PROCEEDINGS OF THE 10TH INTERNATIONAL CONFERENCE "LINGUISTIC RESOURCES AND TOOLS FOR PROCESSING THE ROMANIAN LANGUAGE" 18-19 SEPTEMBER 2014

Editors:

Mihaela Colhon

Adrian Iftene

Verginica Barbu Mititelu

Dan Cristea

Dan Tufiș

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

Department of Computer Science and the Faculty of Letters

University of Craiova

This volume was published with the aid of
the Faculty for Computer Science,
“Alexandru Ioan Cuza” University, Iași, Romania

ISSN 1843-911X

PROGRAM COMMITTEE

Costin Bădică, Faculty of Automation, Computers and Electronics, University of Craiova
Tiberiu Boroș, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Verginica Barbu Mititelu, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Corneliu Burileanu, Electronics and Telecommunications Faculty, University "Politehnica" of Bucharest
Mihaela Colhon, Faculty of Mathematics and Natural Science, University of Craiova
Ruxandra Cosma, Faculty of Foreign Languages and Literatures, University of Bucharest
Dan Cristea, Faculty of Computer Science, "Alexandru Ioan Cuza" University and Institute for Computer Science, Romanian Academy, Iași
Ștefan Daniel Dumitrescu, Institute of Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Daniela Gîfu, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
Corina Forăscu, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași and Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Gabriela Haja, "A. Philippide" Institute for Romanian Philology, Romanian Academy, Iași
Ion Iancu, Faculty of Mathematics and Natural Science, University of Craiova
Adrian Iftene, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
Diana Zaiu Inkpen, School of Information Tehnology and Engineering, University of Ottawa, Canada
Elena Irimia, 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, the United States of America
Vivi Năstase, Fondazione Bruno Kessler, Trento, Italy
Nicolae Panea, Faculty of Letters, University of Craiova
Ionuț Pistol, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
Elena Isabelle Tamba, "A. Philippide" Institute for Romanian Philology, Romanian Academy, Iași
Horia-Nicolai Teodorescu, Institute for Computer Science, Romanian Academy, Iași and Gheorghe Asachi Technical University of Iași
Amalia Todirașcu, Universitate Marc Bloch, Strasbourg
Dan Tufiș, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Adriana Vlad, Electronics and Telecommunications Faculty, University "Politehnica" of Bucharest and Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest

HONORARY PRESIDENT
Dan Claudiu Dănișor, Rector of the University of Craiova

ORGANISING COMMITTEE
Amelia Bădică, Faculty of Economics and Business Administration, University of Craiova
Costin Bădică, Faculty of Automation, Computers and Electronics, University of Craiova
Alex Becheru, Faculty of Automation, Computers and Electronics, University of Craiova
Anca Diana Bibiri, Faculty of Computer Science and Departament of Interdisciplinary Research - Human and Social Field, "Alexandru Ioan Cuza" University of Iași
Dumitru Bușneag, Faculty of Mathematics and Natural Sciences, University of Craiova
Mihaela Colhon, Faculty of Mathematics and Natural Science, University of Craiova
Nicolae Constantinescu, Faculty of Mathematics and Natural Science, University of Craiova
Mirel Coșulschi, Faculty of Mathematics and Natural Science, University of Craiova
Dan Cristea, Faculty of Computer Science, "Alexandru Ioan Cuza" University and Institute for Computer Science, Romanian Academy, Iași
Daniela Dănciulescu, Faculty of Mathematics and Natural Science, University of Craiova
Daniela Dincă, Faculty of Letters, University of Craiova
Corina Forăscu, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași and Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest
Lucian Gâdioi, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
Daniela Gîfu, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
Ion Iancu, Faculty of Mathematics and Natural Science, University of Craiova
Adrian Iftene, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
Petru-Adrian Istrimschi, "Alexandru Ioan Cuza" University, Iași
Nicolae Panea, Faculty of Letters, University of Craiova
Mădălin-Ionel Pătrașcu, Faculty of Computer Science, "Alexandru Ioan Cuza" University and Institute of Romanian Philology "A. Philippide", Romanian Academy, Iași
Dan Popescu, Faculty of Automation, Computers and Electronics, University of Craiova
Cecilia-Mihaela Popescu, Faculty of Letters, University of Craiova
Diana Trandabăț, Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
Cristiana Nicola Teodorescu, Faculty of Letters, University of Craiova
Dan Tufiș, Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest

TABLE OF CONTENTS

FOREWORD

The series of events organised by the Consortium of Informatisation for the Romanian Language (ConsILR) has reached its 10th edition this year. With a history that goes back to 2001, the ConsILR series of events has evolved in these 13 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, the event takes place at University of Craiova, an academic institution that shares the cultural, moral, scientific and educational European values.

The traditional organizers of the Conference *Linguistic Resources And Tools For Processing The Romanian Language* are the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iași and three institutes of the Romanian Academy: the Research Institute for Artificial Intelligence "Mihai Drăgănescu", Bucharest, the Institute for Computer Science of the Iași branch and the Institute of Romanian Philology "A. Philippide" of Iași. This year our event is held under the auspices of the Academy of Technical Sciences of Romania (ASTR) and it is locally organized by two institutions of University of Craiova: Department of Computer Science and Faculty of Letters.

Our conference has reached an anniversary edition, which in itself is significant because it demonstrates the keen interest for the computational study of Romanian. Besides the already established researchers, we gladly notice new and young researchers, seriously involved in this fascinating research and development area.

The itinerant CONSILR conference has been organised this year in Craiova and this is not by accident. The AI research group at the University of Craiova is very active and gained its leading position for more than a decade. We are very happy to be in this beautiful town with a vivid academic life and express our gratitude to the Rector of the University, the members of the Faculty of Automation, Computers and Electronics, the Faculty of Mathematics and Natural Science, Faculty of Economics and Business Administration and the Faculty of Letters, from University of Craiova, as well as to all other members of the Organising and Scientific Committees. Also, we would like to acknowledge and thank for all contributions, without which this volume would not exist. Special thanks are, obviously, due to our invited lecturers.

We think that the quality of the selected papers makes the present volume, alongside the volumes from previous editions, an interesting source of information on what is happening in Romanian natural language scientific and industrial community, a collection of articles very useful for researchers, AI, NLP and Linguistics professors and students and anybody who is concerned with language use, especially Romanian, in the electronic media.

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/2014/),

thanks to MEDIAEC – the Multimedia Laboratory of the "Alexandru Ioan Cuza University".

Iași, București, Craiova September 2014

The editors

CHAPTER 1
INVITED CONTRIBUTIONS

LANGUAGE TECHNOLOGIES IN HEALTHCARE

GALIA ANGELOVA

Linguistic Modelling Department
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Sofia, Bulgaria
galia@lml.bas.bg

Extended Abstract

Information Extraction (IE) is the dominating text analysis approach that is currently applied in the biomedical domain: due to the complexity of the narratives, shallow analysis is performed in order to extract automatically important entities, skipping the remaining text fragments. The IE systems operate further on the extracted text units relying on partial text understanding. Performance evaluation is done in terms of *precision* and *recall*, two widely-used indicators for the extraction accuracy and success. Here we present recent achievements in automatic IE from hospital discharge letters and outpatient records in Bulgarian language. Currently we deal mostly with records of diabetic patients, having in mind that diabetes is a chronical disease of major social importance. Several types of entities are essential for the IE tasks we tackle: *(i)* patient's principal diagnosis and diagnoses of the accompanying diseases; *(ii)* names of drugs, admitted by the patient, in particular those drugs that are discussed in free text descriptions with their dosage, frequency and route of admission; *(iii)* values of clinical tests and lab data that are documented in the patient records as free texts; *(iv)* patient status descriptions; *(v)* opinions of specialists concerning the patient status and diagnoses; *(vi)* family history and risk factors (e.g. smoking status, hypertony, etc.). Extracting these entities and events, the IE components need to cope with the negation. Timing of events is essential as well, so building timelines and temporal models is another important challenge for the medical IE applications.

Vocabulary. Starting with IE for Bulgarian patient records, we had to develop a relevant set of Bulgarian language resources: *medical terminology* including names of diseases (we used the Bulgarian version of ICD, the International Classification of Diseases), as well as *key phrases* that are used as typical names of entities and stable collocations. These phrases have to be learnt from the patient records' texts since they are not included in medical dictionaries. A language-independent extractor of phrasal units was developed (Boytcheva, 2012) that examines the frequency distribution of *N*-grams and suggests domain-specific collocations for final expert inspection. In this way meaningful collocations were learnt, like '*visible age'*, an important diabetic patient attribute, with its values *'corresponding to the real/passport/calendar one'*, *'about the real/calendar one'* and so on. Another important *attribute-value* pair is the *skin-characteristic* pair, as well as the *thyroid-status* pair, etc. This domain-specific vocabulary was acquired from the text of 6200 hospital discharge letters.

Grammar rules. In addition to the extraction of phrasal units, we needed to develop grammar rules for shallow analysis of text fragments that contain potentially interesting entities. In general, these rules are regular expressions that help the system to group alpha-numeric literals into meaningful text units. The extractor of lab data and clinical

test values analyses the paragraphs where these values are enumerated without predetermined order and without standardised names of the indicators (Tcharaktchiev et al., 2011). The extractor recognises at first the indicator (i.e. the *name of the tested characteristic*), as well as the *value* related to the corresponding indicator. The *measure* and *interval limits* are desirable features, and the *time*, *condition* and *explanation* of further details are optional features. The extractor copes with *(i)* the variety of name writings (abbreviations, omitted words in the name, joined words in the name, typos), *(ii)* various symbols used as separators, *(iii)* the varying format of the numeric values, *(iv)* arbitrary replacements of Cyrillic and Latin letters which look identical and *(v)* ambiguity in the lab data recognition and the scoping of phrases related to certain indicators that have specific values. As an illustration, we show a rule for packing tokens into a structural group:

<(> <n> <v> <s> <v> <)> => <N> which means the following:
Find a sequence of tokens which:
starts with '('
followed by a phrase signalling referential values <n>,
followed by a number <v>,
followed by a separator <s>,
followed by a number <v>,
followed by a ')'.
If all tokens occur in the given order than this expression defines the group <N>.

This simple rule is used to group the literals *'(norm – 8,7-42)'* in the text fragment *'testosterone -3.2 (norm – 8,7-42)'*. The rule has 18 variants reflecting the various separators and delimiters learnt from a training set of 1000 discharge letters; it is used for the shallow analysis of lab data in 6200 hospital discharge letters of diabetic patients.

Similarly, using rule-based shallow analysis, values of blood sugar are extracted (Nikolova, 2012), as well as names of drugs, dosage, frequency and admission route (Boytcheva, 2011) and temporal expressions in Bulgarian discharge letters (Boytcheva et al., 2012). Negation in the Bulgarian clinical texts is also treated by shallow IE analysis (Boytcheva et al., 2005).

Corpora of medical records. Our first corpus of clinical narratives contains 6200 anonymised discharge letters of patients diagnosed with diabetes and other endocrinal diseases, provided by the University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev" (USHATE), Medical University Sofia, Bulgaria. These letters are texts with length of 2-3 pages. Due to centralised national regulations, they consist of predefined sections like *Diagnoses*, *Anamnesis (Case History)*, *Patient Status*, *Lab data & clinical tests* and *Debate* which are available in 100% of the records in the corpus. Discharge letters are written in telegraphic style, discussing mostly positive findings, and can be treated more successfully by shallow analysis instead of full sentence parsing. Latin terms are commonly used, written in Latin alphabet (3% of all words in the corpus) or transliterated to Cyrillic alphabet (34% of all words). Spelling errors are common, too.

The current corpus with patient records contains more than 37.9 million pseudonymised reimbursement requests (outpatient records) submitted to the National Health Insurance

Fund (NHIF) in 2013 for more than 5 million patients, including 436000 diabetic ones. These records are semi-structured files with predefined XML-format, produced by the General Practitioners (GPs) and the Specialists from Ambulatory Care for every contact with the patient. They contain sufficient text explanations to summarise the case and to motivate the requested reimbursement, so IE is again the best analysis tool. Most important patient indicators like *Age*, *Gender*, *Location*, *Diagnoses* are easily seen since they are stored with explicit tags. The *Case history* is presented quite briefly in the *Anamnesis* section as free text with description of previous treatments, including drugs taken by the patient beyond the ones that are to be reimbursed by NHIF. The values of *Clinical tests and lab data* are enumerated in arbitrary order as free text in another section. The *Prescribed treatment* is described under a special tag. Only the drugs prescribed by the GPs and reimbursed by the NHIF are coded, the other medication is described as free text. Latin terms are relatively rarely used and spelling errors are rare too, compared to the discharge letters written as hospital documentation. An outpatient record might include about 160 tags.

Applications. Our initial tasks in research projects were oriented towards extraction of isolated entities and their attributes, such as status of patient skin, neck, thyroid gland, limbs and patient age (Boytcheva et al., 2010). These IE prototypes achieved accuracy of 83-92%.

Automatic recognition of temporal expressions helped to identify the drugs admitted by the patient at the moment of hospitalisation, i.e. at hospital stay Day 0 (Boytcheva at al., 2011). The extractor recognises about 350 drugs with accuracy of 90.17%, which are not prescribed via the Hospital Pharmacy, but are taken by the patients during the period of hospitalisation. This adds value to the data recorded in the Hospital Information System, where Day 0 is typically not reflected, and helps to search for Adverse Drug Events.

In (Boytcheva and Angelova, 2012) we present a prototype that constructs timelines of events that are described in the *Anamnesis* of hospital discharge letters; there are two timelines organised for the absolute and relative temporal markers. All clauses between two temporal markers are called "an episode" where drugs, diagnoses and patient conditions are recognised with accuracy higher than 90%.

The drug extractor, initially tested on discharge letter texts (Boytcheva, 2011), is elaborated to tackle drugs in the outpatient records. For diabetic patients, currently the extractor handles 2239 drugs names included in the NHIF nomenclatures. Recent extraction experiments deal with large-scale analysis of the outpatient records of 33641 diabetic patients. The drug extractor finds in the *Anamnesis* section drug names, daily dosages, frequency and route of admission with precision 95.2% and recall 93.7%, and replaces the drug names by their ATC[1] codes.

The extractor of values of lab tests and clinical examinations was elaborated too and now it copes successfully with the outpatient record texts. Major difficulties encountered in the entity recognition are due to the large variety of expressions

[1] Anatomical Therapeutic Chemical (ATC) Classification System for the classification of drugs, see http://www.who.int/classifications/atcddd/en/

describing the laboratory tests and clinical examinations. It tackles more than 40 types of clinical tests and works with precision that exceeds 98%.

Our recent efforts are focused on extraction of entities from the outpatient record Repository, in order to facilitate the construction of the Bulgarian diabetic register (Nikolova et al., 2014). Integrating language technologies and a business intelligence tool, we discover potential diabetic patients who are not formally diagnosed with diabetes. Diabetes can be suggested, for instance, by the values of the patients' clinical tests, or because the patients admit drugs that treat higher levels of blood sugar, or because some medical specialist has doubts that the patient might have diabetic complications (e.g. diabetic retinopathy). The Repository of outpatient records is pseudonymised, i.e. we can track the multiple visits of the same patient to his/her GP, hospital, etc. In addition, we can identify the important risk factors and the family history, in case they are described in the outpatient record text. Finally, the records of potential diabetic patients are classified according to the hypothesis "*having diabetes*" with precision 91.5% and the findings are delivered to the medical authorities for further checks. Discovering several hundreds of potential diabetic patients shows the importance of automatic text analysis in large medical Repositories.

Acknowledgements

The contribution summarises the results jointly obtained with Dr. Svetla Boytcheva, Dr. Dimitar Tcharaktchiev MD, Ivelina Nikolova, and Dr. Zhivko Angelov. The research work presented here is partially supported by the projects *EVTIMA* DO 02-292 (Effective search of conceptual information with applications in medical informatics), funded by the Bulgarian National Science Fund in 2009-2012; *PSIP* (Patient Safety through Intelligent Procedures in Medication) grant agreement 216130, funded by FP7 ICT in 2010-2011; and *AComIn* (Advanced Computing for Innovation) grant agreement 316087, funded by FP7 Capacity in 2012-2016.

References

Boytcheva, S. (2011). Shallow Medication Extraction from Hospital Patient Records. In: Koutkias, V. et al. (Eds.), *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, IOS Press, *Studies in Health Technology and Informatics* series 166, 119-128.

Boytcheva, S. (2012). Structured Information Extraction from Medical Texts in Bulgarian. *Cybernetics and Information Technologies 12:4*, 52-65, available at http://www.cit.iit.bas.bg/CIT_2012/v12-4/4Boicheva4-2012-Gotovos.pdf

Boytcheva, S., Angelova, G. (2012). A workbench for temporal event information extraction from patient records. *Proceedings of the 15th Int. Conference on Artificial Intelligence: Methodology, Systems, and Applications AIMSA 2012*, Springer, *Lecture Notes in Artificial Intelligence* 7557, 48-58.

Boytcheva, S., Angelova, G., Nikolova, I. (2012). Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the*

Association for Computational Linguistics, Avignon, France, 77-81, available at http://www.aclweb.org/anthology-new/E/E12/E12-2016.pdf

Boytcheva S., Nikolova, I., Paskaleva, E., Angelova, G., Tcharaktchiev, D., Dimitrova, N. (2010). Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records. In: V. Fomichov (Ed.), *Special Issue on Semantic Technologies, Informatica (Slovenia)*, Issue 4, December 2010, 269-278.

Boytcheva, S., Strupchanska, A., Paskaleva, E., Tcharaktchiev, D. (2005). Some Aspects of Negation Processing in Electronic Health Records. *Proceedings of the Int. Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries*, associated to RANLP-2005, Borovets, Bulgaria, 1-8.

Boytcheva, S., Tcharaktchiev, D., Angelova, G. (2011). Contextualization in automatic extraction of drugs from Hospital Patient Records. In A. Moen at al. (Eds) *User Centred Networked Health Case*, Proceedings of MIE-2011, the 23th Int. Conf. of the European Federation for Medical Informatics, Norway, IOS Press, *Studies in Health Technology and Informatics* series 169, 527-531.

Nikolova, I. (2012). Unified Extraction of Health Condition Descriptions, *Proceedings of the NAACL HLT 2012 Student Research Workshop*, Montreal, Canada, 23-28, available at http://aclweb.org/anthology-new/N/N12/N12-2005.pdf

Nikolova, I., Tcharaktchiev, D., Boytcheva, S., Angelov, Z., Angelova, G. (2014). Applying Language Technologies on Healthcare Patient Records for Better Treatment of Bulgarian Diabetic Patients. In: G. Agre et al. (Eds.): *Proceedings of AIMSA 2014*, Springer, *Lecture Notes in Artificial Intelligence* 8722, 92–103.

Tcharaktchiev, D., Angelova, G., Boytcheva, S., Angelov, Z., Zacharieva, S. (2011). Completion of Structured Patient Descriptions by Semantic Mining. In: Koutkias, V. et al. (Eds.), *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, IOS Press, *Studies in Health Technology and Informatics* series 166, 260–269.

ENUNCIATIVE ATTITUDES IN ROMANIAN TOTALITARIAN DISCOURSE

CRISTIANA NICOLA TEODORESCU

Faculty of Letters, University of Craiova, Romania
cteodorescu05@yahoo.fr

1. Introduction

The years 1947-1989 have marked Romanian society and Eastern European societies in general, by means of a political and ideological system transposed at the linguistic level through what is nowadays known as "wooden language".

How is wooden language seen in Romanian communist totalitarian discourse of 1948-1989? Which are the corresponding structures, discursive forms and their typology? These are some of the questions we aim at providing an answer to.

Research Objectives. Our objective is twofold: *linguistic*, for the comprehension and description of wooden discursive mechanisms and, at the same time, *therapeutical*, with a view to discarding or at least mitigating a harmful ballast of Romanian language and not only.

Premises. While attempting to discover the characteristics of Romanian wooden language, we agree with Tatiana Slama-Cazacu, who defines it as "a sub-system of a language, including mostly lexical elements, but also idioms, fixed expressions, clichés, with a determined meaning in the context of a certain 'authority', used to a wide extent in a stereotypical and dogmatic manner, as the expression of an ideology (or a simulacrum of ideological, economic, technological, political, cultural, etc. sub-systems that have a certain power or authority), imitated, but also imposed by the political power […] and then disseminated by repetition, by means of frequent usage in the various oral or written media, thus annihilating the recipients' reflection, who end by surrendering to a collective suggestion, with the real intent or at least the achieved effect being to impose authority […], to implant a different way of thinking and, in general, to hide, to mask the true reality, if it is not favourable" (Slama-Cazacu, 1991: 4). We start from the feeling that Romanian communist totalitarian discourse, as expressed in the analysed period, is dominated by predictability, an institutionalisation of clichés, a huge torrent of words, an obsessive stereotypy, whereby the "already said" becomes an anesthetising and hypnotic repetition. Its evolution seems to be marked by incommunicability, by the depreciation of words, which progressively lose their "sovereign virginity and dignity and prematurely enter the darkest routine" (Negrici, 1978: 110). In the analysis of forms of Romanian totalitarian language, we start from the hypothesis that we are dealing with an opacifying discursive phenomenon, whereby, despite the promoted egalitarian values, the enunciator has a position of force, ignoring the recipient, despising him/her and, what seems to be the worst, treating him/her as an object.

Questions for Analysis. Our analysis and interpretation is built around the following questions: how can the enunciator of such discourse have a discursive behaviour that ignores the recipient's voice? How can we explain the astonishing "loneliness" of this type of discourse, despite its generalisation by means of compulsory quotes in other types of texts (literary, scientific, didactic, etc.)? What distance and tension does the

enunciator of the political discourse establish, by means of the intermediary of his/her utterance, between himself/herself and his/her message, between himself/herself and the world? Which are the linguistic facts mobilised by the enunciator, that result in the opacification of the message, in the establishment of a non-cooperating discursive inequity between discursive partners? Which is the discursive typology of analysed texts?

2. Theoretical framework

We have chosen discourse linguistics as an interpretative paradigm, since the reintroduction of "man into language" and of the subject into enunciation opens the difficult issue of the enunciator's position and behaviour and results in Strawson's axioma: "One cannot expect to understand language, as theoreticians expect to, if one does not understand discourse. One cannot expect to understand discourse if one does not take into account the purpose of communication". Within the extremely wide framework of discourse linguistics, we have chosen the theoretical and methodological approach provided by the French School of Discourse Analysis, a generous subject, due to its openness towards history, ideology and politics.

2.1. The corpus

The corpus subject to the compared analysis includes three texts:

Gheorghe Gheorghiu-Dej, *Political Report of the Central Committee to the Congress of the Romanian Workers' Party*, February 21, 1948;

Nicolae Ceauşescu, *Report of the Central Committee of the Romanian Communist Party Regarding the Party's Activity between the 8th and 9th Congresses of the Romanian Communist Party,* July 19, 1965;

Nicolae Ceauşescu, *Report of the Central Committee of the Romanian Communist Party Regarding the Current Status of Romanian Society, the Activity of the Central Committee between the 13th and 14th Congresses, the Accomplishment of the Programme for Economic and Social Development during the 9th Five-Year Term and in Perspective, until 2000-2010, for the Secure Accomplishment of the Construction Programme of a Multilaterally Developed Socialist Society and Romania's Advancement Towards Communism*, November 20, 1989.

These are verbally presented written discourses, defended every five years in front of delegates to the congresses, which were fully reproduced in contemporary press and have become extended sources of subsequent quotes and replays. The enunciative source of these texts is the party's general secretary, who proposes an analysis of the previous period and an orientation for the years to come, an ideological orientation approved unanimously, in an enthusiastic manner, by the representatives of the people and, implicitly, by the entire population.

2.2. The analysis concepts

The analysis concepts we propose are ***the concept of distance*** and that ***of tension*** (Dubois, 1969). According to Jean Dubois, *the concept of distance* can help us analyse the attitude of the speaker regarding his/her utterance, i.e. the relative distance established by the speaker between his/her utterance and himself/herself. The speaker may fully assume his/her utterance. In this case, we are dealing with an identification between the I, subject of the utterance, and the I, subject of the enunciation. On the contrary, the distance may be maximal and, in this case, we would be dealing with a didactic discourse, where the I become the formal impersonal of the utterance. *The concept of tension* helps establish the relationship between the speaker and his/her interlocutor through the intermediary of the utterance. This process expresses what Luce Irigaray refers to as "desire of communication". The analysis of personal pronouns, adjectives and possessive adjectives, the analysis of adverbs, of enunciative verbs, of injunctive and negative structures can provide us with relevant information on the manifestations of the two selected theoretical concepts, i.e. distance and tension. In this article, we shall only present the analysis of personal pronouns, adjectives and possessive adjectives, a range of factors able to prove the construction of a certain type of political discourse.

2.3. Elements of lexicometry

Our analysis implied handling a wide amount of data, so that quantitative aspects may be relevant and subjective appreciations should be minimized. In the description and analysis of the language forms used by Gheorghe Gheorghiu-Dej and Nicolae Ceauşescu as political enunciators, we have used computer-based lexicometry, i.e. the software *Conc: Concordance Generating Program* (version 1.70 beta, 1992), created by John Thompson of the Summer Institute of Linguistics, Academic Computing Department, Dallas.

