS system would yield faulty results.

We differentiate two ways to handle such words. The first method is to attain the above mentioned sub-tasks using a custom set of rules adapted to the foreign language from which the word originates. However, having different sets of rules for more than the native language is challenging.

The second method (that we propose here) is to use transliteration on these foreign words and to convert them to pseudo-native words. This facilitates using a single package of native rules for the tasks of phonetic transcription, syllabification and lexical stress.

The difference between the two methods (figure 1) is that the first method applies phonetic transcription with syllabification and lexical stress rules for the foreign language(s) followed by an adaptation at phonetic level between the two languages, while the second method uses transliteration to produce a pseudo-native word and then uses the native rule sets for attaining the final goal.

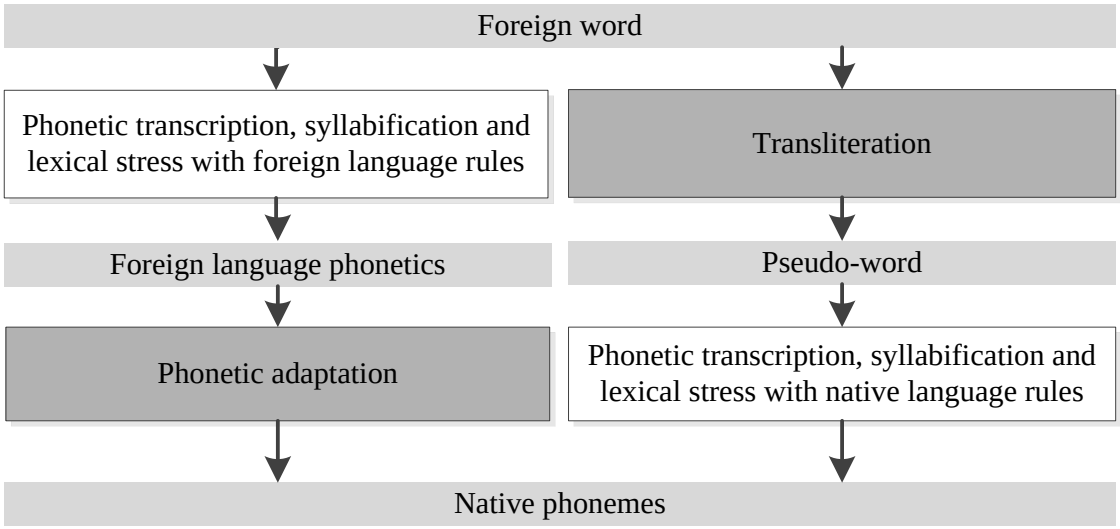


Figure 2: Foreign word handling in TTS synthesis

There are several reasons for using transliteration in TTS synthesis. First we argue that constructing or attaining lexicons for the TTS tasks of syllabification, lexical stress and phonetic transcription is more demanding than building transliteration lexicons. Secondly the ML methods used for generating the prosody of the TTS system are trained using native utterances. It is likely that using custom rules for generating syllabification and lexical stress on foreign words would generate previously unseen data which impedes the correct functioning of such methods. Applying syllabification

and lexical stress on the pseudo-words obtained by transliteration is likely to produce pronunciations different from those generated by a native speaker (of the foreign language). However, in practice, this is not an understanding issue since many non-native speakers could pronounce such foreign words similarly, misplacing the lexical stress and making adaptations at the phonetic level.

4.2. Detecting which words require transliteration in TTS

One common problem with both approaches to foreign word adaptation for TTS synthesis is detecting when an OOV word is a foreign word and also what is the source language of that word. One partial solution to the problem is to use a lookup table of word-forms for each foreign language that the system has transliteration rules for. Such a list is easier to attain than a list of fully processed words and it can be done by crawling through documents written in specific languages. Any OOV word found by the TTS system has to be checked against these precompiled lists and once the word occurs in the lexicon of some language it can be transliterated to a native pseudo-word using the source language specific rule set. It is also important to keep a separate word-form list for the native language and to check if the word is not inside this list (some words may have identical orthographies in more languages: e.g. “mi-nus” is written identically in both Romanian and English). The later list is important for determining when not to apply transliteration.

There are however cases where a word or a group of words do not appear in any lexicon (such may be the case of uncommon proper nouns). Based on the fact that some orthographic symbols (especially those that have diacritics) or groups of symbols are uncommon in certain languages the assumption that a word should be transliterated can arise from testing for such occurrences. For example, characters such as ‘y’ or groups like “ck” are highly uncommon for Romanian.

The entire process is summarized in figure 2.