3. Relations with the surrounding environment and individuals: the category of person

Having or not having the floor or, more accurately, in discourse, showing or not showing that the enunciator assumes his/her words are elements of interest for our analysis: we want to know who speaks in Romanian communist discourse, on behalf of whom and who he/she speaks to. This vocation of inter-individual link that defines language functioning is primarily translated through the linguistic expression of the category of person. To this purpose, Anna Jaubert (1990: 9) shows that the linguistic expression of the category of person can range from the fullest deletion of communication participants (i.e. stories) to their more or less obvious presence (i.e. speeches), where I and YOU construct, by means of their mutual relation, the figurative framework of enunciation. Saying "I" implies an expression of one's position as enunciator and, by default, an assertion of the existence or presence of a recipient, YOU, liable to accept or enter communication.

Analysing in our corpus the functioning of personal substitutes[1], we shall interpret the relations established, by means of the intermediary of the utterance, between the enunciator and the surrounding world, between the enunciator and his/her recipient. Moreover, we shall also establish the relations of inclusion or exclusion into/from a group and/or the surrounding world and, implicitly, the phenomena of distancing and tension seen at the level of the utterance.

3.1. A Quantitative Approach

A brief quantitative analysis of the forms of personal pronouns appearing in our corpus shows:

In GGD 48, personal pronouns in the 1st person can be seen quite strongly (64 occurrences in singular and plural). The consideration of the communicative partner (2nd person, singular and plural) is very low (4 occurrences). References to the surrounding environment or other individuals are highly marked by the use of the personal pronoun of the 3rd person, singular and plural, with 207 occurrences.

In NC 9, the quantitative situation of personal pronouns is slightly changed: the enunciator does not manifest himself at the discourse level (a lower number of forms in the 1st person, singular and plural – only 5 occurrences, a lower number of forms in the 2nd person – a single occurrence, with 153 occurrences of pronouns for absent individuals).

In NC 14, the enunciator expresses his presence again (11 occurrences for the 1st person, singular and plural), with the interlocutor being completely absent (no occurrence of the 2nd person, singular and plural). The category of absent persons is not more significant either: the number of 44 occurrences is lower than the numbers present in the two other texts.

For the Romanian language, the forms of personal substitutes are not enough to measure the distance set by the enunciator between himself and the recipient or between himself and the environment. For the analysis of such relations, the number of personal verbs in the 1st and 2nd person (singular plural) with implicit subject is highly relevant, since, thus, we shall have a global image of the tension and/or distance established between discursive positions:

Table 1: Global image of the tension and/or distance established between discursive positions

Implicit subject	**1st person, singular**	**2nd person, singular**	**1st person, plural**	**2nd person, plural**
GGD 48	3	2	95	5
NC 9	1	-	82	5
NC 14	56	1	281	1

[1] See Pierre Achard's sociolinguistic analysis of the functioning of personal pronouns in the document of the Congress of People's Deputies of the Soviet Union, 1989, « Registre discursif et énonciation : introduction sociolinguistique à partir des marques de personnes. Le Congrès des Députés du Peuple d'URSS en 1989 », in *Langage et société*, No. 71, 1995, pp. 5-34.

We shall corroborate the two categories of obtained data, in order to present the textual expression of the category of person in the analysed texts. It can be seen that the category of person, reduced to the 1st and 2nd person, singular and plural, according to the conception of Benveniste, is expressed at a textual level either through the specific forms of personal pronouns, or through verbal endings. The compared situation of the three analysed texts is the following:

Table 2: Compared situation of the three analysed texts

Text/person	I	YOU (singular)	WE	YOU (plural)
GGD 48	6	2	156	9
NC 9	2	-	86	9
NC 14	57	1	291	1

The forms for absent persons (personal pronouns of the 3rd person, singular and plural) have decreasing values in the analysed texts.

Table 3: Forms for absent persons

Forms for non-person	Occurrences
GGD 48	207
NC 9	153
NC 14	76

The forms for absent persons have positively (economy, production, Marxist-Leninist theory, literary criticism, constitution drafts, the party, the socialist state, etc.) or negatively (imperialist politics, the royal family, the capitalist system, small manufacturers, American imperialists, aggressive forces, the vassals of capitalists, etc.) connoted content.

3.2. Interpretation of quantitative data

Analysing the obtained numbers, we remark the insistence of discursive expression of the enunciator of the text NC 14, unlike the other texts. What most strikes is the insistence shown by the enunciators of the *Reports* in hiding themselves behind a collective WE that dilutes responsibility and also marks indifference towards the recipient, who is not directly addressed within the discourse.

All these data indicate, on the one side, the maximum distance established by the enunciator of the communist discourse between himself and the environment and, on the other side, the minimum tension between the enunciator and the recipients of this type of discourse.

Table 4: Interpretation of quantitative data

Forms	**Discursive value**	**Discursive effect**
1st person, singular	The subject of the utterance is absent	Maximum distance
2nd person, singular	Non-differentiated and non-individualised, global consideration of the interlocutor	Refusal of individualisation, maximum distance, absent tension
3rd person, singular	Various contents, positive or negative connotations	Desire to include and uniqueness of the group
1st person, plural	Ambiguous contents (We = the general secretary of the party = a plural of majesty?, the Central Committee, all communists, the entire people…)	I becomes a collective WE. Transparency, didacticism, desire of full inclusion in the improvement-oriented group of communists
2nd person, plural	Ambiguous contents (all communists, all the delegates to the congress, the entire people…)	Minimum distance, absent tension, refusal of individualisation
3rd person, plural	Various contents, positive or negative connotations	Desire to include and uniqueness of the group

4. The Globalising Obsession of Possession: Possessive Adjectives

This desire of globalisation, of full inclusion within the improvement-oriented group of communists, of expression of its uniqueness and force, also resorts to the intensive use of possessive adjectives, interpreted by Patrick Charaudeau as "a procedure of establishing the more or less hierarchized dependence of an entity from a person" (Charaudeau, 1992: 203). The analysis of the frequency of possessive adjectives indicates an inflationist usage.

The forms *our*, *ours (nostru, noastră, noştri)* appear as follows in the analysed texts: GGD 48 = 156, NC 9 = 19 and NC 14 = 276 occurrences.

The process of establishing the dependence of an entity from a person can refer to all kinds of nouns:

[+concrete, -animated, +designable] *our* cities, *our* villages…

[+human, ±collective] *our* working class, *our* people, *our* talented creators…

[-concrete, -animated, +designable] *our* country, *our* party, *our* system, *our* homeland…

In this vision, ALL becomes OURS: the homeland, the people, the party, the achievements, the cause, the pride… as the absence of responsibility and commitment of social actors is a feature of this type of discourse. Due to this globalising obsession of possession, the tension between communicative poles is minimised, and the enunciator

actually misses one of the discursive goals he has aimed at, i.e. increasing the commitment of social actors, resulting in a huge increase of the gap between words and reality, specific to wooden language. This obsession of globalisation does not leave room for tension and contrary arguments, since the place of others is only considered in terms of agreement and obedience.

4.1. The Adjectival Delirium

Another element that can provide clues on the distance and/or tension established by the intermediary of the discourse between the enunciator, the recipient and the referential world is represented by subjective adjectives. Since "everything is relative in the use of adjectives", Catherine Kerbrat-Orecchioni (1980) distinguishes among several categories of adjectives: objective and subjective, affective and evaluative subjective adjectives and axiological and non-axiological evaluative adjectives.

Trying to analyse the adjectives (Kerbrat-Orecchioni, 1980: 83-84) (actual adjectives and adjectives proceeding from verbal participles) appearing in the analysed texts from this perspective, we first observe their high frequency and the preference for subjective adjectives.

Affective adjectives "enounce at the same time a property of the object they determine and an emotional relation of the speaker regarding such object. To the extent that they involve an affective commitment of the enunciator, to the extent that they express his/her presence in the utterance, they are enunciative" (Kerbrat-Orecchioni, 1980: 86).

Such adjectives seldom appear in the analysed texts, precisely because they are excluded from texts tending towards objectivity. Very few examples in NC 9 and NC 14: *dear* comrades, our *beloved* homeland.

Non-axiological evaluative adjectives represent a very wide class that, according to Kerbrat-Orrecchioni's description, includes "all adjectives which, without enouncing judgments of value or affective commitment of the locator (at least in terms of strict lexical definition; of course, in context, they can acquire affective or axiological nuances), involve a qualitative or quantitative evaluation of the object denoted by the noun they determine" (Kerbrat-Orecchioni, 1980): *historical* importance, an *impetuous* rhythm, *great* successes…

If the very low number of affective adjectives can be explained by the desire of objectivity, we explain the high frequency of non-axiological evaluative adjectives by the fact that their use is related to the (subjective) idea of the locator regarding the norm of evaluation of a given category of objects.

A lexical phenomenon, interesting in terms of its high frequency in the analysed corpus, is also included in this category: *the expression of intensity by means of adjectives with the prefix ne-*: *nemijlocit, nezdruncinat, nedeghizat, nestingherit, neţărmurit, neclintit, neabătut*[2]*...*

Axiological evaluative adjectives "are doubly subjective", expressing a positive or negative judgment of value on the object denoted by the noun they determine. This is a very interesting class for our analysis, precisely for this double subjectivity appearing

[2] direct, unbreakable, unconcealed, free, infinite, unconditional, immovable…

due to their use, that varies according to "the specific nature of the subject of the enunciation, reflecting his/her ideological competence" (Kerbrat-Orecchioni, 1980: 91) ; to the extent they express the speaker's favourable or unfavourable position towards the designed object.

The corpus analysis outlines the high frequency of this type of adjectives: a *simple* decision, the *true* objectives, a *large* action, the *great* experience, a *revolutionary* attitude, the *logical* response, *deep* changes…

The axiological adjectives found in the analysed corpus express the speaker's favourable or unfavourable position towards the designed object. It is he who appreciates the objective reality, according to a personal scale of values. A deeply subjective appreciation which, in the case of a text that aims to be objective and analytical, is deeply manipulative, by imposing a single evaluative scale.

The intensive use of axiological and non-axiological evaluative adjectives marks the maximum distance established by the enunciator between himself and the actual world, as the enunciator positions himself within the discourse as the single entity able to judge reality.

Another means of increasing distance is the opacifying game of adjectival anteposition. Even though in Romanian the normal, non-emphatic, order of words favours the postposition of adjectives, we can see a clear tendency towards adjectival anteposition, even when such adjectives express a non-gradable quality and when they determine nouns marked as [-human]. The phenomenon, remarked by Maria Manoliu-Manea (1993) in the Romanian press of the communist period, seems to be characteristic for Romanian communist language. We can see the recurring apparition of combinations such as: Adj. + N: the *simple* decision, the *fake* socialist Bevin, the *true* objectives, the *true* light, the *great* victory, the *great* Lenin, the *great* socialist power...

We can see in the texts NC 9 and NC 14 a preference for doubly anteposed adjectival combinations and the expression of absolute superlative by repetition of the same adjective: Adj. 1 + Adj. 2 + N: a *new and strong* development, *the most correct and most advanced* society, an *extended and prolonged* debate…, Adj. 1 + Adj. 1 + N: *new and new* conditions/forms, achievements, *new and new* questions…

In the *Reports* to the 9th and 4th Congresses, we can see the tendency of double and even triple adjectival determination, as if a single adjective were not able to express all the qualities of an object. The obsessive-explanatory desire of enunciators is so marked, that this double/triple adjectival determination results in frequent pleonastic effects and striking hyperboles: N + Adj. 1 + Adj. 2: a *modern, developed* industry; a *modern, scientific* organisation; a *better, equal* life…

If one wants to establish the collocations between adjectives and the nouns they are related to, one can see the existence of some adjectives highly open to a large number of combinations (*large, powerful, new, entire…*), but also the existence of some adjectives reserved for certain combinations marked as [+affective]: *immense* + love, trust, *deep* + attention, satisfaction, *particular* + attention, effort, results, *powerful* + development, *great* + victory, achievements, success, attention, responsibility, objectives, productivity, *high* + rhythm, level, productivity…

This preference for adjectival multiplication, this "Stakhanovism of glamorous adjectival multiplication" (Negrici, 1995: 12), reminds us of Louis Veillot's remark regarding Jules Favre's discourse of admission to the French Academy: "He cannot do without adjectives, he always exhibits them in pairs, frequently in triads and even in clusters, and only omits the accurate one" (Marouzeau, 1950: 142). This adjectival delirium is seldom, as shown by Marouzeau, an effective means of reinforcing a certain expression, and most frequently results in a dispersion of attention. We consider that this preference for adjectival anteposition and for double, triple adjectival determination is one of the most important marks of the enunciator's subjective attitude, who imposes his own vision of the world and reality, as well as a violation of the axioma of Quality (Grice, 1979), by orienting the recipient's attention towards the "quality" of the object, not towards the object itself. The distance separating communicative actors is highly increased.

5. Conclusions

All these data show that the Romanian communist discourse is constructed in a *didactic* manner, establishing a *maximum distance* between the enunciator and the actual world and *minimum tension* between the poles of political communication.

Tabel 5: Maximum distance and minimum tension in the Romanian communist discourse

Concepts of analysis	Content elements	Discursive effects	Type of achieved discourse
Maximum distance	- practically, the enunciator does not appear in his discourse - inflationist usage of the collective WE - non-axiological and axiological evaluative adjectives	- absence of the subject of the enunciation - refusal of individualisation - absence of responsibility - desire of inclusion in the improvement-oriented group of communists - homogenisation of the group - obsession of possession	- didactic political discourse - scarcely marked enunciation
Minimum tension	- low dynamics of the pair WE/YOU	- ignored recipient, absence of co-enunciation phenomena - manipulative tendencies - a position of submission of the others' voice, who only have to accept, by reiteration, the discursive positions of enunciators	- didactic political discourse

We can say that Romanian communist totalitarian discourse has a uniform, rigid, self-sufficient structure, resistant in time and also in space. A structure that under certain conditions (single Power) becomes pathological.

References

Achard, P. (1995). Registre discursif et énonciation : introduction sociolinguistique à partir des marques de personnes. Le Congrès des Députés du Peuple d'URSS en 1989. *Langage et société*, 71, 5-34.

Benveniste, E. (1996). *Problèmes de linguistique générale.* Paris: Gallimard.

Kerbrat-Orecchioni, C. (1980). *L'énonciation de la subjectivité dans le langage*. Paris: Armand Colin.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Paris: Hachette Education.

Dubois, J. (1969). Enoncé et énonciation. *Langages*, 13, 100-110.

Grice, H.P. (1979). Logique et conversation. *Communication*, 30, 57-72.

Jaubert, A. (1990). *La lecture pragmatique*. Paris: Hachette Supérieur.

Kerbrat-Orecchioni, C. (1980). *L'énonciation de la subjectivité dans le langage*. Paris: Armand Colin.

Manoliu-Manea, M. (1993). Metamorfozele timpului în discursul politic românesc. In *Gramatică, pragmasemantică şi discurs*, Bucharest: Litera, 248-253.

Marouzeau, J. (1950). *Précis de stylistique française.* Paris: Masson et Cie Editeurs.

Negrici, E. (1978). *Figura spiritului creator*. Bucharest: Cartea Românească.

Negrici, E. (1995). Gugumănii înaripate. *România Literară*, 40, 12.

Slama-Cazacu, T. (1991). "Limba de lemn". *România Literară*, No. 24, Year XXIV.

Thom, F. (1978). La langue de bois. *Commentaire Julliard*, Paris.

MACHINE TRANSLATION – A LOOK INTO THE FUTURE

DAN TUFIȘ

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
tufis@racai.ro

Abstract

The article reviews some of the major challenges of Machine Translation brought up by the data-driven revolution and open-source initiatives. It presents a short-term vision on the future Translation as a Service (TaaS) based on recent research results and modern trends in the management and exploitation of massive volumes of data.

1. Introduction

Machine Translation is one of the oldest dreams in the history of computer science with its first traces in the investigations carried out, as early as 1946, by the computer scientist A.D. Booths, who, in a memorandum dated February 12, 1948, wrote:

"A concluding *example, of possible application of the electronic computer, is that of translating from one language into another. We have considered this problem in some detail and it transpires that a machine of the type envisaged could perform this function without any modification in its design.*" (Weaver, 1949)

The generally recognized "baptist" of the domain is Warren Weaver who authored the manifest document called "Translation" (Weaver, 1949) in which he raised interesting questions and suggested "four types of attack" of the future research and development on machine translation. His statement is of perfect actuality: "and it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken, as a necessary preliminary step" (ibid).

The history of Machine Translation had ups and downs, but beginning with the early 1990s, the statistical models developed at the IBM TJ Watson Research Center (Brown et al., 1993) and implemented in their Candide project started a resurrection of the scientific and commercial interests in this area. The spectacular progress in information and communication technology as well as the larger and larger available electronic bilingual data sets opened unprecedented possibilities to statistically "attack" the thorny problem of automatic translation.

2. The Data-Driven Revolution

The last 20 years have seen an incredible progress in speed, quality and volumes of translated documents. In 2012 Franz Och, the head of Translation Development at Google, had a nice post on Google's blogspot[1]: "*In a given day we translate roughly as much text as you'd find in 1 million books. To put it another way: what all the*

[1] http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html

professional human translators in the world produce in a year, our system translates in roughly a single day."

More and more open source software programs have become available (parallel data crawlers, aligners, decoders, optimizers) as building blocks for new translation systems. The release of the APIs for the major web translation systems (Google Translate and Bing) supported translation projects all over the world, so that, practically in all countries (at least in EU), programs (many of them funded by EC) were launched to create infrastructural language and tools resources for building translation systems covering as many language pairs as possible. A fundamental role in the European boost of MT research and development should be undeniably attributed to the open-source statistical MOSES MT engine, with a first version released in 2007 and continuously developed since then (Koehn et al., 2007).

As most users already noticed, the quality of Web translations, irrespective of the service providers (Google, Microsoft, Yahoo, Language Weaver-SDL, Yandex and many others[2]), differs significantly, depending on the language pair and the type of text (we do not take into account here the literary texts). Why does this happen? A quick answer will point to the disparity among the language resources for the language pair and/or the domains of documents to be translated, as well as to the limitation of the current processing algorithms. This explanation is valid irrespective of the translation model (rule-based, data driven or hybrid ones). The disparity among the language resources does not refer only to the quantity of linguistic data, but also to the linguistic idiosyncrasies. Take, for instance, the honorifics system or the pro-drop character of one language, absent from the other language. Only these cases will create huge problems in an accurate translation, requiring access to large context (beyond one sentence - as current technology considers) and possibly to socio-cultural knowledge. On the other hand, translation of a technical or a legal document (supported by adequate terminological resources) is easier than translation of an open-domain document (e.g. an editorial in a news magazine). This is because in the first case, in general, language is more controlled (has a lower perplexity) than in the latter case.

A long standing slogan, especially in the data-driven MT community is that "more data is better data", thus, significant efforts have been invested in collecting huge quantities of web data, both monolingual and multilingual. The "web as corpus" initiative, whose evangelist is Adam Kilgarriff (2001), had and still has a major impact on the methodologies and algorithmic developments for MT. The lack of sufficient parallel data for most of language pairs generated a keen interest in collecting comparable corpora and extracting from them parallel or quasi-parallel pieces of text (Rapp, 1995; Rapp, 1999; Diab and Finch, 2000; Fung and Yee, 1998; Koehn and Knight, 2000; Munteanu and Marcu, 2002; Skadiņa et al., 2012). The huge available web data is extremely heterogeneous in terms of formats and language quality and new investigation directions were opened and thoroughly addressed: designing language-pair and topic oriented crawlers, cleansing large volumes of collected data (boiler-plate

[2] See, for instance, http://en.flossmanuals.net/open-translation-tools/ch035_web-translation-systems/

tools, recognizing MT translated texts[3], balancing text types), building language modellers able to deal with very large quantities of data and new types of decoders (factored phrase-based and/or syntax based), implementing new optimisation techniques, domain adaptation of language and translation models, translation quality evaluation, translation quality estimation, combining human post-editing with real-time SMT models adaptation, and many others. The processing flows implementing a translation service, from beginning to the end, become more and more sophisticated. For instance, pre-processing modules linguistically annotate the input texts, take care of the true casing or re-order the source words to follow as much as possible the natural word-order in the target language, while post-processing modules may take care of re-establishing the correct casing, correcting some translation errors (as in cascaded translation approaches (Ehara, 2011; Tufiș and Dumitrescu, 2012), update the language and translation models.

The current state-of-the-art in MT brings together text and speech input/output and all the known problems in text translation are multiplied by the problems in speech recognition and speech synthesis. High accuracy of free speech recognition (mainly in noisy environments[4]) and the naturalness of the synthesized speech are just two major milestones for present and future research, the achievement of which are not so far away. The task of MT advancement towards the objective of Fully Automatic High Quality Machine Translation (FAHQMT), text & speech, remains one of the most difficult (if not the hardest) problem of computer science. Yet, the last 20 years of constant improvements in ICT technology power and prices (e.g. cheap computational power available via GPGPUs – General-purpose computing on graphics processing units), the availability of vast amounts of electronic data, open source platforms for experimenting new ideas in MT, significant progress in Artificial Intelligence (especially in Statistical Machine Learning) and the keen interest of the funding agencies and commercial companies are sufficient arguments to believe that FAHQMT will be a reality in the near future.

3. Modern Trends

One of the lessons learnt from the empirical research and many conducted experiments, including ours (Tufiș and Dumitrescu, 2012; Dumitrescu et al., 2013; Boroș et al., 2013; Tufiș et al., 2013a; Tufiș et al., 2013b), was that the quality of statistical MT (SMT) depends to a great extent on the match between the training data and translated texts in terms of terminological overlap, subject domain, genre and text type. That is to say, the use of translation models learnt from clean texts similar to the ones supposed to be translated is one of the best ways to increase the translation quality. Given that the Information Retrieval (IR) technology used for text classification has already reached a very high level of accuracy, a natural path to a better and better MT quality is to construct statistical models for as many as possible domains of discourse and dynamically use the ones which are the most appropriate for the given input texts.

[3] This concern is related to the fact that more and more bitexts are obtained by machine translation. Reusing translated texts for training SMT engines would reinforce the translation errors and would make room for new ones.

[4] See demo on https://www.youtube.com/watch?v=tBNpglPHQsY

Appropriateness here refers to the similarity between the training data from which the models were learnt and the text to be translated (Tufiş, 2009). One step further is promoted by TAUS (https://www.taus.net/) based on their existing cloud-based computing infrastructure developed for translation memories sharing and intelligent access to high-volume translation resources. This infrastructure is used on an industrial scale by translation companies, departments and individual translators around the world, covering a wide variety of text styles. Essentially, according to this vision, which has been recently articulated into the FT2MT project proposal[5], the future MT "factories" will turn from using large collections of classified training corpora and translation memories towards using translation and language models already optimized, ready to be included or combined in dynamically customized translation systems.

In the MT research community (and not only), the most popular platform is the open-source Moses statistical MT engine. Maintained and contributed by a large community of experts, MOSES framework includes innovative methods and algorithms while allowing for controlled experiments along the FT2MT vision lines. However, the cleaning, clustering and filtering of data and the development of the Moses training modules remain costly, complex and time-consuming. Building an optimized translation model can take days for a real-world MT system and demands several gigabytes of RAM. Although the hardware is becoming faster and more affordable, there are not enough experts with knowledge and skills for the tasks of data selection and cleaning and the selection of the appropriate tools. Therefore, the transition from repositories of domain specific corpora or translation memories to public repositories of domain specific statistical models and the release of platforms for on-the-fly building of customized translation systems will make a tremendous step forward, not only by saving thousands and thousands of CPU processing hours, but will open the pathway to translation technology for a large number of non-SMT experts who could invaluably contribute to experimenting, testing, evaluating and, finally, improving the MT technological knowledge.

Mapping an input text to the most adequate translation model is not a simple task, especially when this has to be done in real time. This comes to incorporating into the processing flow a module able to predict quality score for a machine translated text without access to reference translations. The possibility to automatically build machine translation systems for any translation combination requires further investigations on how the cloud-based environments ensures that user-specific and project-specific model combinations will be created in real time, and the resulted engines will perform instant on-line translation. One way to achieve these objectives, within the reach of Moses' capabilities, is to optimize its engines to work in parallel on different parts of the translated texts.

Among the most advanced research results in the areas of statistical machine translation on which the future translation environments may and should build, one can mention: models combination, domain adaptation, automated MT evaluation and/or estimation, integration of new methods, approaches and architectures, new statistical and linguistic models. In the following we brief on some of these topics.

[5] FT2MT - Fast Track to Machine Translation. This H2020 proposal, from which we re-used and adapted some ideas and references, is currently under evaluation and, thus, not publicly available.

Combination of models. After the decision was made on which bilingual and target monolingual models are most appropriate for the current translation task, the models have to be combined to create one MT engine. Model combination is not trivial at all since each model generates its own hypotheses (based on different features) which have to be conciliated towards an optimal solution. Various approaches have been developed for bilingual models (translation and reordering models), like linear and log-linear interpolation (Foster and Kuhn, 2007), fill-up (Bisazza et al., 2011) and backoff (Niehues and Waibel, 2012), and for monolingual models (language model), like linear and log-linear interpolation, and mixture (Foster and Kuhn, 2007). Because the models combination has to be computed in real time, preferences should be in favour of those approaches which minimize the computational effort. For instance, the language models linear combination should be preferred to the mixture approach because the former does not require to create a new model, but only to access the existing ones.

Additionally, a system combination strategy may also be considered: several smaller MT engines may be created, run in parallel for the text to be translated, and finally combining their outputs, exploiting well-known decision-making techniques (e.g. those based on confusion networks (Matusov et al., 2006)).

Domain adaptation. Domain adaptation is a recent intensive research area and it refers to overcoming the mismatches between training data and input data to be translated at run time. Depending on how and when adaptation data is acquired and adaptation is performed, different types of domain adaptation can be defined. *Supervised* versus *unsupervised* adaptation are differentiated based on whether the adaptation set includes or not target values of the predictor. *Offline adaptation* assumes that adaptation data is received and processed only once before the system starts operating, while *incremental adaptation* and *online adaptation* assume that adaptation data is received and processed in an iterative way, while the system is in use, respectively, in batches or single instances. In a post-editing scenario, this could correspond to adapting MT before it starts to be used (offline adaptation), whenever a sufficient amount of post-edited data has been collected (incremental adaptation), and after each sentence is post-edited (online adaptation).

Domain adaptation applies to major statistical models involved in an MT system: the language model (LM), the translation model (TM) and the reordering model (RM). The investigations on LM adaptation have already matured, mainly due to the research carried on in the field of automatic speech recognition (DeMori and Federico, 1999; Bellegarda, 2004). On the other hand, research on TM and RM adaptation has started more recently (Nepveu et al., 2004), especially within the statistical phrase-based MT approaches (Koehn et al., 2003). RMs are effective for local reordering, having hard times on dealing with long-distance reordering (as in German-English language pair). One of the most effective ways to improve the performance of RMs is to initially re-order the words of the source text to follow the natural order of the target text and apply the translation engine to this modified source text. This kind of input text transformation has been shown to improve the quality of final translation[6].

[6] For a detailed technical presentation and evaluation see: Maria Nădejde – Syntactic Word Reordering for Statistical Machine Translation, MSc Thesis, Universität des Saarlands, October, 2011, http://homepages.inf.ed.ac.uk/s1065915

More often than not, in offline adaptation so-called foreground models are created from the adaptation data and then they are combined with the available background models, built from the original training data. There are various techniques for combining the foreground and background translation models: mixture model (linear or log-linear) (Foster and Kuhn, 2007), fill-up method (Bisazza et al., 2011), back-off method (Niehues and Waibel, 2012), log-linear combination (Koehn and Josh, 2007), merging method (Nakov, 2008), factored method (Niehues and Waibel, 2012), ultraconservative updating (Liu et al., 2012), etc.

While combination of models applies adaptation on dense features (entire phrase-tables), in (Hasler et al., 2012) offline TM adaptation was performed on sparse features, that is by modifying the scores or probabilities of single phrase pairs occurring in the background TM. Adaptation techniques based on sparse feature were recently explored also in the online adaptation setting by Bertoldi et al. (2013), Green et al. (2013), Waeschle et al. (2013) and Denkowski et al. (2014).

Bach et al. (2009) described experiments and evaluations incrementally adapting a speech-to-speech translation system day-to-day from collected data, while in (Cettolo et al., 2013) adaptation based on batches of fresh post-edited data demonstrated significant improvements of an MT system. Irrespective of the adaptation methods, significant improvements have been reported (Yasuda et al., 2008; Matsoukas et al., 2009; Foster et al., 2010; Moore and Lewis, 2010; Axelrod et al., 2011) based on preliminary data selection on the training data that is the extraction of sentences similar to the adaptation data. Although some available training data is left out, this step permits to build more compact and focused models from the background data, which can be then combined with the foreground models.

System tuning. This process may be conducted differently, depending on the availability or not of a reference bi-text (so called development data). The optimisation of the MT engine when development data are available is the easy case and standard techniques for tuning are applied, like MIRA (Watanabe et al., 2007) or MERT (Och, 2003), which aim at maximizing an objective quality measure like BLEU (Papineni et al., 2002), NIST, TER/TERP, or METEOR (Banerjee and Lavie, 2007).

When no bilingual development data is available (the interesting and more realistic case), optimisation of the MT engine is less precise, because the missing references prevents computing the objective quality measure to be maximized. In such a case, the quality measure may be replaced by setting model weights proportionally to the similarity of the model profiles, to the source profile, or by automatic quality estimation procedure (see for instance (Șoricuț and Echihabi, 2010) and the papers on the last three WMT Quality Estimation Shared Tasks[7]).

4. Conclusions

Machine Translation is making impressive progress as more and more pressure is put by the "global internet village". The recently published report "Strategic Research Agenda for Multilingual Europe 2020" (Rehm and Uszkoreit, 2013) by the MetaNet consortium

[7] www.statmt.org/wmt12/quality-estimation-task.htm, www.statmt.org/wmt13/quality-estimation-task.html, www.statmt.org/wmt14/quality-estimation-task.html

places the "translation cloud" as the first priority research theme for the European programs on language technology domain. There is no digital democracy unless the language barriers are removed. The huge commercial potential of language technology, in general, and machine translation in particular, motivates a very active and productive involvement of the big companies, strongly competed on the translation market by new and creative start-ups. With all its current imperfections, the objective of "assimilation translation" today is to a large extent satisfied. Resolution of the more ambitious goal of the "translation for dissemination", equivalent to FAHQMT, is not far away!

References

Axelrod, A., He, X., Gao, J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), Edinburgh, UK, 355-362.

Bach, N., Hsiao, R., Eck, M., Charoenpornsawat, P., Vogel, S., Schultz, T., Lane, I., Waibel, A., Black, A. W. (2009). Incremental Adaptation of Speech-to-Speech Translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies* (NAACL HLT) Conference: Short Papers, Boulder, US-CO, 149-152.

Banerjee, S., Lavie, A. (2007). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, June 2007, 228-231.

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech Communication*, 42:1, January, 93-108.

Bertoldi, N., Cettolo, M., Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. *In Proceedings of the MT Summit XIV*, Nice, France, 35-42.

Bisazza, A., Ruiz, N., Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. *In Proceedings of the International Work- shop on Spoken Language Translation* (IWSLT), San Francisco, US-CA, 136–143.

Boroș, T., Dumitrescu, Ș. D., Ion, R., Ștefănescu, D., Tufiș, D. (2013). Romanian-English Statistical Translation at RACAI. In *Proceedings of the 9th International Conference "Linguistic resources and tools for processing of the Romanian language", 16-17 mai, 2013,* Miclăușeni, Romania, 81-98.

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. In *Computational Linguistics - Special issue on using large corpora*: II, 19:2, June, MIT Press Cambridge, MA, USA, 263-311.

Cettolo, M., Bertoldi, N., Federico, M. (2013). Project Adaptation for MT-Enhanced Computer Assisted Translation. In *Proceedings of the MT Summit XIV*, Nice, France, 27-34.

DeMori, R., Federico, M. (1999). Language Model Adaptation. In *Computational Models of Speech Pattern Processing*, NATO ASI Series, Springer Verlag, 280-301.

Denkowski, M., Dyer, C., Lavie, A. (2014). Learning from Post-Editing: Online Model Adaptation for Statistical Machine Translation. In *Proceedings of European Association for Computational Linguistics* (EACL), 395-404.

Diab, M., Finch, S. (2000). A statistical word-level translation model for comparable corpora. In Joseph-Jean Mariani, Donna Harman (eds.): *RIAO 2000, 6th International Conference*, College de France, France, April 12-14, 2000. Proceedings. CID 2000, 1500-1508.

Dumitrescu, Ș. D., Ion, R., Ștefănescu, D., Boroș, T., Tufiș, D. (2013). Experiments on Language and Translation Models Adaptation for Statistical Machine Translation. In Dan Tufiș, Vasile Rus, Corina Forăscu (eds.) *Towards Multilingual Europe 2020: A Romanian Perspective*, 205-224.

Ehara, T. (2011). Machine translation system for patent documents combining rule-based translation and statistical postediting applied to the PatentMT Task. *Proceedings of NTCIR-9 Workshop Meeting*, December 6-9, 2011, Tokyo, Japan, 623-628.

Foster, G., Goutte, C., Kuhn, R. (2010). Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), Cambridge, US-MA, 451-459.

Foster, G., Kuhn, R. (2007). Mixture-model Adaptation for SMT. *In Proceedings of the Workshop on Statistical Machine Translation* (WMT), Prague, Czech Republic, 128-135.

Fung, P., Yee, L. Y. (1998). An IR approach for translating new words from non-parallel, comparable texts. *In Proceedings of the Conference of the Association for Computational Linguistics* (ACL), 414-420.

Green, S., Wang, S., Cer, D., Manning, C. D. (2013). Fast and Adaptive Online Training of Feature-Rich Translation Models. *In Proceedings of European Association for Computational Linguistics* (EACL), 311-321.

Hasler, E., Haddow, B., Koehn, P. (2012). Sparse lexicalised features and topic adaptation for SMT. *In Proceedings of the International Workshop on Spoken Language Translation* (IWSLT), Hong-Kong (China), 268-275.

Kilgarriff, A. (2001). Web as corpus. In *Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, Shereen Khoja (eds). Proceedings of the Corpus Linguistics 2001 conference*, Lancaster University (UK), 29 March - 2 April 2001, 342-344.

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. *Annual Meeting of the Association for Computational Linguistics* (ACL), demonstration session, Prague, Czech Republic, 177-180.

Koehn, P., Josef Och, F., Marcu, D. (2003). Statistical Phrase-Based Translation. *In Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies* (NAACL HLT) Conference, 48-54.

Koehn, P., Josh, S. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. *In Proceedings of the Workshop on Statistical Machine Translation* (WMT), Prague, Czech Republic, 224-227.

Koehn, P., Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using EM algorithm. *In Proceedings of the Conference of the American Association for Artificial Intelligence* (AAAI), 711-715.

Liu, L., Cao, H., Watanabe, T., Zhao, T., Yu, M., Zhu, C. (2012). Locally Training the Log-Linear Model for SMT. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), Jeju Island, Korea, 402-411.

Matsoukas, S., Rosti, A. V., Zhang, B. (2009). Discriminative Corpus Weight Estimation for Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), Singapore, 708-717.

Matusov, E., Ueffing, N., Ney, H. (2006). Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. *In Proceedings of European Association for Computational Linguistics* (EACL), 33-40.

Moore, R. C., Lewis, W. (2010). Intelligent Selection of Language Model Training Data. In *Proceedings of the Annual Meeting of the Association of Computational* (ACL): Short Papers, 220-224.

Munteanu, D., Marcu, D. (2002). Processing Comparable Corpora With Bilingual Suffix Trees. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), 289-295.

Nakov, P. (2008). Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing. *In Proceedings of the Workshop on Statistical Machine Translation* (WMT), Columbus, US-OH, 147-150.

Nepveu, L., Lapalme, G., Langlais, P., Foster, G. (2004). Adaptive Language and Translation Models for Interactive Machine Translation. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), 190-197.

Niehues, J., Waibel, A. (2010). Domain adaptation in statistical machine translation using factored translation models, Proceedings of the European Association for Machine Translation (EAMT) (http://www.mt-archive.info/EAMT-2010-Niehues.pdf).

Niehues, J., Waibel, A. (2012). Detailed Analysis of different Strategies for Phrase Table Adaptation in SMT. *In Proceedings of the Conference of the Association for Machine Translation in the Americas* (AMTA), San Diego, US-CA.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the Association for the Computational Linguistics* (ACL), 160-167.

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of*

the Association for the Computational Linguistics (ACL), Philadelphia, July 2002, 311-318.

Rapp, R. (1995). Identifying word translation in non-parallel texts. *In Proceedings of the Conference of the Association for Computational Linguistics* (ACL), 320-322.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German Corpora in non-parallel texts. *In Proceedings of the Conference of the Association for Computational Linguistics* (ACL), 519-526.

Rehm, G., Uszkoreit, H. (editors). (2013). *META-NET Strategic Research Agenda for Multilingual Europe 2020.* Springer.

Skadiņa, I., Aker, A., Glaros, N., Su, F., Tufiş, D., Verlic, M., Vasiljevs, A., Babych, B. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. *In Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), Istanbul, Turkey, 21-27 May, 2012.

Șoricuț, R., Echihabi, A. (2010). TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proceedings of the Association for the Computational Linguistics* (ACL), Uppsala, Sweden, June 2010, 612-621.

Tufiș, D. (2009). Going for a Hunt? Don't Forget the Bullets! In N. Calzolari, P. Baroni, N. Bel, G. Budin, K. Choukri, S. Goggi, J. Mariani, M. Monachini, J. Odijk, S. Piperidis, V. Quochi, C. Soria, A. Toral (eds.). *The European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe*, Vienna, 12th and 13th February 2009, 40-42.

Tufiș, D., Boroș, T., Dumitrescu, Ș. (2013b). The RACAI Speech Translation System. *In Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue* (SPED 2013), Cluj-Napoca, 16-19 October, 1-10.

Tufiș, D., Dumitrescu, Ș. D. (2012). Cascaded Phrase-Based Statistical Machine Translation Systems. *In Proceedings of the 16th Conference of the European Association for Machine Translation* (EAMT), Trento, Italy, May 28-30, 2012, 129-136.

Tufiș, D., Ion, R., Dumitrescu, Ș. (2013a). Wikipedia as an SMT Training Corpus. *In Proceedings of the International Conference on Recent Advances on Language Technology* (RANLP), Hissar, Bulgaria, September 7-13, 702-709.

Waeschle, K., Simianer, P., Bertoldi, N., Riezler, S., Federico, M. (2013). Generative and Discriminative Methods for Online Adaptation in SMT. *In Proceedings of MT Summit*, Nice France, 11-18.

Watanabe, T., Suzuki, J., Tsukada, H., Isozaki, H. (2007). Online large-margin training for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), 764-773.

Weaver, W. (1949). Translation. *Carlsbad*, New Mexico, July 15, 1949, 12 pages (www.mt-archive.info/Weaver-1949.pdf).

Yasuda, K., Zhang, R., Yamamoto, H., Sumita, E. (2008). Method of Selecting Training Data to Build a Compact and Efficient Translation Model. In *Proceedings of the International Joint Conference on Natural Language Processing*, Hyderabad, India, 655-660.

USING XML LANGUAGE FOR INFORMATION REPRESENTATION IN DIGITAL FORMAT

GEORGE CRISTIAN BÎNĂ

SyncRO Soft SRL
george@oxygenxml.com

Extended Abstract

Syncro Soft is a software company located in Craiova that develops the *oXygen XML Editor* tool for helping people work with XML and related standards.

XML stands for *eXtensible Markup Language* and has many applications in different domains, including digital humanities and linguistic areas. XML is similar to HTML, in the way that it marks information with specific tags and attributes, for instance identifying what a link is and to what it links to, what a list is, what a paragraph is, etc. – but it goes beyond HTML by allowing you to specify your own names, your own markup. Thus you can define your semantic model and the meaning for those tags and attributes and then use them to associate that semantics to the information you encode in your own XML vocabulary. A number of standards were defined by different organisations, built on top of XML, to specify such vocabularies where specific tags are defined to encode particular meanings.

A very popular standard in the academic community used for encoding documents in digital format is TEI, *Text Encoding Initiative*. TEI is an XML based framework that defines the structure and meaning of tags that can encode in digital form information about existing documents or manuscripts, dictionary information, etc.

oXygen XML Editor allows to package together a set of resources and configuration information to provide support for an XML vocabulary. TEI is one of the XML vocabularies for which such a package already exists, thus providing ready-to-use support for working with TEI documents. This includes defining a specific set of TEI modules that you want to use and getting the corresponding XML schema files, creating such documents and checking them to be valid according to the schema files and also publishing these documents in different formats.

The TEI support in *oXygen* can be further customized for a specific use case or a specific group of users. An example project for this is a customisation for encoding the Romanian dictionary in TEI XML format where a customisation of the TEI editing support is used to provide a user interface tuned for dictionary entry.

This contribution will introduce you to the XML and TEI open standards that can be used to encode documents, it will present some example projects that use these standards and it will provide an overview of how *oXygen XML editor* supports XML and TEI.

CHAPTER 2
LANGUAGE PROCESSING RESOURCES

STATISTICS OVER A CORPUS OF SEMANTIC LINKS: "QUOVADIS"

ANCA-DIANA BIBIRI[1], MIHAELA COLHON[2], PAUL DIAC[3], DAN CRISTEA[3, 4]

[1] *"Alexandru Ioan Cuza" University of Iași, Department of Interdisciplinary Research in Social-Human Sciences, anca.bibiri@gmail.com*
[2] *University of Craiova, Department of Computer Science, mghindeanu@inf.ucv.ro*
[3] *"Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science, paul.diac@info.uaic.ro*
[4] *Institute for Computer Science, Romanian Academy – the Iași branch, dcristea@info.uaic.ro*

Abstract

"QuoVadis" is a corpus of entities and semantic relations between them, built on the *Quo Vadis* book authored by Henryk Sienkievicz. The process of producing the corpus is briefly presented in the paper. Then, levels of annotation, partly automatic, party manual, are summarised and basic statistics over the corpus are reported. The corpus is freely offered to the scientific public interested to study anaphora resolution and entity linking in the Romanian language.

Key words — anaphora, corpus, entity linking, manual annotation, Romanian resources.

1. Introduction

A book is a world of its own, a creation of its author. Nonetheless, it is re-created each time a reader reads it, discovering its action, characters, plot and facts, while also imposing a personal interpretation on them. As language users, readers make connections between characters easily and subconsciously, recognize the referential correspondences of each character that build up linguistic chains, have intuitive predictions about the evolution of the plot and of the acting characters. Thus, above a basic level of primary information conveyed by the text, an inner world of understanding and linking with the general experience of the reader is elaborated.

In general, there should be a common, basic level of understanding of a text, which should reside in the decoding of the primary messages it conveys, by means of deciphering the mentions of characters and the relations linking them. These relations can be of a referential (anaphoric) and a non-referential nature. Referential relations involve identification of multiple mentions that refer the same character or related properties. For example, reading a Romanian version of "Quo Vadis", the book of the Nobel prize laureate Henryk Sienkievicz, a reader would be able to discover that *Vinicius*, *el* (he), *Marcus*, *tânărul patrician* (the young patrician), *ruda lui Petronius* (Petronius' relative), *un tribun militar* (a military tribune), etc. all refer only one character of the book, whose name is Marcus Vinicius.

Of a non-referential nature are other types of relations holding between the characters of the novel. For instance, the affective relationships between Lygia and Vinicius, that are contradictory at the beginning, develop onto a beautiful love-story. Other examples are

the dominance relations between the emperor Nero and his courtiers, or the kinship relation between Lygia and his adoptive parents, Aulus and Pomponia Graecina.

In this paper we describe a corpus encoding mentions of persons, gods, groups of persons and gods, and body parts of persons and gods, as well as semantic relations linking these entity types, classified in 4 categories: referential, affective, kinship and social. The corpus uses as a hub document a Romanian version of the already mentioned *Quo Vadis* text[1]. The selection of this particular book should be attributed to reasons that include: the density of characters and relations, freeness of copyright and, not the least, the fact that, the novel being translated in so many languages, we envisaged the possibility to exploit the semantic annotations in the Romanian version for other languages, by applying exporting techniques.

The motivations for creating this corpus were two-fold. First, among the Romanian textual resources, a gold corpus that could be used for training programs to reproduce human expertise in the recognition of entities and their correlations, including anaphoric and semantic links is still lacking. Second, the activity itself of building the corpus was organised as a complex annotation exercise in the benefit of our master students in Computational Linguistics. The experience gained during the process of annotation and the organisation of this extremely elaborate and time-consuming task helped us to acquire a level of know-how that, we belief, can be exploited in future large-groups textual annotations tasks.

2. Studies in anaphora and semantic links on Romanian

Huang (2000) defines anaphora as "a relation between two linguistics elements, wherein the interpretation of one (called an anaphor) is in some way determined by the interpretation of the other (called an antecedent)". A referential expression needs a source for its saturation. The notion of anaphora should be considered a contextual one, because many referential expressions are, only by themselves, therefore without a context, ambiguous with respect to what can be considered an antecedent. This is why a corpus, and not a collection of anaphor-antecedent pairs, is the proper resource to train a recognition program.

We have found relatively few studies about anaphora concerning the Romanian language. In the following we list a number of studies concentrating on Romanian anaphora from a linguistic formal level: some contrastive approaches between Romanian and French are discussed by Tasmowsky (1990); types of replays anaphora are investigated by Iliescu (1988); approaches to discursive anaphora are inventoried by Manoliu-Manea (1993); Pană Dindelegan (1994) brings into discussion the neutral value of the pronoun *o* (functioning as pro-form, when an antecedent is referred, vs. non-substitute, when it is not bound with the nominal substitute) and the functions of clitics in Romanian; a thorough analysis of syntactic anaphora is elaborated by Dobrovie-Sorin (1994/2000); means of anaphoric realisations and expression and a typology of anaphora is reported by Zafiu (2004); in *Gramatica limbii române* (the

[1] Translation by Remus Luca and Elena Linţă and published at Tenzi Publishing House in 1991.

Grammar of the Romanian Language), the compendium realised by the Romanian Academy, anaphoric and cataphoric phenomena in Romanian are inventoried in the section dedicated to the *Sentence* (2005/2008, 2nd volume); finally, a discursive approach to anaphora and cataphora can be found in (Oroian, 2006). However, none of these studies use a corpus as an organised repository of examples.

All computational approaches to Romanian anaphora, instead, place corpora, even if of small dimensions, at the base of the tools built and the reported evaluations. We mention here some of these studies: the phenomenon of null anaphora – by Mihăilă et.al. (2010), anaphora resolution – by Pavel et al. (2007); importing anaphoric links from English – by Postolache et al. (2006); the relation between referentiality and discourse structure in Veins Theory, with empirical evaluation also on Romanian, among others in (Cristea et al., 1997; Cristea, 2009).

There are extremely many studies presenting corpora containing annotations of entities and semantic links, in other languages than Romanian, many of them for English: the MUC (The Message Understanding Conference) and the ACE (Automatic Content Extraction) corpora (Doddington et al., 2004), or ARRAU (Poesio & Artstein, 2008). For other languages: AnCora corpus – for Catalan and Spanish (Recasens, 2011), DAD – for Italian (Navarretta & Olsen, 2008), COREA – for Dutch (Hendrickx et al., 2008), etc.

Our work, we believe, could be of real support to both theoreticians and experimentalists. As limited as it is now (one genre, one author, and an exact temporal placement), the QuoVadis corpus could provide evidences for their hypothesis, could be a source of positive and counter examples, or could be used to trains programs designed to recognise entities and relationships among them.

3. Building the corpus

The process of developing the corpus was an elaborate and time consuming activity. A research theme to annotate the *Quo Vadis* novel with semantic relations was discussed with the first year master students in Computational Linguistics at the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iași in the autumn-winter university semester of 2012. Since then, all along 4 semesters, annotation conventions were presented and refined in lab sessions, special annotation cases were analysed, students having to work by themselves or in pairs. More versions were produced and improved iteratively, while, at the end, the process being left only in the hands of the most talented annotators. The whole activity is thoroughly described in (Cristea et al., 2014). The experience gained within the group of students was used by the first author to produce an entirely new version of the corpus, described in (Bibiri, 2014). The statistics presented in this paper reflect this final version of the QuoVadis corpus.

Before manual annotation, the book, in cleaned text format, was submitted to a chain of pre-processing steps, by accessing the web services of the NLP-Group@UAIC-FII[2], see also (Simionescu, 2012): the text was first segmented at sentence boundaries, then

[2] http://nlptools.infoiasi.ro/

tokenised, then POS-tagged and lemmatised, and finally a chunker marked noun phrases (NPs) and their heads. Marking noun phrases was necessary, as referential expressions (the textual realisations of entities) have been selected by the annotators among the NP elements already marked. In the identification of entities, heads are important clues, because annotated recursive NPs should have distinct heads and when more recursive NPs have the same head, only the longest is considered to realise an entity. For instance, in the following recursive referential expressions, heads are underlined: [*oameni* *din* [*toate* *stările* *sociale*]] ([*people* *of* [*all* *walks* *of life*]]). In cases when the entity does not have a lexical realisation (zero anaphora), as in contexts where the subject is not expressed in the stretch of the text, part of the morphological and syntactic features of the included entity are recognisable in the person and number of the verb and the annotation reflects this situation. An example of a null entity notation follows: [*te*]$_1$ [*iubesc*; REALISATION=INCLUDED]$_2$, *Marcus*; here the subject of *iubesc* (*love* – v.) is included in the predicative form of the verb, unlike in the English version, where the subject is always expressed: [*I*]$_2$ *love* [*thee*]$_1$, *Marcus.*

As for relations, they always hold between two arguments and, with the exception of referential relations, they are signalled by a word or an expression (the trigger). An important concern in the creation of the corpus was to mark as span of a non-referential relation the minimal stretch of text in which the relation is expressed. This concern is applied with an eye open for the future, thinking at systems presumed to generalise patterns from relations instances, thus drilling a recognition process. It is clear that the shorter the stretches of text that contain relations, the higher the probability to infer correct patterns out of the annotated examples. The length of a relation is the minimal linear span that includes the two arguments and the trigger. Excepting referential relations, where arguments can sometimes be quite distant in text from one another, relations are usually expressed locally, within a sentence, a clause, or even a noun phrase. Directionality of referential relations is always marked from the right argument to the left one, even if not necessarily the closest linearly in the case of coreferentiality.

Example:

Dacă [*Nero*]$_1$ *ar fi poruncit să fii răpită pentru* [*el*]$_1$, *nu te-ar fi adus la Palatin.*

If [*Nero*]$_1$ *had given command to take thee away for* [*himself*]$_1$, *he would not have brought thee to the Palatine.*

where the identical indexes show a coref relation.