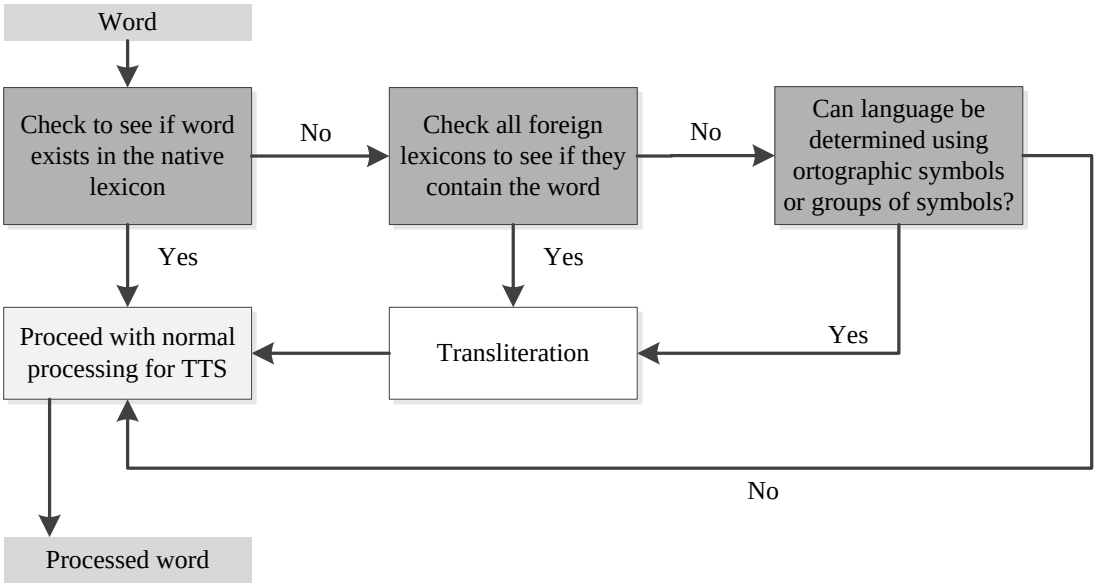


Figure 3 : OOV word handling in TTS synthesis

4.3. Perception base search

As explained earlier in this article perception based search is a method for finding persons, street names, cities, etc. based on the phonetic perception of what that word "sounds like". Using the proposed method we obtain possible spellings for a word written "as heard" in the native language of a user.

To give an example, let's suppose that we know nothing about a city except the fact that it sounds like 'YI AE N T S YI AW'. There are no other information regarding neither the country nor the language in which it should be written and therefore no information about what orthography to use in order to find out more about this location. Perception based search allows obtaining the exact spelling (in the source language) for this location just by typing in the word in the ones native language. A Romanian native speaker would just input the word as "ienţiau", an English native speaker would enter "ientsiaw" and a Bulgarian native speaker would type "ябнцяу". The answer would be "燕郊", which is a location situated in north-east of China in the province Hebei.

To our knowledge, the closest method to the one proposed here is described in (Krishnan et al, 2009), (WO Patent WO/2009/005,961, 2009). As we will show, there is an important difference between using phonetic representations (their method) and directly mapping at orthographic level:

1. A non-native speaker's perception of what a word sounds like is influenced by the phonetic inventory of his native language and it is not 100% accurate because not all languages share the same inventory;
2. Conversion rules from orthography to sounds and back are complex and there are cases where there is no possible combination of orthographic symbols that would generate the perceived phonetic sequence;
3. Multiple spellings can generate the same phonetic sequence (homophones);
4. As mentioned by (Knight & Graehl, 1997), back transliteration does not share the same flexibility as forward transliteration.

All the above mentioned facts reduce the level of reliability when using phonetic representations to get the similarity between two words originating from different languages. To overcome these problems we propose a different strategy: when given an input string in a native language we transliterate all known locations, names etc. from their source languages into the speaker's native language and we directly compare the resulting strings with the input, using a function such as the Levenshtein Distance. Each language has its unique characteristics that dictate the phonetic inventory, the phonetic transcription rules, the way native speakers perceive words from other languages and the way they would spell these words (which is a process accompanied by information loss). Choosing the direction of transliteration should be based on the highest accuracy obtained by the transliteration of the OOV words. For example, if we search a string in Romanian and we want to check foreign English names against our input string we transliterate from English to Romanian and compare results, not the other way around, because English to Romanian transliteration (regarded by (Knight & Graehl, 1997) as forward transliteration) works a lot better that Romanian to English transliteration (back transliteration). For the same reason, we also use forward-transliteration if the input string is in English.

The method proposed in the Patent uses what is referred to as a phonetically normalized character set for word encoding. They store the words in a database and they use this phonetically normalized encodings to perform search. No details are given on the construction of the phonetically normalized character set or on the models used for converting words into this type of representation.