For the other types of relations, if arguments are non-intersecting, the normal reading of the trigger gives the direction, as in this example of a social type of relation:

<Eliberând>-[*o*]$_1$, [*Nero*]$_2$…

[*Nero*]$_2$, *when he had <freed>* [*her*]$_1$…,

where the relation holds as: [2] superior-of [1], on the ground that only a superior can free someone, and the trigger (marked between angle parentheses) is *Eliberând* (*freed*) In case of nested arguments the direction is, by convention, from the external argument towards the inner one, as in this example of a kinship type of relation:

[*<sora>* [*lui*]$_2$]$_1$

[[*his*]$_2$ <*sister*>]$_1$

Here the relation is marked as [1] sibling-of [2] and the trigger is the word *sora* (*sister*).

For the manual annotation, the PALinkA tool[3] was used (Orăsan, 2003). Six XML types of elements record all manual annotations: `ENTITY` – making entities, `TRIGGER` – marking relations' triggers, and `REFERENTIAL`, `AFFECTIVE`, `KINSHIP` and `SOCIAL` – marking the eponymous relations. In all, we annotated 9 subtypes of referential relations, 7 subtypes of kinship, 11 subtypes of affective and 6 subtypes of social relations, described in (Cristea et al., 2014).

Following is a more complex example:

... cui i-ar fi putut trece prin minte că [*un patrician*]$_1$, [*nepot şi* [fiu de [*consuli*]$_4$]$_3$]$_2$, *ar putea să se găsească printre gropari*

Besides, into whose head could it enter that [*a patrician*]$_1$, [*the grandson* [*of one consul*]$_5$]$_2$, [*the son* [*of another*]$_7$]$_6$, *could be found among servants, corpse-bearers*

Here (on the Romanian version) there were annotated: [2] coref [1], [2] kinship:grandchild-of [4]; [3] kinship:child-of [4] (in the English version, the notation would have been: [2] coref [1], [7] coref [1], [2] kinship:grandchild-of [5], [6] kinship:child-of [7]).

4. Statistics and discussions over the corpus

The figures in Table 1 present the corpus at a glance.

Table 1: General statistics over the corpus

#sentences	7,281
#tokens, punctuation included	146,822
#tokens summed up under all relations	171,029
#entity mentions	24,636
#referential relations	22,301
#AKS relations (Affective + Kinship + Social)	755
#triggers	752

Figure 1 shows the histogram of lengths of non-referential relations. The good news is that most of the relations have very short length spans. We might even risk to conjecture that the accuracy of a relations recognition program, after being trained on this corpus, will very much fall over this curve.

In the annotation of entities and referential links we were interested to re-compute referential chains. A referential (anaphoric) chain is made up of the complete set of referential expressions coreferring the same entity, ordered in the linear unfolding of the text. To overcome the lack of inter-annotator agreement tests, a set of software filters were designed and run on the XML file, each error triggering new correction phases.

[3] PALinkA (accessible at http://clg.wlv.ac.uk/projects/PALinkA/) was created by Constantin Orăsan in the Research Group in Computational Linguistics, at the School of Law, Social Sciences and Communications, Wolverhampton.

The final goal was to obtain a one-to-one mapping between the set of referential chains and that of entities of the book, being they singular or collective. Thus, a correct chain should include all and only the mentions of one character of the book.

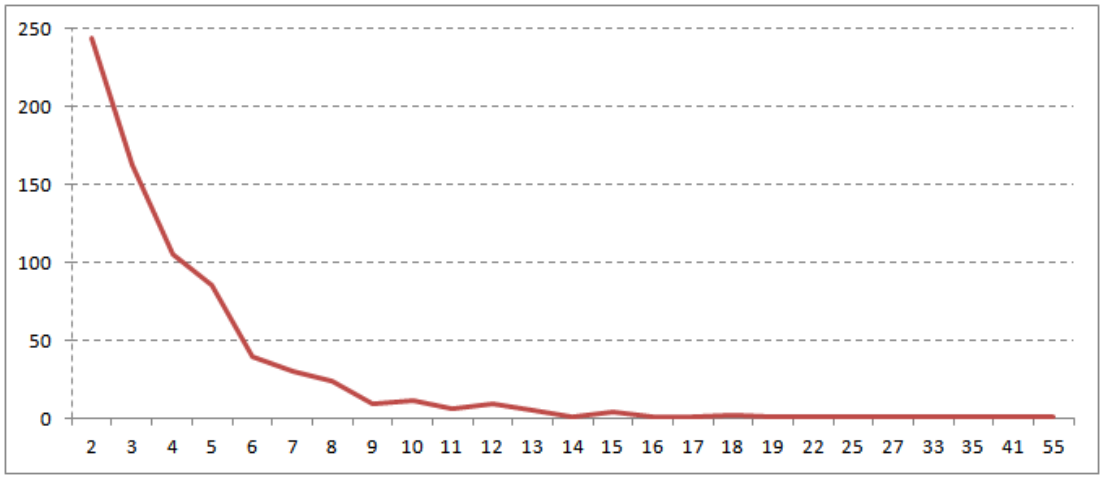


Figure 1: Relations occurrences (y-axis) in correlation with the relations' span length, in number of words (x-axis)

The corpus allows for quantitative and qualitative insights over the characters that are part of the novel. For instance, Figure 2 shows how many times the top 8 characters are referred in the text and in how many relations they occur (this diagram reflects the importance of the characters in the novel, the main ones being Vinicius and Lygia, followed by Nero, Petronius, Chilon, Ursus, Christos and Apostle Peter).

Semantic graphs coding affective and social interactions of characters as well as kinship relations can also be drawn[4]. A semantic graph grouping two affective relations are displayed in Figure 3: the love relation between Vinicius and Lygia and the other characters of the novel, and the worship relationship versus Christos from his adepts. Figure 4 shows two affective relations: this graph displays the feeling of fear that Nero's loyalists and obedients have against him, and the feeling of hate arising between characters as conveyed by the plot.

Finally, we show in Figures 5 and 6 the complex interactions of the main character of the novel: the Roman patrician Marcus Vinicius. Figure 5 reflects the AKS relations between him and other characters of the book. The thickness of arrows suggests the frequency of mentions of relations. As seen in Figure 6, *subordination* and *love* are the dominant relations in the book. Simplifying a lot, they show at a glance what the novel is all about: a love story in a time of predominant social subordination.

[4] Graphs were realised with the NodeXL open-source template for Microsoft® Excel® that automatically generates graphical representations for network edge lists stored in worksheets.

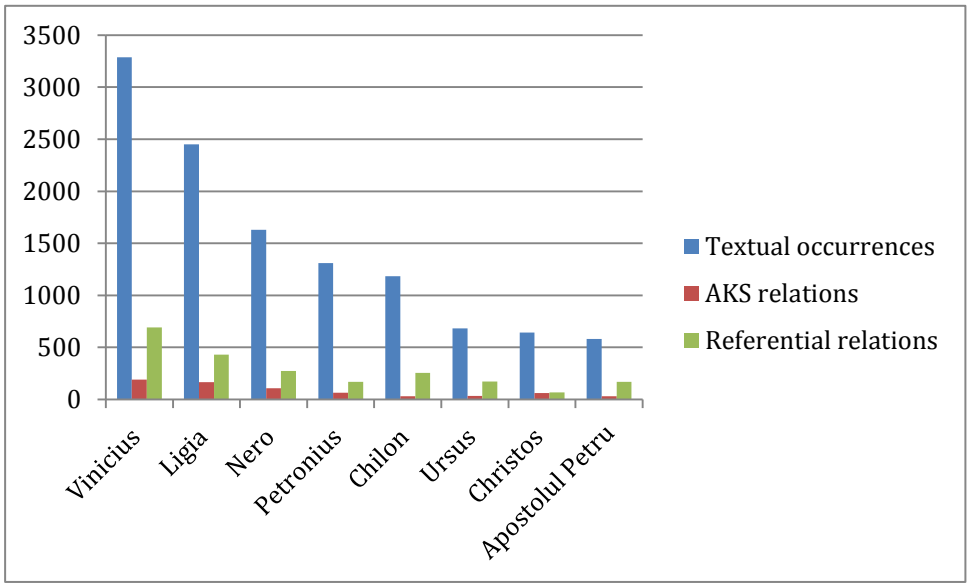


Figure 2: Occurrences of the top 8 characters in the novel

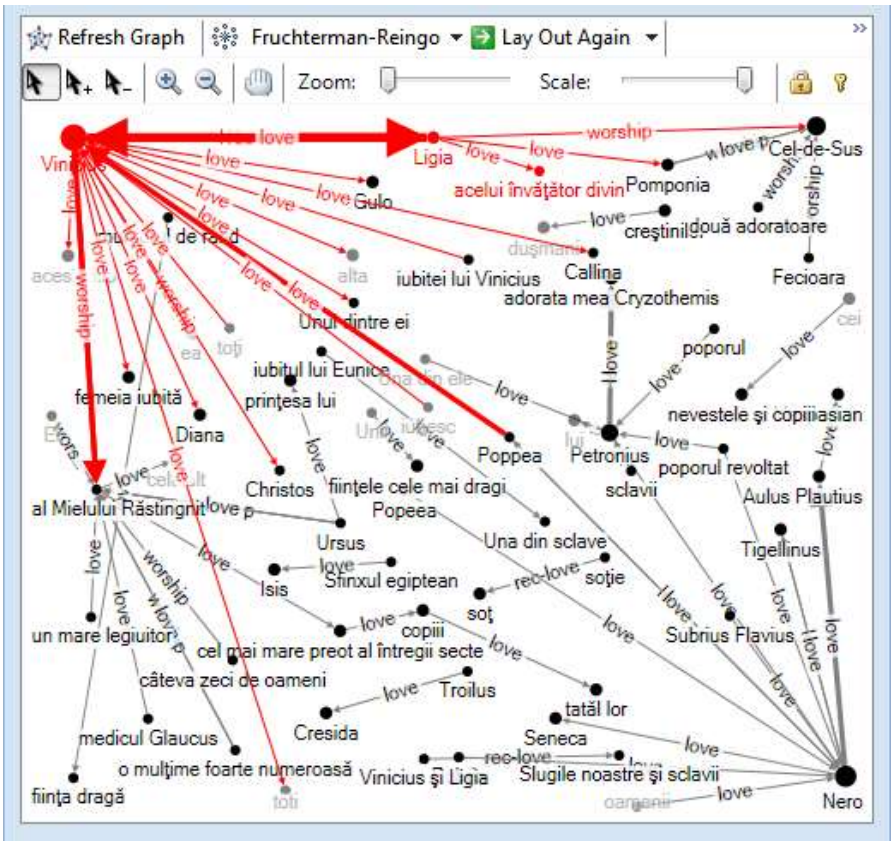


Figure 3: Affective relations `love` and `worship` in the corpus

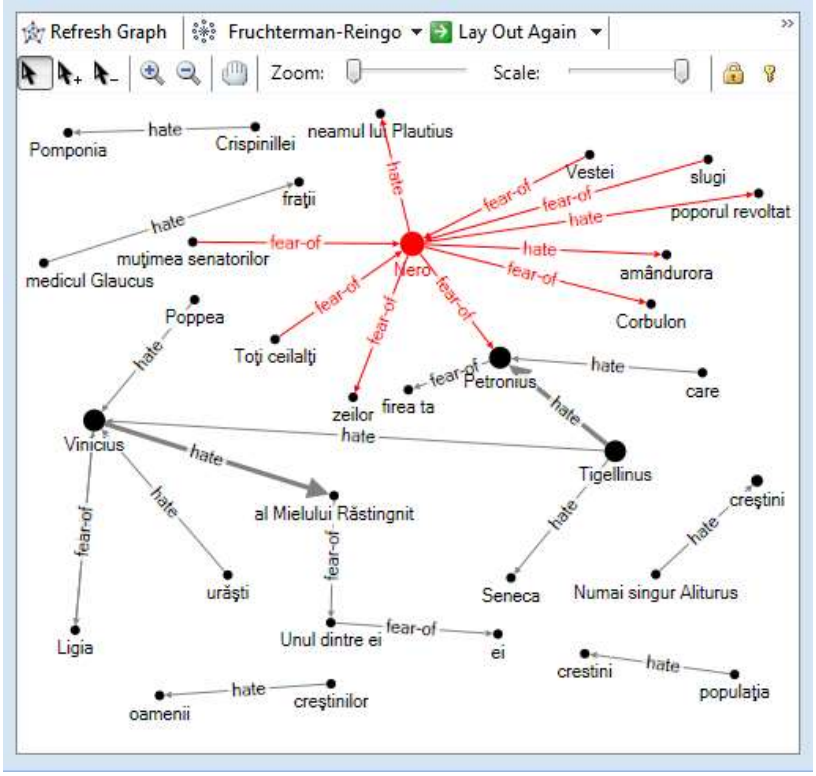


Figure 4: Affective relations fear and hate in the corpus

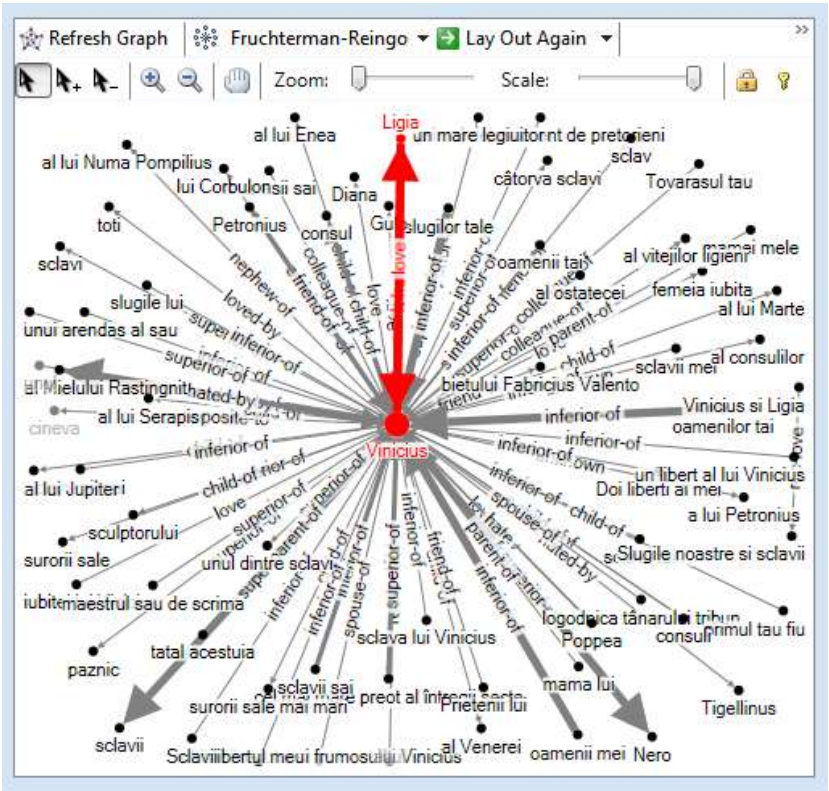


Figure 5: Vinicius' links with other characters

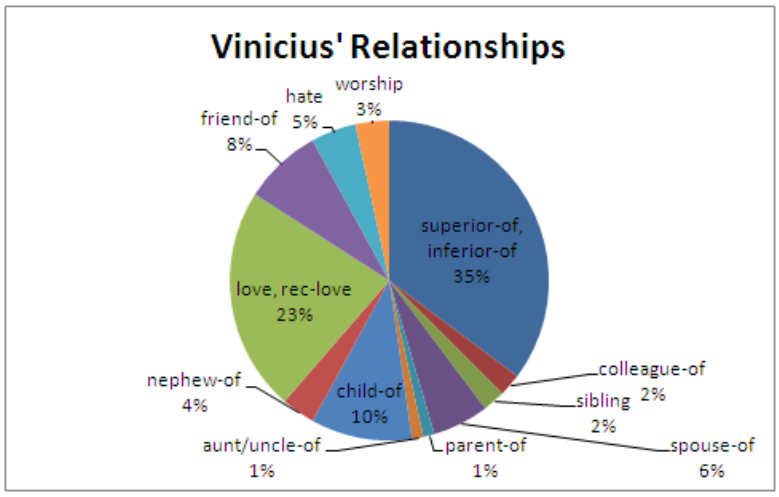


Figure 6: The distribution of semantic relations involving the character Vinicius as one of the arguments

5. Conclusions

We presented in this paper "QuoVadis", a corpus of entities and semantic links. Lack of financial resources did not allow redundant annotation and calculation of inter-annotator agreement. This aspect, of a tremendous importance if we aim at achieving a high enough accuracy to qualify it for a 'gold' corpus, is planned to be solved in the near future, by organising sample-based evaluations. For the time being, to raise the quality of annotation we have made use of software filters, described in (Cristea et al., 2014). For instance, it is highly improbable that a coreferential link contains only pronouns, and it is highly probable that all common nouns in a chain have the same number and gender morphological values. Errors found in the listings generated by these filters were manually corrected during repeated correction phases. Moreover, a special interface was designed to visualise the coreference links[5].

In books, relations are either stable (for instance, if not contested, those describing family links) or may evolve or even completely change, as the story unfolds. To give some examples, the sexually motivated interest that Vinicus shows to Lygia and the lack of interest or even disgust that she has for him evolve both into love; the friendship of Petronius versus Nero depreciates in hate; and Vinicius' lack of understanding versus Christ develops into worship. However, we do not record in this variant of the corpus time frames, so these dynamics are impossible to be caught now.

The following work will concentrate on more directions: first to certify the accuracy of the corpus by organising inter-annotator agreements tests, then to train programs to recognise entities' mentions, to test an anaphora resolution platform for the Romanian language (Cristea et al., 2002a) and to improve it, then to recognise semantic relations belonging to the classes referential, affective, kinship and social, and, finally, even to try experiments of semantic inferences that would exploit combinations of relations (for instance, the difference between the paternal love of Lygia versus her adoptive parents and the one she develops versus Vinicius).

[5] http://nlptools.infoiasi.ro/QuoVadisVisualization/

The corpus is freely available for research at http://nlptools.infoiasi.ro/Resources.jsp and will also be included in CoRoLa, the reference corpus of contemporary Romanian language.

Acknowledgements

Part of the work described in this paper was done in relation with CoRoLa – the computational reference corpus of contemporary Romanian language, a joint project of the Institute for Computer Science in Iași and the Research Institute for Artificial Intelligence in Bucharest, under the auspices of the Romanian Academy. The annotation conventions used in the corpus represent largely work done in preparation of the project "MappingBooks – Let me jump in the book!", financed within the PARTENERSHIP programme of the 2013 Competition (PCCA 2013), in a joint consortium made up of the "Alexandru Ioan Cuza" University of Iaşi, SC SIVECO Romania SA and „Ștefan cel Mare" University of Suceava. We thank our master students in Computational Linguistics from the "Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science, who along the university years 2012-2014 have annotated and then corrected the first version of the "QuoVadis" corpus.

References

Bibiri, A. D. (2014). An Annotated Corpus of Entities and Semantic Relations, M. S. thesis, Master in Computational Linguistics, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași.

Cristea, D. (2009). Motivations and implications of veins theory: a discussion of discourse cohesion. In *International Journal of Speech Technology*, 12:2-3, September, 2009, 83-94.

Cristea, D., Gîfu, D., Colhon, M., Diac, P., Bibiri, A. D., Mărănduc, C., Scutelnicu, L. A. (2014 – to appear) Quo Vadis: A Corpus of Entities and Relations, in Nuria Gala, Reinhard Rapp and Gemma Bel Enguix (eds): *Language Production, Cognition, and the Lexicon*, Springer International Publishing Switzerland.

Cristea, D., Postolache, O., Dima, G.E., Barbu, C. (2002a). AR-Engine – a framework for unrestricted coreference resolution. In *Proceedings of Language Resources and Evaluation Conference*, LREC 2002, Las Palmas, VI, 2000-2007.

Cristea, D., Dima, G.E., Postolache, O., Mitkov R.(2002b). Handling complex anaphora resolution cases. In *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium*, Lisbon.

Dobrovie-Sorin, C. (1994/2000). The Syntax of Romanian. *Comparative Studies in Romance*, Berlin-New York, Mouton de Gruyter; Sintaxa limbii române. Studii de sintaxă comparată a limbilor romanice (Rom. translation), Editura Univers, București.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R. (2004). The Automatic Content Extraction (ACE) program – Tasks, data, and evaluation. In *Proceedings of Language Resources and Evaluation Conference*, LREC 2004, Lisbon, 837–840.

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A. M., Van Der Vloet, J., Verschelde, J. L. (2008). A Coreference Corpus and Resolution System for Dutch. In *Proceedings of Language Resources and Evaluation Conference*, LREC 2008, Marrakech.

Huang, Y. (2000). Anaphora. A cross-linguistic approach, Oxford University Press, Oxford.

Kamp, H., Reyle, U. (1993). *From Discourse to Logic*. Dordrecht: Kluwer Academic Publishers.

Lappin, Y.S., Leass, H.J. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics 20:4*, 535-561.

Manoliu-Manea, M. (1968). Sistematica substitutelor din româna contemporană standard, Editura Academiei R.S.R., București.

Mihăilă, C., Ilisei, I., Inkpen, D. (2010). Zero Pronominal Anaphora Resolution for the Romanian Language. In *Proceedings of Language Resources and Evaluation Conference*, LREC 2010, 17-23 May, Valletta.

Navarretta, C. Olsen, S.A. (2008). Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of Language Resources and Evaluation Conference*, LREC 2008, Marrakech.

Oroian, E. (2006). Anafora şi catafora ca fenomene discursive, Cluj-Napoca, Editura Risoprint.

Orăsan, C. (2003). PALinkA: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, 39-43.

Pană Dindelegan, G. (1994). Pronumele «o» cu valoare neutră și funcția cliticelor în limba română. In *Limbă și Literatură*, XXXIX, 1, București, 9-16.

Pavel, G., Postolache, O., Pistol, I.C., Cristea, D. (2007). Rezolutia anaforei pentru limba română. In Corina Forăscu, Dan Tufiş, Dan Cristea (eds.): *Lucrările atelierului „Resurse lingvistice şi instrumente pentru prelucrarea limbii române, Iaşi, noiembrie 2006"*, Editura Universităţii "Alexandru Ioan Cuza" Iaşi, România.

Poesio, M., Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *Proceedings of Language Resources and Evaluation Conference*, LREC 2008, Marrakech.

Postolache, O., Cristea, D., Orăsan, C. (2006). Transferring Coreference Chains through Word Alignment. In *Proceedings of Language Resources and Evaluation Conference*, LREC-2006, Geneva, May.

Recasens, M., Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. In *Proceedings of Language Resources and Evaluation Conference*, LREC-2010, La Valetta, 44(4):315–345.

Simionescu, R. (2011). Hybrid POS Tagger. In *Proceedings of the Workshop "Language Resources and Tools with Industrial Applications"*, Eurolan 2011 Summer School, Cluj-Napoca.

Tasmowski-De Ryck, L. (1994). Référents et relations anaphoriques. In *Revue Roumaine de Linguistique*, XXXIX, nr. 5-6, București, 456-478.

Zafiu, R. (2004). Observații asupra anaforei în limba română actuală. In Pană Dindelegan, Gabriela (coord.), *Tradiție și inovație în studiul limbii române, Actele celui de-al 3-lea Colocviu al Catedrei de Limba Română*, Editura Universității din București, București, 239-252.

*** (2005/2008) Gramatica limbii române, Editura Academiei, București.

"QUOVADIS" RESEARCH AREAS – TEXT ANALYSIS

MIHAELA COLHON[1], PAUL DIAC[2], CĂTĂLINA MĂRĂNDUC[3], AUGUSTO PEREZ[4]

[1]University of Craiova – România, mghindeanu@inf.ucv.ro
[2] "Al.I.Cuza" University, Computer Science Faculty, Iaşi – Romania, paul.diac@info.uaic.ro
[3] "Iorgu Iordan – Al. Rosetti "Institute of Linguistics, Bucharest – Romania, catalina_maranduc@yahoo.com
[4] "Al.I.Cuza" University, Computer Science Faculty, Iaşi – Romania, augusto.perez@info.uaic.ro

Abstract

We want to achieve here an application of the created resource by the annotation of the novel *Quo Vadis* by Henryk Sienkiewicz. We tried first to obtain the unification of the AKS relations with the triggers detection program. Then we tried to find out how the computer and the statistical programs can help in the correct interpretation of the novel. We noticed, following the *love* relational subtype, that the hypothesis that the novel is an erotic one is not true. This relational subtype covers a variety of feelings, depending on the poles between which it is established, if these poles are relatives, friends or collective characters. The erotic plan is subordinated to the social one, where the Imperial camp opposes the Christian one. The characterisation of the 2 camps can be done by selecting certain relation types or subtypes. The individual characters can also be analysed according to the number of relations, to their type and to their constant or contradictory character, related to a temporal axis which can be traced, in this linear novel, by the id of the sentence.

Key words — affect relations, Christians camp, Imperial camp, inclusion relations, kinship relations, social relations, trigger detection.

1. QuoVadis Project Presentation

This research is based on the results of an ample project (Cristea and Ignat, 2013), developed in the last 2 years at the Computer Science Faculty of "Al. I. Cuza" University, which had as objective the annotation of the referential relations from the morpho-syntactic annotated text of the novel translated into Romanian, as well as, the annotation of the affective, social or kinship relations.

Being a work known worldwide, it has numerous versions in different languages. So, the described structures, especially those of the AKS (Affective-Kinship-Social) type, can be easily exported into an aligned text from another language. This is how we can explain the fact that, although a Romanian version of the novel was annotated, we used English annotation tags. The novel annotation was done and corrected manually by a group of students from the Computational Linguistics MA Programme. First, the people or deities' names entities were annotated. The annotated relations have been of different types:

1. The annotated referential relations have been: *coreferential*, *member-of*, *has-as-member* (inverse), *is a...*, *class-of* (inverse), *part-of*, *has-as-part* (inverse), *subgroup-of*, *has-as-subgroup* (inverse), *has-name*, *name-of* (inverse).
2. Kinship: *parent-of*, *child-of* (inverse), *grandparent-of*, *grandchild-of* (inverse), *sibling* (symmetrical), *ant-uncle-of*, *nephew-of* (inverse), *cousin-of* (symmetrical) *spouse-of* (symmetrical), *unknown*.
3. Affective: *love*, *loved by* (inverse), *rec-love* (symmetrical), *hate*, *hated by* (inverse), *rec-hate* (symmetrical), *fear*, *fear by* (inverse), *rec-fear* (symmetrical), *upset* (symmetrical), *friendship* (symmetrical), *worship*, *worshiped by* (inverse).
4. Social: *inferior-of*, *superior-of* (inverse), *colleague-with* (symmetrical), *opposite to* (symmetrical), *in cooperation with* (symmetrical), *in competition with* (symmetrical).

The establishing of the coreferential relations between entities led to the formation of some coreferential chains that correspond to every character of the novel. We consider the present work as being finished. The result is that the computer can identify, according to the entity's id, the chain which belongs to and the character to which the reference is made and then the relations in which the respective character is involved during the novel.

2. The Research Objectives

The project deals with literature, a field less researched by computer scientists. The novel is a coherent simulation of data from reality, offering a structure similar to the real world, a possible fictional world. If we cannot extract explicit statements of the relations from reality, we can extract them instead from the fictional world of the novel. Remembering key words which synthesize what is specific in a relation, the computer can learn them in order to recognize the same type of relation in any real or fictional context.

What we want is to see to what extent the statistical data obtained with the aid of computers and based on the annotations can facilitate the understanding of the ideas and the structure of this famous novel that received the Nobel Prize for literature. The novel is a complex structure with many plans. We can say that the three areas/directions considered by the annotated relations, AKS, represent, on the whole, the plans of the novel.

3. Related Work

In (Hansen et al., 2011) there is a report about an analysis of the novel with the help of a graph in NodeXL. You can find the *Les_miserables_example.xlsx* graph on the official site of the novel. It contains 254 relations and 77 nodes. The 77 nodes represent the characters of the novel which, in our case, are represented by referential chains. The nodes have different sizes according to the importance of the character in the structure of the novel. In order to calculate it, the authors of the study start from the hypothesis that the more a character appears in scenes, the more important it is. Each scene has the shape of a relation; it represents a side of the graph which connects the nodes of the

participants in the scene. Cosette and Valjean participate together in 31 scenes, the maximum number of co-appearances in this novel.

Here, however, the surface structure of the novel rather than the deeper one is what matters. For other types of novels the criterion is irrelevant. In the structure of the novel *Les Miserables*, Cosette is not one of the main characters.

In our case, the starting point was an annotated text on which different programs had already been applied in order to obtain statistical data. The annotated relations were listed from all the four types, so that we only needed to transform them in the *.xlsx* format. We must notice that the annotation has isolated 370 referential trees that represent not only individual characters but also collective characters (Christians, people, noblemen) or groups of characters which occasionally form themselves in the course of the story line (for example, "Vinicius and his fellows", formed of Vinicius, Chilon, Croton, who tries to kidnap Ligia).

We started from the hypothesis that the more important a character is in the structure of the novel, the bigger is its coreferential chain and the more relations it has with characters from all the story lines. According to the specialists' definition, the main character is the one that can be found in all the story lines, establishing a connection between them and the cohesion of the novel.

However, because we have worked with different annotators and the types and subtypes list of relations has been modified many times during the 2 years of the project, there are differences in annotation. We have already made the unifications and the corrections of the referential relations. We only have to unify the annotation of the AKS relations type.

For this, a trigger detector was built which memorized the triggers' lemmas annotated during the narration and these were looked for in the entire novel, detecting their unannotated appearances. These built up a list of suggested triggers in certain contexts, which were viewed and then validated or not. The validated triggers were added to the annotated triggers list and then one graph for each type of annotated relation was generated.

The idea of the triggers detector and of the program that learns to annotate relations has a wide application both to the analysis of this novel and of other texts, fictional or not.

The program produced a list of trigger suggestions which must be checked by a human annotator. We have obtained a list of 5136 triggers which include also those 757 manually annotated ones, which served for the training of the program.

We detected 1427 affective relations, 1149 of which were manually validated, including the 225 manually annotated. Often, instead of adding new relations, there is an accurate evidence of how many times relations are repeated. Figure 1 is the graph of those relations before and after the triggers were detected by the program and validated:

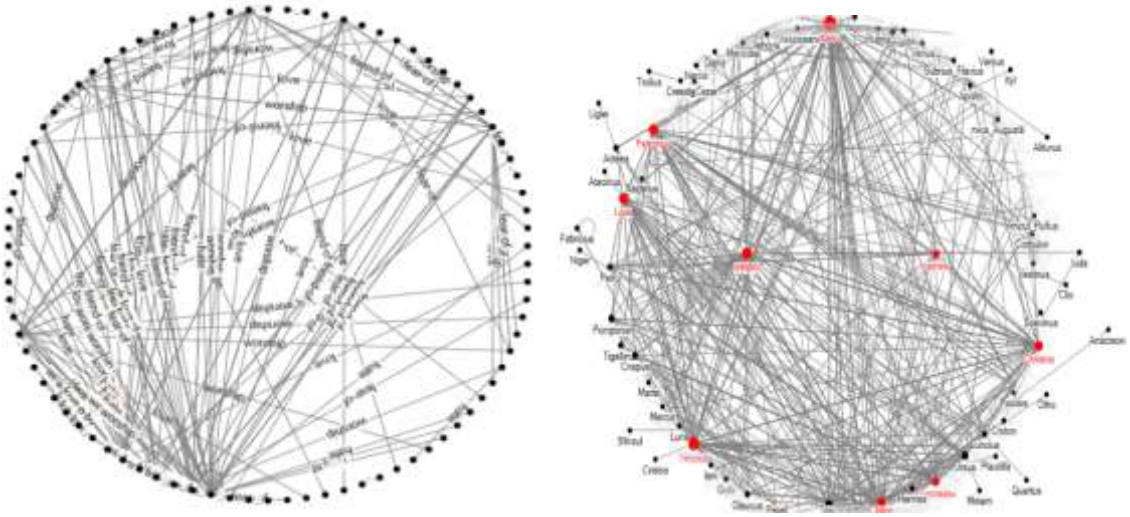


Figure 1: The affect relationships after and before the automatic detection of triggers

The number of kinship relations detected was 624, out of which 317 were validated, including 235 manually annotated. 3085 social relations were detected, 2136 of which were validated, including 297 manually annotated. The validation percentage for affective relations is 80.51%; for kinship is 50.80%; for social is 69.23%. Following the application of the program, the total number of AKS relations increased from 757 to 3602.

To increase the validation percentages, we will fix the manner in which the triggers are placed in the program, taking into consideration the negative or affirmative form of the verb and its prepositional character or its reflexivity. For example, *(s)he doesn't love* triggers *hate*, not *love*; *se roagă* ("is praying") (reflexive verb in Romanian) is the trigger for *worship*, while *roagă* ("ask") is not; *ține la* ("care about") (prepositional verb in Romanian) is the trigger for love, while *ține pe* (be married with), *ține în* ("hold in") are not.

4. The Novel Structure

We started from the hypothesis that the novel is a romance novel. We generated a special graph for the affective relations of the subtype love. But this subtype is not unitary, it contains both erotic feelings and love feelings among friends or relatives. More important, the significance is different when it concerns collective characters, where we can talk about the feelings characteristic to the Christian doctrine, the love for humanity.

The graph of the relation love shows that the hypothesis is not true. The analysed novel is a historical evocation. The historical data combine with fiction and generate a vision which can be applied to any conflict generated by dictatorship. The great couples of forces in conflict are: I. Nero Dictatorship; II. The Christianity Expansion. This (social) story line is the most general one and can be traced at the collective characters level.

For this reason we have selected among the referential annotations those which include the characters in one or more subgroups and groups; these were integrated, in their turn, into bigger groups and finally fell into one of the 2 contender camps. We are going to

detail here the member-of, has-as-member (inverse) and subgroup-of, has-as-subgroup (inverse). These relations establish the way in which the characters integrate themselves in bigger or smaller groups or collectivities.

In the graph named *Inclusions* we colored in blue the characters which fall into the Roman imperial world and in red the characters which fall into the Christians camp. The characters who are undecided or in relation with the two camps were colored in purple. We can observe that their number is quite reduced, the two worlds are isolated, they coexist without interrelating.

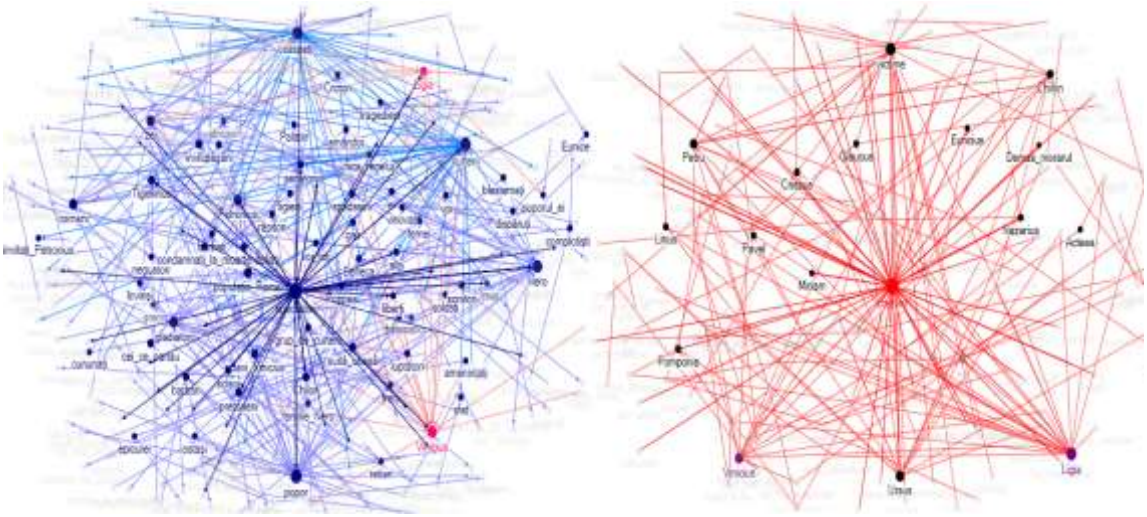


Figure 2: The relationship graphs between the "blue" and respectively, the "red" characters

We are going to intersect these graphs, the blue one and the red one, with graphs resulted from the listing of the 3 types of relations, thus obtaining the type of social, kinship and affective relations which characterizes each of the 2 camps.

4.1. Social Relations

The superiority-inferiority social relations will significantly characterize the imperial plan which is strongly layered between superiors, the Noblemen, and inferiors, the People (thousands of Romans which assist to the city's arson and to different bloody shows). We can say that a gap is created. We will observe that only a few relations are established between characters from the 2 social layers. Both of them are characterized by cruelty, oblivion, hate, fear, desire for vengeance, riot and immorality.

Social relations are the most numerous and that shows the predominant character of the narrative plan. Hierarchical relations are marked in black – mostly, the relations between Nero and other characters from the dictatorship plane are governed by hierarchies.

Relations of opposition – colored in brown – are those established between executioners and victims (the red nodes are for the Christians), that is, between the two ends of the graph. Relations in-competition-with, marked by green lines, are established, especially, between courtiers: Tigellinus, Petronius, Seneca. Christians are characterized by relations of equality, fellowship, marked in pink, and by cooperative relations, marked in blue, so that their side is full color graphics; except for Vinicius, who, although

belonging to the imperial world, migrates, in the second part of the novel, towards Christians; consequently, many relationships in pink and blue start from him.

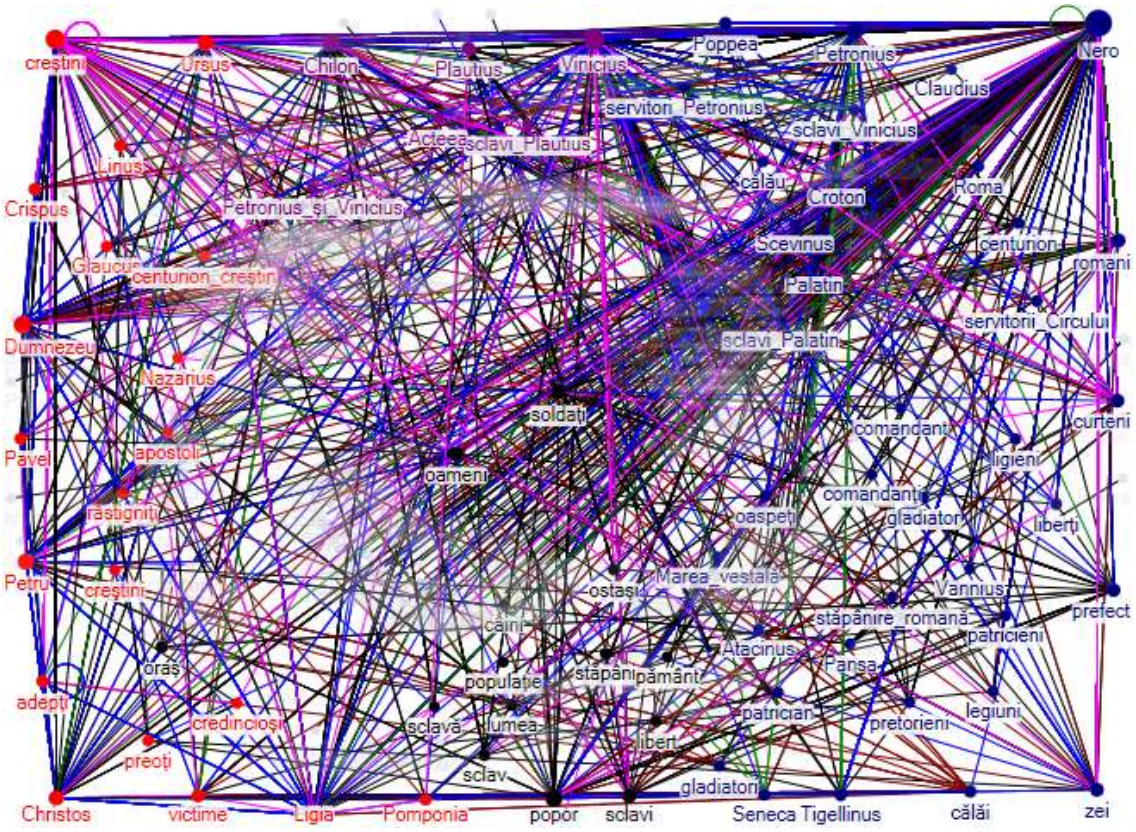


Figure 3: The social relationship graph. The code of colours: *colleague-of*: pink; *in-competition-with*: green; *in-cooperation-wih*: blue; *inferior-of* and *superior-of*: black; *opposite-to*: brown

4.2. Kinship Relations

There are not too many kinship relations in this novel. Nevertheless, their way of structuring is considered a model of Christian camp in which we have few superior relations of the type: *the greatest of the Christians*, *the apostles*, *the bishop* and more numerous equality relations structured as in a family. The brothers and the sisters can be triggers of the relation of kinship type, *sibling* subtype, when words are used in their proper meaning or they can be triggers of the social relation type, *in cooperation-with* subtype, when words are used figuratively. The superiors have for the inferiors a father – son like position; that is why the polysemy of the word *parent*, that can also mean "priest", has been established in language. This is why the developing of the automatic learning machine of the trigger detection meets numerous challenges and still needs a human valuator.

Polysemic words and triggers make the validation percentage of kinship triggers automatically identified be the lowest, that is, only 50.8%. To use this type of relations for building a family tree, it would be necessary to remove all figurative relations. Relations such as "son of Osiris", "son of Mars" focus around Vinicius, who does not have actual kinship relations with anyone other than Petronius.

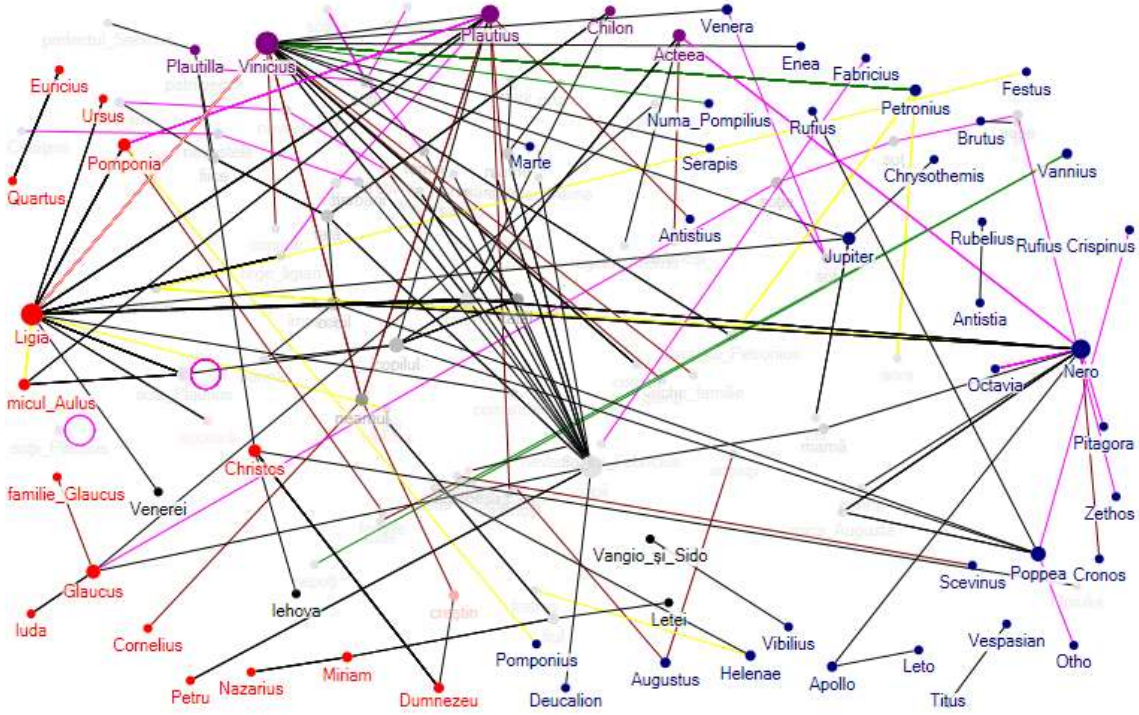


Figure 4: The kinship graph.The code of colours: *nephew-of*: green; *sibling*: yellow; *unknown*: brown; *parent-of* and *child-of*: black; spouse of: pink

4.3. Affective Relations

These relations, very numerous in the novel, actually represent a plan subordinated to the social one. The most numerous affective relations are of the *love* subtype. They are often mistaken for the feelings of the *worship* subtype and for the *friend-of* subtype. In order to make them clear, it is useful to consider what kind of poles the relation has. In the case in which at least one of the poles is a collective character, a group, we can talk about feelings characteristic for the Christian collectivity and not about erotic feelings.

According to our hypothesis, positive feelings should focus on the camp of red nodes of the Christians, and the negative feelings on the other camp. Let us see why this trend – which exists (on the right side, pink, orange and yellow colours are denser) – is not more obvious. As the graph is not directed, the green lines representing fear revolve around Nero, which is the source of all fear. Fear should not characterize the Christian characters, but the graph contains relations throughout the novel, a novel in which Ligia undergoes an evolution; at first, she is afraid of everyone, and, finally, she looks forward to martyrdom.

Most black lines that express hate start from Nero, but also from Vinicius, who, in the first part of the novel, is very aggressive. Ursus, because of his physical force, is a source of fear, and, thus, many green lines surround him. The orange lines representing worship focus, in the imperial world, around the gods, and, in the Christian world, around God and Christ. Chilon, the indecisive character, transfers at some point, to Ursus his hatred for Glaucus. Until his (final) conversion, he is characterized by fear and thirst for revenge. Feelings of love and upset govern all characters, irrespective of camp.

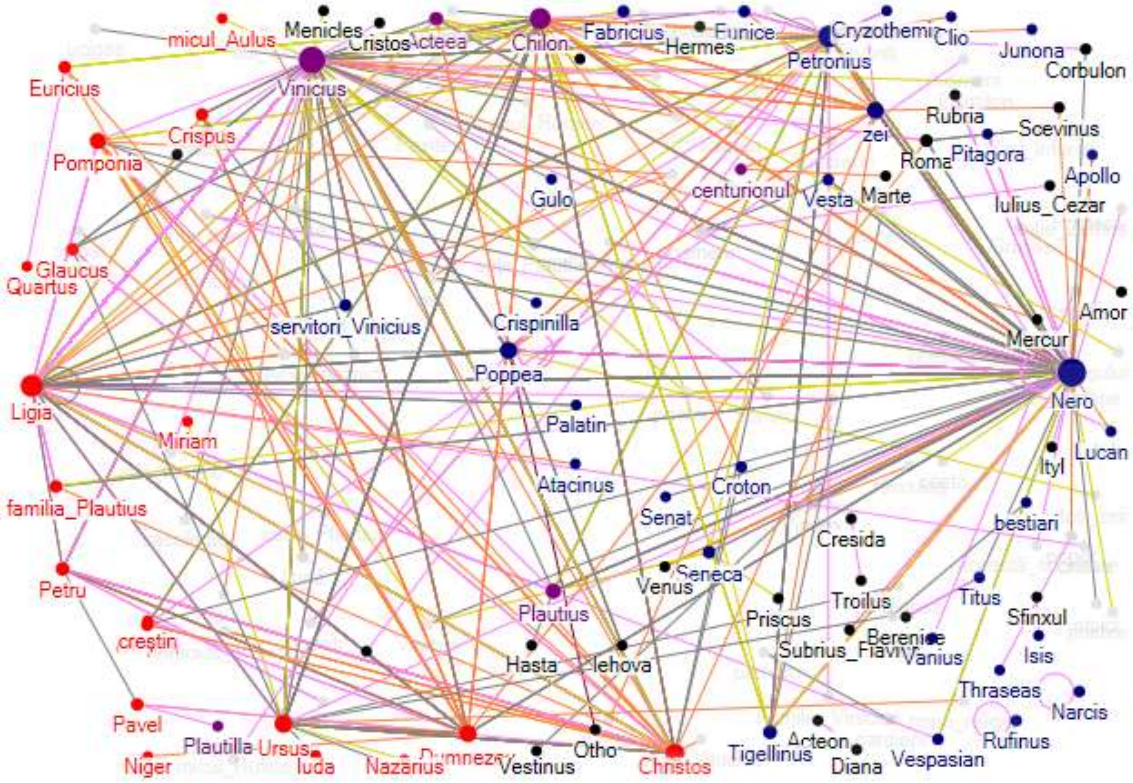


Figure 5: The affect relationship graph. The code of colours: *love*: pink; *worship*: orange; *friendship*: yellow; *hate*: black; *upset*: purple; *fear*: green

5. The Protagonists' Characterisation

This step can be undertaken only by taking into consideration the way the character gets involved in the social plans described above. An important detail is if the character evolves, it changes during the novel or it is static, equal with itself. To check this aspect, we will observe that it is about a classic novel with a linear structure. The method applied here will be the one that considers the ids of the sentences to be a temporal axis. The bigger the numeric difference between them, the greater the distance in time of one character's statement about some contradictory feelings, and this demonstrates the idea of evolution. If the statement of some contradictory feelings is present in sentences with nearest ids, the character is being insincere, which characterizes the imperial world, where there is forced adulation of Caesar. Considering the order of the sentences, a temporal axis cannot be used in modern novels where the story line is disrupted, there being flashbacks, parallel plans, cited documents, etc.

1. The character with the greatest number of relations is Vinicius, strongly involved in both social plans, the imperial one, through Petronius, and the Christian one, through Ligia. The character evolves, passing, in the second part of the novel, from the imperial camp into the Christian one. We can observe that through this his character changes. In the beginning, he selects triggers like *hate*, *anger*, *revenge*, *punishment*, which disappear in the second part of the novel.

We have taken into consideration the moment when Vinicius asserts his friendly feelings for the Christians and demands the Apostles to teach him their doctrine; this moment was recorded in id 3620.