5. Conclusions and future work

We presented a method that can be automatically trained for transliterating between any two pairs of languages and we thoroughly tested our system for English, French, Bulgarian and Romanian. Using a limited context classifier for attaining transliterations for the above mentioned languages is a viable solution.

In section 4.1 we proposed a strategy for handling OOV foreign words, which are one of the plagues of unrestricted TTS synthesis. Although WAR rates are lower for some language pairs, in a true scenario, not all words are OOV and even if transliteration fails and does not produce fully correct transliterated words, the letter accuracy rate is very high, indicating that there, very well, may be only one incorrectly classified letter. This means that using the pseudo-word, even if it is not fully correct, is preferable to using the direct unmodified foreign word.

The transliteration corpora we created will be made publically available for research purposes. Our current priority is increasing the number of lexicons. The next language of interest to us is German.

Using Romanian as a pivot we will add another transliteration lexicon from English to Bulgarian. We plan to exploit the fact that the accuracy of Romanian to Bulgarian transliteration is very high (above 99% letter accuracy rate and 97% word accuracy rate), allowing us the following procedure:

1. Train to transliterate from Romanian to Bulgarian;
2. Use our tool to transliterate the Romanian pseudo-words from the English to Romanian corpus into Bulgarian pseudo-words, thus generating English to Bulgarian mappings.

Evaluating the perception based search methodology poses a series of challenges. In order to correctly asses the performance of the system in real conditions we have to use native speakers of the languages in which the search is performed. The test corpus has to be created manually and it has to contain a significant number of entries in order to correctly asses the system accuracy. Also a comparison with other multilingual oriented search algorithms is required for a thorough validation of the presented idea.

References

Bisani, M., and Ney, H. (2002). Investigations on joint-multigram models for grapheme-tophoneme conversion. *Proceedings of the 7th International Conference on Spoken Language Processing*, 105–108.

Black, A., Lenzo, K. and Pagel, V. (1998). "Issues in building general letter to sound rules", *ESCA Speech Synthesis Workshop*, Jenolan Caves.

Bosch, A., and Canisius, S. (2006). Improved morpho phonological sequence processing with constraint satisfaction inference. *Proceedings of the Eighth Meeting of the ACL-SIGPHON at HLT-NAACL*, 41–49.

CMU (2011). Carnegie Mellon Pronuncing Dictionary. *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

Demberg, V. (2007). "Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion". *In Proceedings of ACL-2007*.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39:1, 1–38.

Hartley, H. (1958). Maximum likelihood estimation from incomplete data, *Biometrics*, 14, 174–194.

Jiampojamarn, S., Cherry, C. and Kondrak, G. (2008). Joint processing and discriminative training for letter-to-phoneme conversion. *Proceedings of ACL-2008: Human Language Technology Conference*, Columbus, Ohio, 905–913.

Jung, S. Y., Hong, L. S. și Paek, E. (2000). An English to Korean Transliteration Model of Extended Markov Window. *Proceedings of COLING*.

Knight, K. and Graehl, J. (1997). Machine transliteration. Proceedings of the Thirty-Fifth *Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Somerset, New Jersey, 128–135.

Krishnan S, H., Bendapudi, P., Gore, A. S. (2009). WIPO Patent *No. 2009005961*. Geneva, Switzerland: World Intellectual Property Organization.

Li, H., Zhang, M. și Su, J. (2004). A joint source-channel model for machine transliteration. *Proceedings of the 42nd ACL Annual Meeting*, Barcelona, Spain, 159–166.

Li, M., Zhang, Y., Zhu, M. and Zhou, M. (2006). Exploring distributional similarity based models for query spelling correction. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, 1025–1032.

Marchand, Y. and Damper, R.I. (2000). A multistrategy approach to improving pronunciation by analogy. *Computational Linguistics*, 26(2):195–219.

Meng, H.M., Lo,W-K., Chen, B. și Tang, K. (2001). Generate Phonetic Cognates to Handle Name Entities. *English-Chinese cross-language spoken document retrieval, ASRU*.

Pagel, V., Lenzo, K. and Black, A. (1998). "Letter to sound rules for accented lexicon compression", *International Conference on Spoken Language Processing*, Sydney, Australia.

Rama, T., Singh, A. K., Kolachina, S. (2009). Modeling Letter-to-Phoneme Conversion as a Phrase Based Statistical Machine Translation Problem with Minimum Error Rate Training. *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009*, Suntec, Singapore, 124–127.

Stalls, B.G. și Knight, K. (1998). Translating Names and Technical Terms in Arabic Text. *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*.

Virga, P., Khudanpur, S. (2003). Transliteration of Proper Names in Crosslingual Information Retrieval. *Proceedings of ACL 2003 workshop MLNER*.

INDEX OF AUTHORS