Tabel 1: Affect Relations from Vinicius

Affect	Id 1-3620 = direction - nr.	Id 3621-7145 = direction - nr.
love	**56:** Ligia – 46; Petronius – 2, Ursus – 2, Christos – 2, Chilon – 2, cross – 1, another – 1.	**52:** Ligia – 35, Christos – 7, humans – 2, Roma – 2, pleasures – 1, brothers – 1, St. Peter – 1, religion – 1, Petronius – 1, Ursus – 1.
worship	**35:** Ligia – 15, gods – 7, Christ – 4, Nero – 3, Plautius – 1, Crispus – 1, Mercur – 1, lares – 1, Christans –1, cross –1.	**34:** Ligia – 14, Christ – 10, God – 7, Nero – 3.
hate	**34:** Nero – 2, Ursus – 3, Glaucus – 2, Christ – 2, Gulo – 2, cross – 2, Roman Empire – 1, Plautius – 2, Petronius – 1, Ligia – 2, religion – 2, Acteea – 2, Moon – 1, St. Peter – 2, Christians – 2, Crispus – 2, Chilon – 2, enemies – 1, killer – 1.	**3:** Nero – 2, pleasures – 1.
fear	**17:** Nero – 3, Ligia – 2, Christians – 2, Chilon – 2, religion – 2, St. Peter – 1, servants – 1, sacrilege – 1, Ursus – 1, death – 1, exaggerations – 1.	**11:** Nero – 3, Ligia – 1, St. Peter – 1, God – 1, Christians – 1, pain – 1, smoke – 1, weakening of faith – 1, evil force – 1.
friendship	**10:** Ligia – 3, – Nero 4, companions – 1, Chilon – 1, Petronius – 1.	**3:** Plautius – 1, Petronius – 2.
upset	**3:** Christ – 1, Poppea – 1, Chilon – 1.	–

The constancy of love for Ligia is worth noticing; it has shades of worship, equal on both intervals. Nero's adulation of – by which Vinicius complies to court customs – also remains unchanged. The adulation of the deities, over the first interval, becomes an adulation of the Christian divinity, in two of its hypostases, as God and as Christ. That transformation is not sudden, there is sporadic devotion to Christ during the first interval. The friendliness shown to Petronius is also constant.

A spectacular transformation occurs in respect to the negative relations: hate, prominent in the first interval, when Vinicius is an aggressive, impulsive character, fades away almost completely in the second interval. Fear diminishes a little and there is a shift from the fear of other characters towards a fear of abstract entities like pain, weakening faith, the force of evil.

This kind of analysis could be continued at the level of social relations, where, presumably, in the second interval, there would be a decrease in prevailing hierarchical relations.

2. Petronius is also a main character, the referee of the imperial plan, knowing the other plan through his nephew. He is the keeper of the values of the old world: beauty, poetry, culture. He uses this "power" elegantly. But he is also the keeper of the defects of this world, the most important being insincerity and adulation. His statements about the emperor are contradictory at a small distance in time.

For instance, we consider the 39 affective relations between Nero and Petronius. Of these, 16 are directed from Nero to Petronius: 2 are negative and 14 are positive. However, Petronius commits suicide at Nero's orders. Out of the 23 targeted relations from Petronius to Nero, 3 are negative and 20 positive (love, friendship, worship). Sentences 1580-1581 assert three relations: love, friendship and upset. In sentence 5418, Petronius states his friendship, and, in sentence 5433, his hate of Nero. Therefore, the relation is dishonest on both sides.

3. The character that evolves the most radically is Chilon, the traitor converted in the end and who became a martyr. Given the emotional relations that are directed from Chilon towards the others, the total number of relations detected was 69 – out of which, 25 were relations of worship, 24 of fear, 10 of friendship, 6 of love and 4 of hate. Chilon's fear is directed towards Christians – 11 relations, towards Ursus – 5, Vinicius – 4, deity, victims, Glaucus, Christ. His feeling of worship is directed towards the following: deities – 8 relations, Nero – 6, Christ – 4, God – 3, Croton, Hermes, Ursus, Vinicius, Ligia. His feeling of friendship is directed towards Euricius – 3 relations, Christians – 2, Vinicius – 2, the prefect, Tigellinus, friends. In general, contradictory feelings are stated alternatively, showing dishonesty. Chilo's conversion occurs in the sentence with id 6601, after which feelings of worship are stated which are directed towards Christ, then, the character refuses to abjure and accepts martyrdom.

Such characterisation could be done with regard to other characters in the novel, too.

6. Conclusions

Our research redundantly demonstrated that the annotation of the referential and AKS relations in the novel *Quo Vadis* by Henryk Sienkiewicz is a viable linguistic resource which can be used in a variety of studies. An important direction will be the automatic learning annotation of these relations in other texts, fictional or not, which we have just set up now through the triggers detection program.

Acknowledgements

The authors are grateful to the leader of the project, Prof. Dan Cristea, as well as to the people that participated to painstaking and meticulous annotation of the novel, among whom we mention: Anca Bibiri and Andreea Gagea. Beside Paul Diac, other programmers who participated in the elaboration of the programs for the storage, statistics, displaying and listing of the data were Radu Simionescu and Andrei Scutelnicu.

References

Anechitei, D., Cristea, D., Dimosthenis, I., Ignat, E., Karagiozov, D., Koeva, S., Kopeć, M., Vertan, C. (2013). Summarizing Short Texts through a Discourse-Centered Approach to a Multilingual Context. In *Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems*, Heidelberg: Springer.

Cristea, D., Dima, G. E., Postolache, O., Mitkov R. (2002). Handling Complex Anaphora Resolution Cases. In *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium*, Lisbon.

Cristea, D., Ignat, E. (2013). Linking Book Characters. Toward a Corpus Encoding Relations Between Entities. In *Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue*, Cluj-Napoca.

Cristea, D., Gîfu, D., Colhon, M., Diac, P., Bibiri, A., Mărănduc, C., Scutelnicu, L. A. (2014). Quo Vadis: A Corpus of Entities and Relations. In *Language Production, Cognition, and the Lexicon, Text, Speech and Language Technology*, Springer International Publishing. (N. Gala et al. eds.)

Gala, N., Rey, V., Zock, M. (2010). A Tool for Linking Stems and Conceptual Fragments to Enhance Word Access. In *Proceedings of LREC*, Malta.

Girju, R., Badulescu, A., Moldovan, D. (2006). Automatic Discovery of Part-Whole relations. In *Computational Linguistics*, 32(1), 83-135.

Hansen, D., Shneiderman, B., Smith, M. A. (2011) *Analyzing Social Media Networks with NodeXL. Insights from a Connected World*, Amsterdam-Boston-Heidelberg-London-Oxord–Paris, Elsevier Publishers.

Hjørland, B. (2007). Semantics and Knowledge Organization. In *Annual Review of Information Science and Technology*, 41, 367-405.

Kamp, H., Reyle, U. (1993). *From Discourse to Logic*. Dordrecht, Kluwer Academic Publishers.

Kawahara, D., Kurohashi, S., Hasida, K. (2002). Construction of a Japanese Relevance-Tagged Corpus. In *Proceedings of LREC*.

Masatsugu, H., Kawahara, D., Kurohashi, S. (2012). Building a Diverse Document Leads Corpus Annotated with Semantic Relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, 535-544.

Mitkov, R. (2003). Anaphora Resolution. In *The Oxford Handbook of Computational Linguistics*, Oxford University Press, (R. Mitkov ed.), 266-283.

Postolache, O., Cristea, D., Orăşan, C. (2006). Transferring Coreference Chains through Word Alignment. In *Proceedings of LREC*, Geneva.

Rao, D., Mc. Namee, P., Dredze, M. (2012). Entity Linking: Finding Extracted Entities in Aknowledge Base. In *Multisource Multilingual Information Extraction and Summarization, Springer Lecture Notes in Computer Science*. Berlin, Springer. (T. Poibeau, H. Saggion, J. Piskorski, R. Yangarber eds.).

Rosenfeld, B., Feldman, R. (2007). Using Corpus Statistics on Entities to Improve Semisupervised Relation Extraction from the Web. In *Proceedings of the 45th Annual Meetingof the Association for Computational Linguistics*, Prague, 600-607.

Vlăduţescu, S. (2013). Communicational Basis of Social Networks. *International Journal of Management Sciences and Business Research*, 2(8), 1-5.

Zock, M., Ferret, O., Schwab, D. (2010). Deliberateword Access: An Intuition, a Roadmap and Some Preliminary Empirical Results. *International Journal of Speech Technology*, 13, 201–218.

THE PROVISIONAL STRUCTURE OF THE REFERENCE CORPUS OF THE CONTEMPORARY ROMANIAN LANGUAGE (COROLA)

VERGINICA BARBU MITITELU, ELENA IRIMIA

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest, Romania, {vergi, elena}@racai.ro

Abstract

Within the context of creating the reference corpus of the contemporary Romanian language (CoRoLa) as the main objective of a priority project of the Romanian Academy, carried on by the Research Institute for Artificial Intelligence "Mihai Drăgănescu" and the Institute of Computer Science, we outline the provisional structure of the corpus, alongside the reasons behind our decisions. We also enumerate the contributors that have agreed to become our partners and to help us feed each domain by giving us access to texts. The current state of our project (as for September 1st, 2014) and several available quantitative data are presented in the final part of the paper.

Key words — reference corpus, Romanian, corpus structure.

1. Introduction

In 2012 the Romanian Academy Research Institute for Artificial Intelligence from Bucharest started a project for defining a powerful infrastructure for collecting texts and speech, annotating them, making them available for searching by those interested, and also making public various statistics based on them. In 2014 this initiative was joined by the Institute of Computer Science from Iași, in a larger priority project of the Romanian Academy: the Reference **Co**rpus of Contemporary **Ro**manian **La**nguage (CoRoLa).

A reference corpus is designed to provide comprehensive information about a language (Sinclair, 1996). In order to attain this aim, it has to contain all relevant language varieties and the characteristic vocabulary. To the best of our knowledge, there is no paper that describes the proportion of various functional styles and of scientific domains in the (whole) mass of written texts of any language. In order to establish the characteristics the texts to be included in our corpus should have, we were interested in the structure of similar corpora in the world.

A core of CoRoLa already exists (Mititelu et al., 2014), in the form of a balanced corpus (Ion et al., 2012). In spite of its balanced character, the domains it contains (i.e., news, medical, legal, biographies and fiction) are not a basis for designing the structure of the future CoRoLa.

The paper is organized as follows: we present the outline of CoRoLa in section 2. We compare the structure of various reference corpora around the world in section 3. Afterwards, we outline the provisional corpus structure, alongside the way we foresee to find texts for feeding each domain.

2. The Outline of CoRoLa

CoRoLa will be a big corpus, containing more than 500 million word forms that will be collected until 2017. All functional styles will be represented: imaginative, law and administrative, scientific, journalistic and colloquial style. It will contain both written and oral texts. Each text file will have an associated metadata file and will be sentence-split, tokenized, lemmatized and annotated (at least at the morphological level, but we also envisage a syntactic and even semantic and discourse annotation, see (Bibiri et al., 2014 this volume)). Details about the metadata and the annotation can be found in (Mititelu et al., 2014).

Aiming eventually at a reference corpus, we focus, in a first step, on the contemporary literary language. Contemporary Romanian is the last phase in the evolution of the language, starting, according to specialists, after the Second World War. Due to historic reasons, to the different political, economic and social transformations that marked the community of speakers, we can further divide this period in a post-war (or communist) stage (1945-1989) and the present one (1990-now). The main differences between them are visible at the vocabulary level: words frequencies, words creation mechanisms, borrowings. We want to represent both periods in the corpus, although it is evident that for the latter it is easier, given the existence of texts in electronic format, whereas for the former we need to use printed materials, scan, OCRise and correct them.

The vast part of the corpus will contain texts originally written in Romanian. However, a part of the final corpus will be represented by translations from various domains. Although translated texts may be influenced (at the lexical or the syntactic level) by the originals, this is a phenomenon affecting language and it must be recorded. Also, another part will probably be represented by parallel texts (in which one of the languages will obligatorily be Romanian).

3. The structure of various reference corpora around the world

For identifying the structure of various corpora around the world, we focused on those about which we could find such information and which had at least 100 million words: the British National Corpus (BNC) (http://www.natcorp.ox.ac.uk/corpus/index.xml), the Polish National Corpus (Przepiórkowski et al., 2011), the Czech National Corpus (Čermák & Schmiedtová, 2003), the Hungarian National Corpus (Váradi, 2002 and http://corpus.nytud.hu/mnsz/index_eng.html), the Russian National Corpus (http://www.ruscorpora.ru/en/corpora-stat.html), the Reference Corpus of Contemporary Portuguese (Généreux et al., 2012; http://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc), the Bulgarian National Corpus (Koeva et al., 2012), the Croatian National Corpus (http://www.hnk.ffzg.hr/struktura_en.html).

The German corpus (Kupietz & Keibel, 2009) has an impressive size (over 20 billion tokens); however, they do not target a certain structure of the corpus, but rather gather many texts together, creating thus subcorpora that a user can combine to create his/her own type of needed corpus. As we could not find statistics about the current structure of the corpus, we do not refer to it below.

3.1. Oral vs. written language

Traditionally, the study of language has focused on the written aspect, as it is more relevant linguistically and easier to process technically. Spoken language displays specific phenomena (false starts, anacolutha, etc.) that must be treated consequently for the proper analysis of language. Speech processing is a newer domain and the tools and resources necessary for it are either not so well developed for some languages or absent for others.

The first aspect of interest is the distribution of texts according to the communication medium: oral or written. For the corpora enumerated above the distribution is presented in Table 1.

Table 1: Oral vs. written texts in corpora

Corpus	**Oral component (%)**	**Written component (%)**
British National Corpus	10	90
Russian National Corpus	3.9	96.1
Czech National Corpus	1.3	98.7
Reference Corpus of Contemporary Portuguese	0.8	99.2
Polish National Corpus	10	90
Bulgarian National Corpus	2.6	97.4
Croatian National Corpus	-	100
Hungarian National Corpus	-	100

As one can notice, many of the initiatives of creating big corpora also considered the oral aspect of language. This component is present in various percents, with 10% the highest degree of representation in British National Corpus and Polish National Corpus. When oral components are included in the corpus, they are either transcribed (and the transcription is subject to (roughly) the same processing and annotation as the written text) or they are left unprocessed.

3.2. Informative vs. imaginative texts

Informative texts come from scientific books or from various media (see subsection 3.3) and offer data on different aspects of our lives, mainly society, economy, politics, education, health, entertainment, etc. Imaginative texts are texts from fictional books, mainly in prose, although some projects (for instance the Russian one) also report on a poetry component in their corpus.

As far as the informative component is concerned, the domain distribution varies in statistics presented for the corpora: in some cases the domains are more refined than in others, while for some corpora we could not find relevant data (e.g., for the Portuguese and Bulgarian corpora the statistics we could find mix the medium or the genre with the domains criteria; for Polish, Hungarian and Croatian we could not find statistics about domains distribution). For the data we had at our disposal the situation is presented in Table 2.

Table 2: Domains distribution in corpora

Domain	**British National Corpus**	**Russian National Corpus**	**Czech National Corpus**
Arts & Culture	7.47	6.3	3.48
Social & Political Sciences	35.54	18.7	3.67
Life Style	13.91	10.4	5.55
Business, commerce, economics, finance	8.34	3.3	2.27
Belief & Religion	3.45	4.2	0.74
Law & Security	-	3.6	0.82
Natural Sciences	4.34	0.7	3.37
Applied Sciences	8.15	23.8	4.61
Humanities	1.9	-	-
Administration & Management	-	0.3	0.49
Miscellanea	-	25.5	-

The data are quite bewildering: if social and political sciences are the (far ahead) leading domain in the informative component of British National Corpus, for the Russian National Corpus the leader is the domain of applied sciences, whereas for the Czech National Corpus all the domains in the table represent 25% of the whole corpus, which is dedicated to specialised and technical texts. However, 60% of the corpus is made up of journalism texts, that are not domains distributed, unfortunately.

Table 3 shows the percent of imaginative texts (belles-lettres) in corpora. On average, they represent around a fifth of the whole corpus, although the Russian corpus contains almost 40% of fictional texts and one fifth of the Bulgarian texts are imaginative.

Table 3: Percent of written imaginative texts in corpora

Corpus	**Percent of the total corpus**
British National Corpus	18.75
Russian National Corpus	39.7
Czech National Corpus	15
Reference Corpus of Contemporary Portuguese	0.3
Polish National Corpus	16
Bulgarian National Corpus	25.11
Croatian National Corpus	23
Hungarian National Corpus	20.36

3.3. Written medium

When deciding what texts to include in the corpus, there is another aspect that must be considered: the medium, also with respect to the necessity of covering a broad range of different language styles. This can refer to books, journal articles, magazine articles, newspaper articles, letters, etc. For the corpora examined before, the situation is presented in Table 4. For the Russian, Hungarian and Czech corpora we could not find (sufficient) pertinent data. Čermák and Schmiedtová (2003) only mention 60%

journalistic texts, whereas for the specialised and technical texts, we do not know anything about their medium of origin.

Table 4: Distribution of written medium from which texts are selected

Corpus	Books (%)	Newspapers and magazines (%)	Others (%)
British National Corpus	60	30	10
Reference Corpus of Contemporary Portuguese	0.3	52.2	47.5
Polish National Corpus	21.5	50	28.5
Bulgarian National Corpus	2.5%		97.5 (Internet)
Croatian National Corpus	44	53	3

The category "Others" usually covers different types of texts: brochures, leaflets, manuals, advertisements, letters, memos, reports, minutes, essays, etc. However, for the Croatian National Corpus it is a "mixed" category, and for the Bulgarian National Corpus it covers texts downloaded from internet.

4. CoRoLa's provisional structure and its feeding

In designing CoRoLa's structure we considered mainly the other corpora existing in the world. As stated on the British National Corpus site and as reaffirmed by Čermák and Schmiedtová (2003), when establishing the percents for domains, styles and written medium of the texts we must take into consideration statistics about the readers' preferences. The only study for the Romanian market that we could find was done by IVOX in 2012 (http://ivox.ro/download/get/f/raport-cat-cum-si-ce-citesc-romanii-2012, accessed on August 29th, 2014) and has the shortcoming of not being statistically relevant, as the interviewees were volunteers. Nevertheless, according to it, most people read books (in our terms, imaginative writing) (28.47%) (correlated with the main reason for reading identified in this study, namely for pleasure and relaxation, expressed by 38.85% of the subjects, and with the leading group of 15.59% of people who love reading fiction), a slightly lower percent of people (27.4%) read articles from online magazines, 21.49 percent of people read printed newspapers and magazines, 10.37% of people read online scientific articles and 9.69% of the readers read blogs.

As far as the distinction oral – written form is concerned, we target 10% oral texts for CoRoLa. They will reflect continuous speech and will have the transcribed counterpart, as well. Rador press agency of the Radio Romania will contribute one hour of oral text per day (with the associated transcribed texts) for one year, which totals (when counting only working days) 260 hours of speech.

We will try to obtain the majority of texts from books (60%). We hope to get 30% from newspapers and magazines, while other sources (such as blog posts, brochures, etc.) will contribute 10%.

The written component will be structured according to two different criteria. As far as the functional styles are concerned, the distribution we envisage is presented in Table 5. We target all functional styles: the colloquial is not specifically targeted, although it will

be present in the imaginative texts. Memoirs are not recognized as a functional style, but given their characteristics, we chose to treat them separately. The last column of the table contains the feeders of each style. All those that are enumerated have already signed a written agreement with us to allow us to introduce their texts in the corpus, to process them and make them available for searching for those interested.

Table 5: Styles distribution in the written component of CoRoLa and their feeders

Style	**Percent in the written component**	**Feeders**
Imaginative	25	Humanitas, Polirom, România literară, the journal of Colegiul Național „Unirea" from Focșani, Destine literare
Memoirs	5	Humanitas, Polirom, Editura PIM
Law	10	
Administrative	10	
Science	30	Humanitas, Polirom, Editura Academiei, Editura Universității din București, Editura Economică, Editura Simetria, Muzica, România literară, Editura PIM
Journalistic	20	DCNEWS, România literară, Actualitatea muzicală, România literară, Destine literare

Both great and modest names occur in our list of contributors. It is normal for us to target the important publishing houses, as the readers focus mainly on fiction from books. They can offer "big names" as far as the list of authors is concerned, as well as quality texts, as far as orthography and text format is concerned. However, as the process of persuading publishing houses and media to become our partners in this project is sometimes quite slow (for reasons varying from one potential partner to another), we welcome whomever offers or is easily persuaded by our team or our collaborators to join our efforts.

The law and administrative texts are outside the scope of the copyright law, so we can freely take such texts and add them to the corpus.

As far as the domains are concerned, we propose the distinction in 4 main domains: Arts & Culture, Society, Nature, and Science. Subsequently, they are refined starting from the structure of Romanian Wikipedia, which was reorganized with the type of texts we want and hope that we will be able to collect.

Table 6: Domains distribution in the written component of CoRoLa

Arts & Culture	Literature
	Art History
	Folklore
	Film
	Architecture
	Sculpture
	Painting & Drawing
	Design
	Fashion

THE PROVISIONAL STRUCTURE OF THE REFERENCE CORPUS OF THE CONTEMPORARY ROMANIAN LANGUAGE (COROLA)

		Theatre
		Music
		Dance
		Others
Society		Politics
		Law
		Administration
		Economy
		Army
		Health
		Sport
		Family
		Gossip
		Social Events
		Education
		Social Movements
		Tourism
		Religion
		Entertainment
		Others
Nature		Environment
		Natural Disasters
		Universe
		Natural Resources
		Others
Science	Exact/Formal Sciences	Mathematics
		Informatics
		Logics
		Standards
	Applied Sciences	Medicine
		Archaeology
		Engineering
		Architecture
		Technics/technology
		Aeronautics
		Agronomy
		Metrology
		Criminalistics
		Constructions
		Military Science
		Pharmacology
		Oenology
		Others
	Social Sciences	Pedagogy
		Geography
		Economy
		History

		Psychology
		Sociology
		Ethnology
		Anthropology
		Religious Studies and Theology
		Juridical Sciences
		Linguistics
		Political Sciences
		Philosophy
		Philology
		Others
	Natural Sciences	Biology
		Physics
		Astronomy
		Chemistry

5. Current situation

So far, we have gathered texts from books (both imaginative and scientific ones) provided by publishing houses, from magazines, from an online news website and from two blogs. All texts are electronic: most of them were given to us as .pdf files from which texts had to be extracted (Moruz and Scutelnicu, 2014), several were in .doc format, while those originating on the internet were automatically extracted by us as .txt files. The available .pdf and .doc files have already been converted into 2026 txt files and a metadata file has been associated to each of them manually. A specific application for both creating metadata and extracting text from .pdf files was developed (Moruz and Scutelnicu, 2014), but while this tool was under construction or debugging, we used Arbil (https://tla.mpi.nl/tools/tla-tools/arbil/), which is only a metadata editor, as an alternative. We assured that the output metadata files format was the same for both applications. Moreover, 76868 txt files were automatically extracted from two sites and a metadata file was automatically created for each of them. All these txt files contain more than 60 million word forms (including punctuation).

Texts from newspapers, magazines and blogs were taken entirely. Following international practice, we decided to take no more than 50,000 tokens (word forms and punctuation) from a book. When a book contains less than 50,000 tokens all its text was taken, but we will probably disregard a part of it in the end.

As far as the styles are concerned, we have already gathered imaginative texts, memoirs, scientific and journalistic texts. The domains covered are: Arts & Culture, Society, and Science, whereas the subdomains are numerous: Literature, Art History, Painting & Drawing, Music, Dance, Politics, Economy, Health, Sport, Social Events, Education, Environment, Mathematics, Medicine, Engineering, Architecture, Technics/technology, Constructions, Military Science, Oenology, Pedagogy, Economy, History, Psychology, Religious Studies and Theology, Biology, Physics, Chemistry, etc.

We cannot provide statistics about the coverage of each style, domain, subdomain, etc. at the moment, as we have not started processing the files yet, but this will be our concern in the very near future.

6. Conclusions

In this paper we started with the description of the structure of various big national corpora for which we could find relevant information in order to outline the structure of CoRoLa. We aim at including texts reflecting all functional styles, as well as various domains with their terminology. Given that recent years have witnessed several advances in Romanian speech processing, we intend to include an oral component in the corpus. The majority of the texts will come from books. We presented the current state of our work in the end of the paper.

All the data presented here must be interpreted as provisional. Their reflection in the final corpus is subject to factors (mostly) independent of the teams that work for creating CoRoLa.

References

Barbu Mititelu, V., Irimia, E., Tufiş, D. (2014). CoRoLa – The Reference Corpus of Contemporary Romanian Language. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation - LREC*, Reykjavik, Iceland, 1235-1239.

Bibiri, A.D., Colhon, M., Diac, P., Cristea, D. (2014) Statistics Over A Corpus Of Semantic Links: "QuoVadis". In this volume.

Čermák, F., Schmiedtová, V. (2003). The Czech National Corpus Project: Its Structure and Use. *Lodz Studies in Language*, 7, 207-224.

Généreux, M., Hendrickx, I., Mendes, M. (2012). Introducing the Reference Corpus of Contemporary Portuguese Online. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2237-2244.

Ion, R., Irimia, E., Ştefănescu, D., Tufiş. D. (2012). ROMBAC: The Romanian Balanced Annotated Corpus. In *Proceedings of the 8th LREC*, Istanbul, Turkey, 339-344.

Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, T., Dekova, R., Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0:1, 65-110.

Kupietz, M., Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. *Working Papers in Corpus-based Linguistics and Language Education*, No. 3, Tokyo: Tokyo University of Foreign Studies (TUFS), 53-59.

Moruz, M. A., Scutelnicu, A. (2014). An Automatic System for Improving Boilerplate Removal for Romanian Texts. In this volume.

Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B., Łaziński, M., Pęzik, P. (2011). National Corpus of Polish. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland, 259-263.

Sinclair, J. (1996). Preliminary recommendations on Corpus Typology, Tech. Rep. EAG--TCWG--CTYP/P.

Váradi, T. (2002). The Hungarian national Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Athens, Greece, 385-389.

A CORPUS OF ROMANIAN TEXTS FOR THE INTERCOMPREHENSION IN ROMANCE LANGUAGES

DORINA PĂNCULESCU[1], RODICA VELEA[2]

[1]*University of Craiova, Faculty of Letters, panculescu@yahoo.fr*
[2] *University of Medicine and Pharmacy of Craiova, rodivel13@yahoo.com*

Abstract

In this paper we want to present an analysis of a corpus of texts written in current Romanian, elaborated to be used within Intercomprehension teaching Romance languages activities. Elaborated at the request of the Association for the Promotion of mutual understanding (APIC) headquartered in France, this corpus presents an assembly of non literary texts from press articles dealing with various facts in an accessible language, with a current vocabulary which contains many transparent words, whose meaning is easy to understand either due to the international circulation, to the formal and semantic similarity, either due to some contextual analysis.

Key words — Intercomprehension, mutual understanding, Romanian corpus, transparent words.

1. Introduction

In the current context of globalisation and multilingualism, the promotion of the Romanian language as being equal in rights among the other languages of the European Union appears as an economic, political and didactic target. Encouraged by the European language policy of the Council of Europe supported by important works[1] Europeans, today, can become multilingual, easier than ever, without this statute to impose a complete and thorough knowledge of several foreign languages. This is the concept of 'Intercomprehension'[2] (Teyssier, 2014; Escudé, 2010), which associations as APIC (Association pour la Promotion de l'InterCompréhension) promote and put into practice, applicable especially when speaking about related languages such as Romance languages.

The purpose of this association is to support a methodology for "learning to learn related languages" regarded as equal in importance in a democratic framework of mutual and respectful "listening". The workshops about Intercomprehension in Romance languages are held under the motto: "I speak to you in my language, you answer me in yours, and we understand each other", situation increasingly observed in colloquia and international debates.

[1] *Guide pour l'élaboration des politiques linguistiques éducatives en Europe*. Division des politiques linguistiques, Conseil de l'Europe, Strasbourg, www.coe.int/lang/fr
Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer (CECR). Division des Politiques linguistiques, Conseil de l'Europe, Strasbourg.
[2] APIC : contacts.apic@gmail.com; site : apic-langues.eu

The deliberate choice, as a didactic aim, of a single skill among those necessary for foreign language learning, the understanding (comprehension) of a written text **accompanied by a listening of the oral** reading (performed by a native speaker and registered) is determined by the current actional perspective promoted in the language learning by CECR, which is in favour of the rapid adaptation to frequent and ongoing communication situations, such as collection of information, easy conversation, focusing on vocabulary learning and less on the grammar aspects.

2. The Corpus

As a result of the collaboration with APIC to achieve a corpus of texts in current Romanian in order to use them in the Intercomprehension workshops organized by this association, we compiled a corpus of 10 non literary texts, which must meet the following requirements:

1. *Simple, understandable, attractive texts*

The texts that best match this criterion are the newspaper articles that present various facts with a simple textual structure and a common lexicon containing many 'transparent' words (by the word 'transparent' understanding that word which is intelligible to the speaker of a Romance language either because of the formal and semantic similarity or by neological universalisation or by various strategies of contextual deciphering).

2. *Texts are small*.

These texts, about half a page each, do not require a long and tiring concentration, falling within the margin of time given for their approach, which is of about '20 min. During an APIC workshop, lasting one hour and 30 minutes, 4-5 texts in as many different languages are given for comprehension, in order to achieve the listening, the reading (the mental one) and the understanding of each text in good conditions.

3. *Texts are addressed to a non-specialized public*.

The selected texts are informative ones (presenting news, facts, in a neutral, uncommitted way) and are addressed to a reader with a medium unprofessional culture.

4. *Texts present a vocabulary with a cultural implication*.

These texts include "words with shared cultural meaning" (Galisson, 1991, 1995) (our translation), "a kind of basic behavioural level", meaning those words bearing cultural meanings known by all native speakers.

To avoid a type of limited communication to a mere transmission of information, the chosen texts may contain three types of words with cultural load:

- clichés and stereotypes, most often vehicled by figurative expressions, which transmit views shared by the whole society, having a national character. Examples of clichés are: "Românul s-a născut poet" = the Romanian is born poet or: "Eternitatea s-a născut la sat" = Eternity was born in the village. We also find the names of animals with certain new meanings, metaphorical ones, in a large number of figurative expressions and proverbs belonging to the ethnic Romanian space. In

the same type of figurative structures we can find names of peoples or ethnic coinhabiting groups, as it could be seen in the following examples: "Doar nu dau (sau vin) turcii", it is said in order to moderate the hurry without reason of somebody, "a fi turc(sau ca turcul)" it is said about somebody stubborn who does not want to understand, "a plăti nemțeste" = to pay in equal shares, "a lua (sau a fura) luleaua neamțului" = to get drunk. Such fixed combinations of words are opaque to the reception; they cannot be understood without additional information. In order to make them understood by speakers of a different language, a word by word translation of the structure can be given initially, and then the meaning of the expression can be made clear either by an explanatory sentence or by a similar expression already existing in the native tongue of the receptor.

- the cultural charge results from the association of a place with a specific product, often this combination is generated by the publishing activity or by the media influence. For Romanian, some famous examples are blue of Voroneț, Fat wine of Cotnari, Tear of Ovid (cf. Rădulescu, 2010).
- Cultural charge of some customs, traditions and habits evoked by certain words as the names of the religious or popular holidays, social rituals, etc.

5. *Text-picture association*.

As an important auxiliary, the image facilitates the understanding of the textual type, of the message and of the cultural implicit message. The texts that form our corpus are accompanied by a picture suggested by the source editor, which is the online Observatory, Antena 3, PRO TV.

6. *Texts are exemplary*.

It is a condition of readability, of grammatical and lexical correctness (normative language). The texts come from a trusted source. Their exemplarity character assumes to be homogeneous and representative samples for a particular type of analysis or interpretation according to the definition of the term corpus itself.

In compiling the corpus sent to APIC we met all these requirements. The texts, accompanied by the corresponding images come from the online Observatorul.ro, yahoo.com, romaniaactualitati.ro and they have the following form of presentation:

- the scanned page contains the text;
- the audio version of the text, the reading slightly slower than the normal reading;
- a literal translation of the text in French. In the case of the phraseological expressions,, it provides a functional equivalent expression, a cultural or an idiomatic one (Tenchea, 2008; Rădulescu, 2010).

We shall present two of the non literary texts selected for the corpus and the methods of analysis meant to accomplish the intercomprehension.

Text no. 1

„Ministrul Mihnea Costoiu dezvăluie provocarea-cheie pentru România în viitor

By Andra Dumitru, 06 September 2014

*Ministrul Mihnea Costoiu a declarat, vineri, la deschiderea programului **Academia** BNR, că provocarea-cheie pentru România în orizontul 2020 este dezvoltarea şi dinamizarea inovării prin susţinerea cercetării - dezvoltării.*

*„Susţinerea specializării **inteligente** este unul dintre obiectivele specifice propuse prin Strategia Naţională de CDI. Pentru **ca** România să devină competitivă la nivel regional şi global în orizontul 2020, avem ca obiective crearea unui mediu stimulativ pentru iniţiativa sectorului privat, centrarea unei părţi importante a activităţilor CDI pe problemele societăţii, susţinerea cercetării de excelenţă, atingerea până în 2020 a masei critice de cercetători şi dezvoltarea unor organizaţii de cercetare performante, capabile să devină operatori regionali şi globali", a afirmat ministrul Mihnea Costoiu.*

*„Prin urmare, provocarea-cheie pentru România pentru orizontul 2020 este dezvoltarea şi dinamizarea inovării prin susţinerea cercetării - dezvoltării, stimularea mediului **privat** pentru inovare, susţinerea dezvoltării resurselor umane din CDI orientate spre sectoare cu potenţial de creştere", a punctat ministrul cercetării.*

*Ministrul delegat pentru Învăţământ Superior, Cercetare Ştiinţifică şi Dezvoltare Tehnologică, a participat vineri, 5 septembrie, la Constanţa, la lucrările celei de-a doua sesiuni **din** 2014 a programului „Academica BNR", sesiune dedicată managementului universitar.*

Ministrul Mihnea Costoiu a subliniat importanţa acestui program care certifică implicarea fermă a Băncii Naţionale în sprijinirea învăţământului universitar şi a declarat că astfel de iniţiative au rolul de a stabili şi stabiliza un solid canal de comunicare între BNR şi mediul academic.

La evenimentul de vineri au participat oficiali BNR, în frunte cu guvernatorul BNR, Mugur Isărescu, şi membrii Consiliului de Administraţie al BNR, ministrul Educaţiei Naţionale, Remus Pricopie, reprezentanţi ai Consiliului Naţional al Rectorilor şi ai universităţilor partenere, cadre didactice angrenate în programul „ACADEMICA BNR", precum şi reprezentanţi ai comunităţii financiar-bancare şi afaceri locale şi regionale.

A doua sesiune din acest an a programului „Academica BNR" se desfăşoară în perioada 4-7 septembrie la Universitatea Maritimă din Constanţa şi este organizată de Banca Naţională a României, în parteneriat cu Ministerul Educaţiei Naţionale şi Consiliul Naţional al Rectorilor. »

Source: Romaniaactualitati.ro

Transparent words: *ministru*, noun; *program*, noun; *academie*, noun; *orizont*, noun; *inovare*, noun; *inteligent*, adj.; *specializare*, noun; *strategie*, noun; *competitiv*, adj.; *regional*, adj.; *global*, adj.; *local*, adj.; *stimulativ*, adj.; *inițiativă*, noun; *sector privat*, syntagm; *resurse umane*, syntagm; *mediu privat*, syntagm; *organizare*, noun; *organizație*, noun; *masă critică*, syntagm; *excelență*, noun; *performant*, adj.; *operator*, noun; *management*, noun; *științific*, adj.; *dezvoltare*, noun; *implicare*, noun; *ferm*, adj.; *comunicare*, noun; *oficial*, adj., noun; *eveniment*, noun; *guvernator*, noun; *consiliu de administrație*, syntagm; *consiliu național*, syntagm; *rector*, noun; *universitate parteneră*, syntagm; *comunitate financiar-bancară*, syntagm; *parteneriat*, noun;

educație națională, syntagm; *a certifica*, verb; *a stabiliza*, verb; *a stabili*, verb; *importanță*, noun; *dedicat*, adj.

Cultural words: CDI – research, development, innovation; BNR – National Bank of Romania; MEN – National Education Ministry; National Council of Rectors.

These abbreviations and syntagma refer to European institutions, known by the public. The only difficulty is represented by the words order and the Romanian form.

Morphosyntactic notes. This text belongs to the journalistic style and also contains many terms of legal-administrative style, but it can be easily understood by speakers of a Romance language and even of English**.** It contains a large number of international circulation neological words that are adapted to the phonetic and spelling of the Romanian language.

Statistically, of the 320 words of text, 55 words are transparent ones, some of which have multiple occurrences in the text, resulting in a greater common element for the Intercomprehension (*Minister, objective, development, environment*, etc.)

The Romanian syntax shows a progressive order of the syntactically related terms: determined – determinant (noun – adjective, verb – complement) which is a common element with other Romance languages (French, Italian) and different from the regressive type syntactic order of English.

For transparent phrases like: *national education partner university* etc. it is necessary to specify this rule. In other phrases we can observe linguistic constructions: calques, as *provocarea-cheie,* the Romanian for *the key challenge*.

Neological productive suffixes in Romance languages are found today in Romanian, with the same semantic values: *-iza* for verbs, *-al* for adjectives, *-at* for nouns, etc.

There can establish correspondences between the morphological classes of verbs, for example the type I verbs in Romanian have the suffix *-a*, that corresponds to the suffix *-er* of the French verbs.

For the morphosyntactic value of the terms and for the inflected forms of the words we can refer to other resources of the Romanian language, some of them already computerized, can be accessed by links and others deserve to be computerized and even translated in a language of international circulation as it is the case of *Gramatica practică a limbii române actuale* by Ada Iliescu, or *Dicționar de construcții verbale,* with correspondence in several Romance languages, published in 2002 by a group of teachers from the Faculty of Letters in Craiova, in a CNCSIS grant.

For the vocabulary, we can send through links to on-line DEX and bilingual dictionaries available on the Internet. We believe that the texts of this kind would be a good start in understanding the Romanian language for an adult audience with an average culture.

Text nr. 2

„ Restaurantul în care mănânci fără să plăteşti.” The restaurant where you eat without paying.

Ideea le aparţine unor antreprenori din Londra. Totuşi, dacă vrei să mănânci gratis trebuie să respecţi o condiţie.

Cei care trec pragul restaurantului trebuie să îşi achite consumaţia cu poze. Astfel, oamenii care vin să mănânce la restautantul The Picture House din cartierul londonez Soho sunt puşi să fotografieze farfuriile cu mâncare și să încarce poza pe rețelele sociale cu hastag-ul #BirdsEyeInspiration.

Proprietarii restaurantului s-au gândit să încerce inedita metodă după ce au citit un studiu care arată că peste 50% dintre britanici fotografiază mâncarea înainte de a o mânca.

This text requires a multimodal reading which associates visual information (image analysis) with linguistic information. The information presented is sensation type news with small textual size. Present difficulties are mostly of grammatical nature.

It is observed the presence of some adverbs of manner (*totuși=however, astfel=such*) and some logical discourse connectors (*după ce=after, înainte de=before*). The verbs and the nouns belong to the Romanian fundamental vocabulary, as they are very common words (*a mânca=to eat, a se gândi=to think, a citi=to read, mâncare=food*) or (*restaurant=restaurant, a fotografia=to photograph, farfurie=plate, poză=picture*).

Such words should be explained with the help of bilingual dictionaries and their inflected form by making reference to simple, accessible, logical grammar of the Romanian language.

Transparent words: *restaurant*, noun; *antreprenor*, noun; *idee*, noun; *gratis*, adj.invar.; *a achita*, verb; *condiție*, noun; *consumația*, noun; *cartier londonez*, syntagm; *rețele sociale*, syntagm; *proprietar*, noun; *inedit*, adj.; *metodă*, noun; *studiu*, noun; *britanic*, adj.; *a fotografia*, verb.

The cultural information refers to a new, recently appeared habit of the British people, who photograph the food in their plates when in restaurants, which can be interesting news for the young audience.

3. Conclusions

The objective of elaborating such a corpus is a pragmatic one, seeking recognition, valuing and understanding (comprehension) of the Romanian language in the Romanic context. Romanian poor representation in textual corpora in international databases or even its absence in some Intercomprehension methods as EUROM5[3] make this activity become an urgent necessity.

The interest in the development of one Intercomprehension skill at children of 8-11 years of Romance language countries (France, Spain, Italy, Portugal, Romania) has found expression in the development of some methods of disciplinary learning (biology,

[3] EUROM5, Lire et comprendre cinq langues Romanes (2011). Authors: Elisabetta Bonvino, Sandrine Cadeo, Eulalia Vilaginés Serra, Salvador Pippa. 20 texts into 5 languages (PT, ES, CA, FR, IT), with a methodological presentation and a grammar of lecture.Co-Edition Hoepli (Italy), SGEL (Espagne), La Maison du dictionnaire (France)

for example) that have sites with lessons that can be downloaded from the Internet, such as Euromania.

The adults can also benefit from the learning method of the multilingualism by intercomprehension, following a series of 10 workshops APIC. This form of rapid teaching-learning and with beneficial effects to a wide audience, becoming more mobile, should actually be promoted in other Latin countries.

Close from the typological point of view to the Eastern Romance languages, Romanian can always be understood with a minimal effort by Italian or Catalan speakers, without having a special training of translators or as specialists in philology just by applying the efficient comprehension methods proposed by APIC workshops.

The texts selected to constitute a relatively small size corpus in our project aim at an audience of young adults with a strong sense of knowledge, particularly young people being particularly attracted to breaking news programmes and sensation news.

References

Drăghicescu, J. (2002). *Dicționar de construcții verbale-română, franceză, italiană, engleză*. Craiova: Editura Universitaria.

Galisson, R. (1991). *De la langue à la culture par les mots*. Paris : CLE International (coll. DLE).

Galisson, R. (1995). Lexiculture et enseignement. In *Études de linguistique appliquée*, no. 97, Didier Érudition.

Iliescu, A. (2008). *Gramatica practică a limbii române actuale*. București: Editura Corint.

Rădulescu, A. I. (2010). *Les culturèmes roumains: problèmes spéciaux de traduction*. Craiova: Editura Universitaria.

Teyssier, P. (2004). *Comprendre les langues romanes, méthode d'intercompréhension*. Paris: Chandeigne éds.

Escudé, P., Janin, P. (2010). *Le Point sur l'intercompréhension, clé du plurilinguisme.* Paris: CLE International.

Escudé, P. *Euromania. J'apprends par les langues, manuel européen*. WebSite: www.euro-mania.eu

Țenchea, M. (2008). *Dicționar contextual de termeni traductologici* (French-Romanian). Timișoara: Editura Universității de Vest.

QUANTITATIVE OUTLOOK ON ANGLICISMS IN FOOTBALL-RELATED FRENCH AND ROMANIAN MEDIA

GIGEL PREOTEASA

Universitatea din Craiova, Facultatea de Litere, Craiova – România;
gigelpreoteasa@gmail.com

Abstract

This article is a quantitative approach of English terms in football-related French and Romanian media articles. The analysis we shall carry out has as its starting point the borrowings from the English sports-related language and the extent to which such borrowings permeated the French and the Romanian specialized terminology. We shall equally deal with the reasons underlying such alternative terminology in both languages. The study is based upon commentaries from the daily sports newspapers *L'Equipe* and *Gazeta Sporturilor* during the World Cup matches in Brasil 2014.

Key words — anglicism, football-related, borrowing, lexicometry.

1. Introduction

Football, from a social point of view, is an ever-growing phenomenon, being an element of a widely-shared culture since:

> "*La notion de « culture » a donc bien évolué en un siècle, passant du domaine purement artistique […] et philosophique à celui plus général de la vie collective d'un peuple*" (Vargas et al. (dirs), 2010: 93)

> "*The notion of « culture » has thus evolved in one century, passing from a purely artistic and […] philosophical field to the more general one of the collective life of a people* " (Vargas et al. (eds), 2010: 93).

On the other hand, football is widely disseminated among social environments; in addition, it has quite a distinct expressive and symbolic dimension. Due consideration should be equally given to the easiness with which football terminology penetrates and goes beyond the strict sports-related framework.

Globalisation led to changes in the vocabulary of everyday language in that new terms were created in order to reflect the new emerging realities; it also led to words being borrowed from other languages, such words being either adapted to the target language or borrowed as such from the source language.

Football makes no exception to the changes in terminology: each language created new specialised terms starting from the borrowings from the English language or drew on its own general vocabulary to form a terminology related to the field of football.

This field developed to such an extent that the sports vocabulary is currently employed not only to account for its specific realities, but also to add a more distinctive touch and to increase the expressiveness of the general language, thus going beyond the sports-related activities. Thus, football vocabulary, including borrowings from English,

managed to integrate itself in every person's language culture, becoming terms of current use.

2. Research objective

The widespread use of borrowings in French and Romanian specialised sports language is the starting point of the current study which aims

- on the one hand, at accounting for the extent of anglicisms in football-related sports newspaper articles on a particular event, in a limited period, and,
- on the other hand, at providing explanations for the use of such terms from English.

3. Research questions

Borrowings tend to fall into two categories: first, we have the terms adapting to the target language in terms of their pronunciation, morphology, semantics, undergoing changes to adjust to the linguistic system which adopted them; then, there are the terms which are borrowed as such, preserving their original form in the target language.

Our approach needs to address the following questions: are football-related borrowings from English subject to the same changes at phonetical, morphological and semantic level? Or are they used as such, without any modification whatsoever, in order to preserve their international (globalising) character since quite a large number of football terms in English are used "tel quel" in other foreign languages?

As far as the justification for the use of such borrowings is concerned, the question arises as to whether they are really employed to fill a gap at the lexical or semantic level of the target language or are there any other stylistic or personal reasons on behalf of the commentator which justify the use of these anglicisms?

4. Corpus

The corpus used for the quantitative analysis in the current study is made up of 64 real-time written commentaries of the football matches of the World Cup - Brasil 2014 (12 June-13 July 2014) from the online edition of the sports newspaper *L'Equipe* (www.lequipe.fr), the total corpus of analysis amounting to a 154,549 words.

For the Romanian language, the 64 articles examined were taken from *mondial.gsp.ro*, the number of words analysed being 52,771.

L'Equipe and *Gazeta Sporturilor* are nationwide sports daily newspaper, which gave widespread coverage to the World Cup both in its print and online editions.

5. Elements of lexicometry

In order to manage quite a large amount of data and to avoid any subjectivity during the collection, analysis and interpretation of the data, the quantitative analysis was carried out by means of AntConc 3.2.4w developed by Laurence Anthony in 2011[1].

6. Analysis

Since football was invented in England, its related terminology followed it closely and was adopted all over the world. That is how football terminology managed to permeate the specialized sports discourse in many languages. In the course of time however, the French language created[2], by drawing on its internal linguistic resources, a sports vocabulary of its own to replace the English borrowings and to find equivalents for the new terms generated by the new emerging realities in the world of sport.

This situation has led to a bilingual lexical resource in English and in French, both designating identical or similar realities. There are English football terms which have come to be deeply rooted in the French or Romanian football vocabulary despite there being French and Romanian equivalents for such terms. The football vocabulary needs to be, as shown in the next paragraph, primarily comprehensible and accessible to everyone, that is why the equivalents of English borrowings need to be perfectly adapted to target language and to convey to the same reality in French.

In this respect, it is worth quoting Loïc Depecker (2012) according to whom

> « À l'intérêt un peu marginal que pouvait représenter en 1984 les terminologies du sport, a succédé une prise de conscience de l'intérêt qu'il y a, à tout point de vue – culturel, économique, social, politique – à disposer d'un langage compréhensible et accessible à tous. Les termes du sport circulent en effet partout dans la société, dans les médias, dans le discours de tous les jours».

6.1. Analysis of occurrences

Next I will analyse some football-related terms in English appearing in live (real-life) commentaries in both languages.

A brief description of the tables is provided for ease of reference.

The *English term* in the following tables refers to the anglicism used as such in the French language commentary.

The *French equivalent* points to the equivalent existing in the French language for the English word, equivalent having the same conceptual field and conveying the same meaning as its counterpart in English. Such equivalent can thus replace the anglicism

[1] AntConc 3.2.4w (Windows) developed by Laurence Anthony, Faculty of Science and Engineering, Waseda University, Japan

[2] La Commission générale de terminologie et de néologie (The General Commission for terminology and neology), attached to the Ministry of Sports in France, is responsible for gallicisation of English terms and for defining the sports-related terms, according to Law of 3rd July 1996.

altogether, without any alteration of meaning or possibility of misinterpretation of the original term. The French equivalents are those provided in the general dictionaries and in specialized glossaries[3].

The *English term in Romanian commentary* refers to the English term used as such in the Romanian language commentary, whereas the *Rom(anian) equivalent* concerns the equivalent in Romanian of the anglicism.

The English terms occurring in the French language commentary were looked up in the Romanian commentaries and the results are as follows:

Table 1: Penalty (noun)

	Term	**Occurrences**
English term	Penalty	161
French equivalent	Coup de pied de réparation	0
	Coup de réparation	0
	Tir de réparation	0
English term in Romanian commentary	Penalty	21
Rom. equivalent	Lovitură de pedeapsă	2

Table 2: Score (noun)

	Term	**Occurrences**
English term	Score	94
French equivalent	Marquage	47
English term in Romanian commentary	-	-
Rom. equivalent	Scor (formă adaptată grafic)	8

Table 3: Corner (nom)

	Term	**Occurrences**
English term	Corner	444
French equivalent	Coup de pied de coin	8
	Coup de coin	0
English term in Romanian commentary	Corner	318
Rom. equivalent	Lovitură de colț	115

[3] The French equivalents of the English terms are those provided in Vocabulaire des Sports (2011) and in the UEFA Football Dictionary (2010)

Table 4: Pressing (noun)

	Term	Occurrences
English term	Pressing	28
French equivalent	Pression	55

Table 5: Dribble (noun/verb)

	Term	Occurrences
English term	Dribble	11
French equivalent	Flip-flap	0
	Dribble de l'otarie	0

Note: Out of the 11 occurrences of this term, *dribble* occurs in 6 instances as a verb and in 5 instances as a noun.

Table 6: Star (noun/adjective)

	Term	Occurrences
English term	Star	15
French equivalent	Vedette	4

Table 7: Coach (noun)

	Term	Occurrences
English term	Coach	7
French equivalent	Entraîneur	11

Table 8: Goal-average (noun)

	Term	Occurrences
English term	Goal-average	1
French equivalent	Décompte final	0

Table 9: Staff (noun)

	Term	Occurrences
English term	Staff	15
French equivalent	Officiel (nom)	0
	Personnel (nom)	0

Table 10: Break (faire le break/réussir le break)

	Term	Occurrences
English term	Break	11
French equivalent	-	-

Table 11: Tacle/tacler (noun/verb)

	Term	Occurrences
English term	Tacle (noun)	74
French term	Tacler (verb)	17

Table 12: Gardien/portier

	Term	Occurrences
English term	Goalkeeper	1
French term	Gardien	81
	Portier	91
English term in Romanian commentary	Goalkeeper	40
Rom. equivalent	Portar	0

6.2. Interpretation of results

As the data under scrutiny show, it is quite obvious that the English football-related terms are preferred to their French or Romanian equivalents, where such equivalents are employed. In some cases, however, the French equivalent is used very often, in parallel with the English term: for example the term *marquage* instead of *score*; there are even instances of French terms being used in many more instances than the English equivalent, such as *pression* and *entraîneur.*

Simultaneous use of the English term and of its equivalent is also visible in the Romanian language commentary, where *penalty* is interchangeable with *lovitură de pedepsă* and *corner* with *lovitură de colț,* the tendency to use the English borrowing being quite conspicuous.

Moreover, there are terms in French which are clearly preferred to their English counterpart: that is the case of *gardien* and *portier*, terms used as an alternative to the other term in French, whereas the English term only has one occurrence.

The term *goalkeeper* however enjoys maximum visibility in Romanian since it occurs 40 times and not even an equivalent occurrence in Romanian is provided.

There are English football terms which are clearly preferred to their French equivalents and their occurrences go to prove that: such is the case of *penalty* and *corner*. Such choice is clearly motivated by the concise character of the English term as opposed to the French one, the English terms being deeply rooted in the French sports-related vocabulary.

Differences can be equally found in the grammatical category of the terms. The word *dribble* for example is a verb in English whereas in French it is used as a noun – in English the noun is *dribbling*; the noun in French was used as a basis to create the verb *dribbler*, by adding the verb-forming suffix *-er*. Such is also the case of *tacle,* formed from the English *tackle,* and used to create the verb *tacler,* by means of the same verb-forming suffix.

6.3. Global results

Table 13: Global results

	Occurrences
English terms	861
French equivalents	314
English term in Romanian commentary	379
Rom. equivalents	125

As can be seen from the table above, the English borrowings prevail in the live commentaries of football matches of the World Cup Brasil 2014 in the online edition of the sports newspaper L'Equipe, that is 2.7 times more than their French equivalent terms, whereas in Romanian the English terms are 3.03 times more than their Romanian equivalents.

This analysis may serve as a basis for further studies of the post-match commentaries which, as opposed to real-life (live) online commentaries, may reveal a different set of results considering the different production conditions: absence of time constraints, possibility of recording the match, etc. Such an analysis is also welcome on newspaper articles in print editions of sports papers.

7. Conclusions

Despite the equivalents in the French language, as recommended by *The General Commission for terminology and neology* (*Commission générale de terminologie et de néologie*)[4] which, for purposes of standardisation and development of the French language, has been providing French equivalents to sports anglicisms, English terms are still used in commentaries of football matches.
As far as the terminological gap is concerned, that is the lack of lexical units, lexemes to convey the meaning of the English football term in French, the linguistic policies of the *General Commission for terminology and neology* made available a list of sports-related terms created from the internal linguistic resources of the French language, terms capable of replacing almost any anglicism (Vocabulaire des sports, 2011).

There are quite a few English football terms which have not appropriate equivalents yet; such terms either adapt to the linguistic system of the French and Romanian language,

[4] Specialised commissions for terminology and neology were created in France as a result of the Law passed on 3rd July 1996. Such commissions were attached to every ministry departament and their purpose was to "establish an inventory of the cases in which it is desirable to complete the French vocabulary, taking into account the needs expressed". These specialised commissions then proposed the necessary terms and expressions, mostly equivalents of foreign terms and expressions, accompanied by their definition, to the General Commission for terminology and neology; this body examines the terms, the expressions and the definitions proposed, considers their harmonisation and relevance and requests the approval of the French Academy (http://www.culture.gouv.fr/culture/dglf/terminologie/termino_ enrichissement.htm)

undergoing changes at phonetical, morphological and somatic level, or they are used as such, without any change whatsoever.

As regards the preference for an English term, where there are equivalents in the French or Romanian language, such an option may be justified, in either language, either by the personal choice made by the speaker/enunciator, the commentator in our study, the term being thus subjectively motivated (speaker's expressiveness), or by objective reasons such as concision – which, considering the space and time constraints, is a highly valued quality.

Considering that terms such as *penalty, corner* and *goalkeeper* have a large number of occurrences in both languages (French and Romanian), despite having equivalents that may be used to convey the same concept and consequently capable of replacing the anglicisms, I believe that, beyond the personal choice of the commentator, such a preference is justified by the existence of a *strong core* of anglicisms; such a *core* is made up only of football-related terms, which have taken root in the specialized vocabulary and are used as such, their meaning being widely known and acknowledged, thus leaving no room for an inaccurate interpretation. Such words may form the basic specialized vocabulary of football which is to be found in the languages adopting it, without any difference in their graphical form or any adaptation whatsoever.

The comparative analysis of football-related anglicisms in the two languages, even limited in number for the purposes of our study, may serve as a starting point for a much deeper examination of foreign borrowings in the language of football in Romanian and French and for a further contrastive exploration which may lead to the finding of traits that these languages may or may not share.

Considering the corpus at hand, its limited number of articles, covering only a short span of time, and the precise point in time it occurred, no generalisations may be made, nor tendencies may be identified. The variables in the current case range from the choice of the commentator (who can express its subjectivity or objectivity depending on his likes or dislikes in terms of the football teams), the teams on the pitch (better/poor ranking team, favourite, second runner), the players in the teams (the official line-up) and last, but not least, the particular event subject to comment.

However, this study on a limited number of articles taken from online editions of sports newspapers covering only a limited period of time only serves to point out certain linguistic facts of football language on a given period and on a particular occasion (World Cup); analyses can be equally made on written commentaries with possible slightly or very different results.

References

Depecker, L. (2012). Introduction. *Le langage des sports: identité et typologie*, Le savoir des mots, n° 9, 12, Paris.

Langenscheidt in kooperation mit der UEFA (2010). *Praxiswörterbuch Fußball English-Deutsch-Französisch*, Nyon, Langenscheidt.

Nin, F. (2010). Des gradins du stade aux bancs de la classe: Le football comme médiateur de connaissances d'une langue/culture en LVE. In Claude, Vargas /

Louis-Jean, Calvet/Médéric, Gasquet-Cyrus; Daniel, Véronique/Robert, Vion (dirs) (2010): *Langues et sociétés – Approches sociolinguistiques et didactiques,* Paris, L'Harmattan.

Vocabulaire des sports (2011) Enrichissement de la langue française, Commission générale de terminologie et de néologie (http://www.dglf.culture.gouv.fr/publications/vocabulaires/sports_2011.pdf)

www.lequipe.fr

www.mondial.gsp.ro

http://franceolympique.com/cat/232-terminologie.html

AN ANALYSIS OF WH-WORDS AND INTONATION OF ROMANIAN INTERROGATIVE SENTENCES

VASILE APOPEI, OTILIA PĂDURARU

Institute of Computer Science of the Romanian Academy, Iași branch

{vasile.apopei, otilia.paduraru}@iit.academiaromana-is.ro

Abstract

The present study analyses several types of interrogative sentences that include interrogative and relative pronouns, pronominal adjectives and adverbs (all referred to as wh-words) from the point of view of prosodic phrasing and intonation. An important preliminary step was to develop rules that establish which wh-words are relative and which are interrogative. The intonation analysis was made from the perspective of hierarchy of the prosodic domains. Also we have sought explanations for differences between intonational contours of the utterances of same sentence.

Keywords **—** wh-word, phonological phrase, prosodic domains.

1. Introduction

Technical literature distinguishes between interrogative sentences and questions (Guțu Romalo, 2008). The interrogative sentences are associated with various syntactic structures and semantic mechanisms. Typically, an interrogation behaves like a question type speech act. A question corresponds to a pragmatic reality of speech, initiated to obtain an answer from an interlocutor (Guțu Romalo, 2008). The speakers ask questions (intend to get information), either for covering some lack in cognition or for hearing the interlocutor's answer. The question-answer pair is the backbone of the verbal interaction (*Cine a câștigat meciul de Cupa Davis? − Elveția.* 'Who won this Davis Cup game? − Switzerland.'). In some cases, speakers use interrogations to perform other speech acts: assertive, directive, commissive, expressive; to make an assertion with rhetoric questions (*Cine a mai pomenit aşa ceva? / − Nimeni n-a mai pomenit aşa ceva*. 'Whoever heard of that? Nobody heard that.'). On the other hand, speakers can perform question type speech acts by using structures with a non-interrogative syntax (Guțu Romalo, 2008), as in the case of assertive, imperative or exclamatory sentences (*– Vreau să știu cine ți-a spus asta. / Cine ți-a spus asta*? 'I want to know who told you that. Who told you that?').

The negation *nu* 'no', present in an interrogative sentence, can give rise to rhetoric questions with inverted polarity and signals an affirmation in the corresponding statement (*Cine nu vrea să fie bogat?* Oricine vrea să fie bogat. 'Who doesn't want to be rich? Anyone wants to be reach.'). Sometimes, denial is treated as a discursive strategy which attenuates the aggressiveness of the question (*Nu l-ați văzut pe Ionuț? Nu știți cât e ceasul?* 'Did you see Ionuț? Do you know what time it is?').

The intonational contour of utterances corresponding to interrogative sentences depends on the semantic context and on the emotional state of the speaker. Sometimes, the

emotional questions are accompanied by interjections originated from interrogative pronouns/adverbs which additionally express emotional contextual values (***Ce,*** *n-a venit nimeni?* 'How is that, nobody came?'*;* ***Cum,*** *ai picat la examen?* 'How is that, you failed the exam?') (Guțu Romalo, 2008).

Our intonational contour analysis is based on (Dascălu and Jinga, 2008; Apopei et al., 2006; Jitcă et al., 2014) who have revealed that the boundary tones and the tones corresponding to wh-words depend on the type of question (yes-no question, wh-question). To model the prosodic events, we used a prosodic model developed for Romanian (Jitcă et al., 2009, 2012) which is based on the hierarchy of the prosodic domains (Selkirk, 2005) and the ToBI framework (Backman et al., 2005).

In what follows, we shall analyse several morpho-syntactic contexts of the interrogative sentences that include interrogative and relative pronouns, pronominal adjectives and adverbs, from the point of view of prosodic phrasing.

2. Interrogative sentences

In what follows, we shall use the term 'wh-expression' for a wh-word preceded by 1-2 prepositions (*de la cine* 'from whom', *pentru câte* 'for how many', etc.). In this case, the whole group will be treated as a compound pronoun/pronominal adjective/adverb.

In this section, we analyse a set of interrogative sentences that contain wh-words/ wh-expressions. To reach this goal, we used a POS (part of speech) tagger to establish when a wh-word is interrogative and when it is a relative pronoun, pronominal adjective or adverb. The utterances corresponding to the chosen texts were annotated at the prosodic level using the Ro-ToBI system (Jitcă et al., 2014) and the computer program Praat (Boersma and Weenink, 2014). Within this analysis, we took into account the intonation of the yes-no questions, wh-questions, questions with multiple interrogations and alternative questions. The questions were uttered by two female speakers, in suggested contexts and chosen by speaker. Selected questions aimed to cover a number of situations from (Guțu Romalo, 2008) and for intonation analysis they were grouped into the two broad categories: yes/no questions (the wh-word is not an interrogative pronoun) and wh-questions (the wh-word is an interrogative pronoun).

2.1. Question with the wh-word placed in final position

When the pronoun/pronominal adjective/adverb to be analysed is expressed by a single word, two types of intonation are possible: information-seeking intonation without surprise and information-seeking intonation expressing surprise/astonishment/ indignation (e.g.: elliptic questions ***Cine***? 'Who?', Va sosi ***cine***? 'Who will come?'). The latter interrogation is an example of affective question, when intonation expressing astonishment is more common.

When an interrogative sentence ends in a wh-expression, intonation expresses astonishment most times. (*Va sosi* ***cu cine****?* 'He will come with whom?', *Supărat* ***pe cine****?* 'Angry with whom?', *Le-a luat* ***de la cine****?* 'He took them from whom?'.)

The pronoun/pronominal adjective/adverb to be analysed is interrogative, but the final part of the intonational contour is either ascending-descending (for a pure information-

seeking interrogation without surprise) or ascending (for an interrogation expressing surprise/astonishment/indignation). Figure 1 illustrates the F0 contours of two utterances of the same question: *Va sosi* ***cine***? 'Who will come?'. The first speaker was just seeking for some information, while the second was very surprised after hearing the name of a person who was about to come.

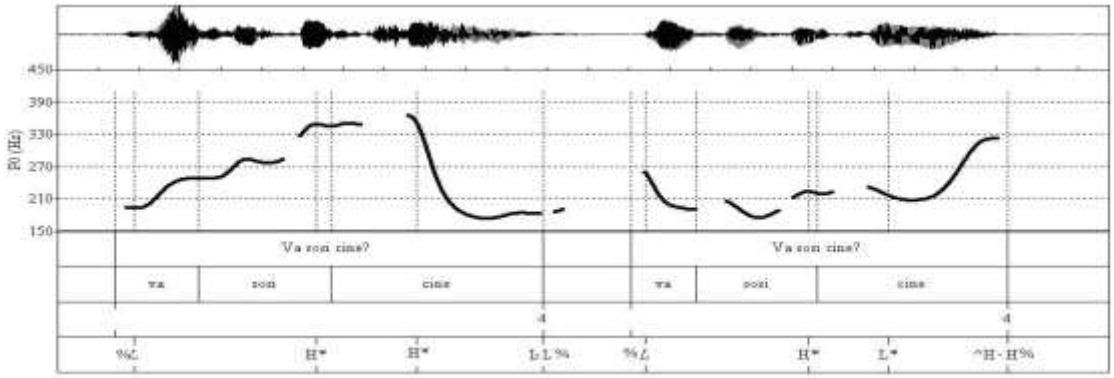


Figure 1: Waveforms, spectrograms, F0 contours and prosodic phrasing of the sentence *Va sosi cine*? 'Who will come?', uttered by two speakers

Typically, such an interrogation is short and expressed by a compact intonational phrase (IP). If minor phonological phrases (MiPs) are present, they do not include the wh-word.

2.2. Interrogative sentences with the wh-word placed in non-final position

2.2.1. Yes-no questions

In yes-no questions, the wh-word/wh-expression is expressed by a relative pronoun/ pronominal adjective/adverb, uttered with a low tone and introduces a relative clause. In an interrogative sentence, a relative wh-word/wh-expression can be identified if a particular morpho-syntactic context (pattern) from a predefined set is detected. This set slightly differs according to the wh-word to be analysed. For the interrogative pronoun *cine* 'who', the set of patterns is the following:

(P1) the wh-word/wh-expression is framed by a pair of verbs, in the indicative or conditional mood (underlined in the next examples), with no comma or conjunction (other than *să* which is part of a verb in the Romanian subjunctive mood) between them (*Știi cumva* ***cu cine*** *merge*? 'Do you know, by chance, with whom he is going?', *Știi* ***unde*** *merge*? 'Do you know where he is going?', *Crezi că vei veni* ***cu cine*** *vrei*? 'Do you think you can come with whom you want?' *Poți să vii acasă* ***cu cine*** *te-am rugat*? Can you come home with whom I asked you to?).

(P2) there is a pair of verbs in the indicative or conditional mood (underlined in the next examples) on the right-hand side of the wh-word, with no comma, other wh-word/wh-expressions or conjunction (including *să*) between them (***Cine*** *va sosi primul va pleca ultimul*? 'Who will arrive first will go last?', ***Cine*** *cântă bine și dansează frumos va lua un premiu*? 'Who sings well and dances nicely will get a prize?').

(P3) the wh-word/wh-expression is followed by a comma (*Nu știu **cine**, dar eşti sigur că va veni cineva? 'I do not know who, but are you sure that someone will come?');*

(P4) there is a 'verb in the indicative/conditional mood – comma – verb in the subjunctive mood' sequence (underlined in the next example) on the right-hand side of the wh-word ***Cine*** *a făcut avere prin fraudă, să pretindă acum respectul oamenilor?* 'Can someone who made a fortune by fraud claim people's respect?'). In this example, typical intonation displays astonishment or indignation.

Most types of yes-no questions include a P1 or P2 pattern. When no comma is present, such a sentence corresponds to one IP composed of two MiPs. An analysis of a sentence including a P2 pattern (***Cine** va sosi primul va pleca ultimul*? 'Who will arrive first will go last?') and a P1 pattern (*Știi cine va sosi primul*? 'Do you know who will arrive first?') is presented in Figure 2. The minor phonological phrases are: (m1) ***Cine** va sosi primul* 'Who will arrive first' and (m2) va pleca *ultimul* 'will go last' for the first sentence and (m1) *Știi* 'Do you know' and (m2) *cine va sosi primul* 'who will arrive first' for the second.

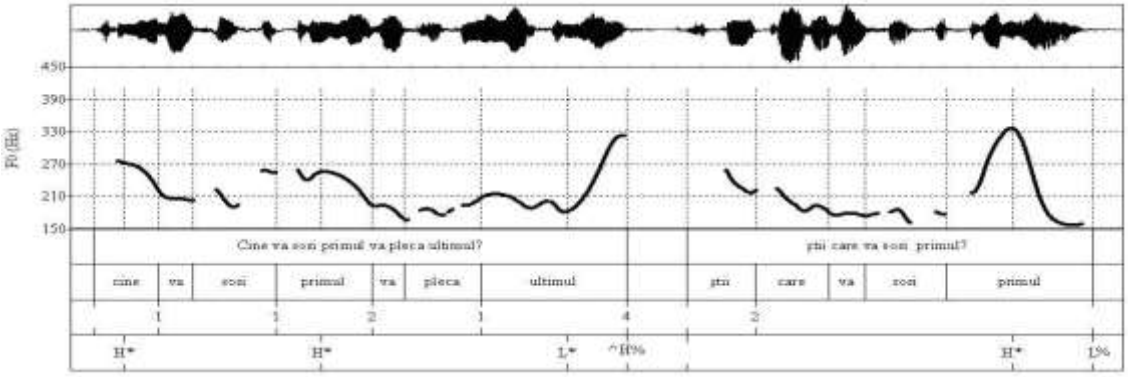


Figure 2: Waveforms, spectrograms, F0 contours and prosodic phrasing of the sentences *Cine va sosi primul va pleca ultimul?* 'Who will arrive first will go last?' and *Știi cine va sosi primul*? 'Do you know who will arrive first?'

The first MiP has a low tonal level, while the latter shows intonation specific to a yes-no question ending in a ^H* L% sequence for neutral utterances and a L* ^H% sequence for utterances expressing surprise. If the last word of the sentence is oxitonic, the boundary tone is always an H% type one and the pitch accent is of L+H* or H* type (Apopei et al., 2006).

2.2.2 Wh-questions with one wh-word/wh-expression

In all interrogative sentences where no morpho-syntactic context from the previously described predefined P1-P4 set is identified, the wh-words/wh-expressions are interrogative pronouns/ pronominal adjectives/adverbs, uttered with a high tone. These interrogations include:

a) Typical wh-questions with one interrogative word: (***Cine** va veni?* 'Who will come?', *Și **cine** crezi că va veni?* 'And who do you think will come?', *Deci, **cine** crezi că va veni?* 'So, who do you think will come?', *El oare **cu cine** va veni?* 'With whom will he come?', *Maria, vecina noastră, **cu cine** crezi că s-a întâlnit*? 'Whom do you

think Maria, our neighbour, met?', *Mă întreb, oare* ***cine*** *va sosi primul?* 'I wonder, who will arrive first?' ***Cine*** *cântă bine și dansează frumos?* 'Who sings well and dances nicely?', ***Cine*** *va veni să cânte primul?* 'Who will come to sing first?'.

When no comma is present, these types of sentences are composed of one or two clauses. The sentences with one clause are uttered in a single IP. The F0 contour of the sentences beginning with a wh-word/wh-expression (e.g. *Cine cântă bine?* 'Who sings well?') is descending, with a H* pitch accent on the wh-word. When the wh-word/wh-expression is preceded by other words, different from personal verb forms (e.g. *El oare* ***cu cine*** *va veni?* 'With whom will he come?'), the F0 contour is ascending-descending, starting from a medium tonal level, with a H* pitch accent on the wh-word/wh-expression. In both cases, the intonational contours end in an L% boundary tone. If the sentence is composed of two clauses (e.g. ***Cine*** *cântă bine și dansează frumos?* 'Who sings well and dances nicely?', Figure 3) the F0 contour corresponds to one IP with two major phonological phrases (MaPs), one for each clause. For the sentence in Figure 3, these MaPs are: (ip1) ***Cine*** *cântă bine* 'Who sings well' and (ip2) *şi dansează frumos?* 'and dances nicely'.

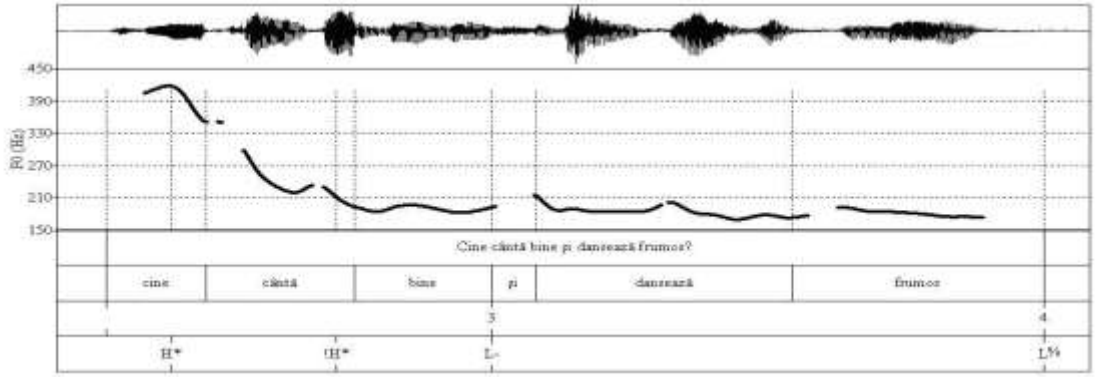


Figure 3: Waveform, spectrogram, F0 contour and prosodic phrasing of the sentence *Cine cântă bine și dansează frumos?* 'Who sings well and dances nicely?'

The interrogative pronoun *cine* 'who', placed at the beginning of the sentence, is clearly focused, while the rest of the intonational contour follows a descending trend followed by a flat trend. If the wh-word/wh-expression is placed (a) immediately after a comma (e.g. *Maria, vecina noastră, cu cine crezi că s-a întâlnit?* 'Whom do you think Maria, our neighbour, met?' Figure 4, first sentence) or (b) after a comma followed by one or more words different from personal verb forms and not followed by other punctuation marks (e.g. *Mă întreb, oare* ***cine*** *va sosi primul?* 'I wonder, who will arrive first?'), then a new IP is initiated after the comma. Usually, the intonation of these statements consists of two MaPs. In Figure 4, for both utterances, the first MaP (*Maria, vecina noastră* 'Maria, our neighbour' for the first sentence and *Vecina noastră* 'our neighbour' for the second sentence) includes the words that precede the wh-expression (*cu cine* 'whom') and is uttered with a low tone to allow a tonal rise on the wh-word and to lead to a wh-question type intonation on the second MaP.

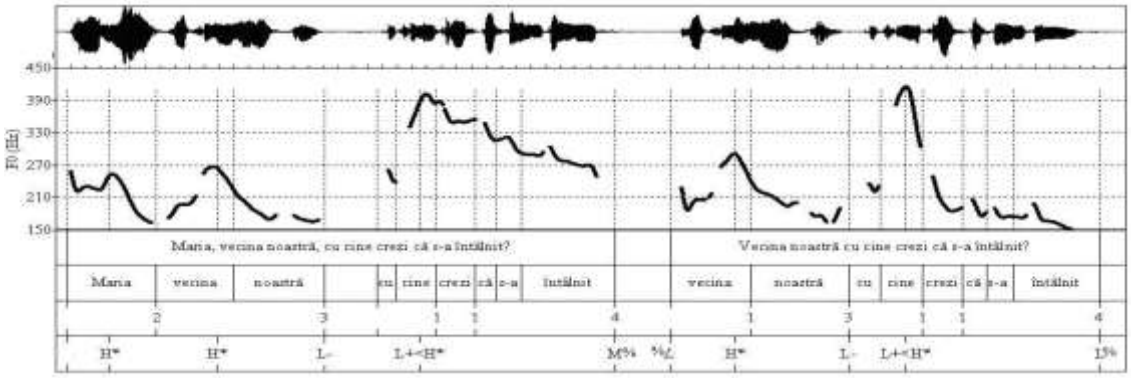


Figure 4: Waveforms, spectrograms, F0 contours and prosodic phrasing of the sentences *Maria, vecina noastră, cu cine crezi că s-a întâlnit?* 'Whom do you think Maria, our neighbour, met?' and *Vecina noastră cu cine crezi că s-a întâlnit?* 'Whom do you think our neighbour met?'

The first MaP has a statement type intonation, with an L- boundary tone. The second MaP has an L-L% type boundary tone when the utterance is neutral and a M-M% boundary tone when the utterance shows surprise.

b) Disjunctive questions (***Cine** va sosi prima, Maria sau Ioana?* 'Who will come first, Maria or Ioana?'*, Mă întreb, **pe cine** îl vei invita întâi, pe Ion sau pe Gheorghe*? 'I wonder, whom will you invite first, Ion or Gheorghe?'). The prosodic phrasing of the utterances corresponding to these types of interrogative sentences will be illustrated by the sentences *Cine va sosi prima, Maria sau Ioana?* 'Who will come first, Maria or Ioana?' and *Cine va sosi prima, Maria care locuiește în Cluj sau Ioana, care locuiește în Timișoara?* 'Who will come first, Maria, who lives in Cluj or Ioana, who lives in Timișoara?' (Figure 5). The differences between the first and the second sentence consist in noun phrase structures.

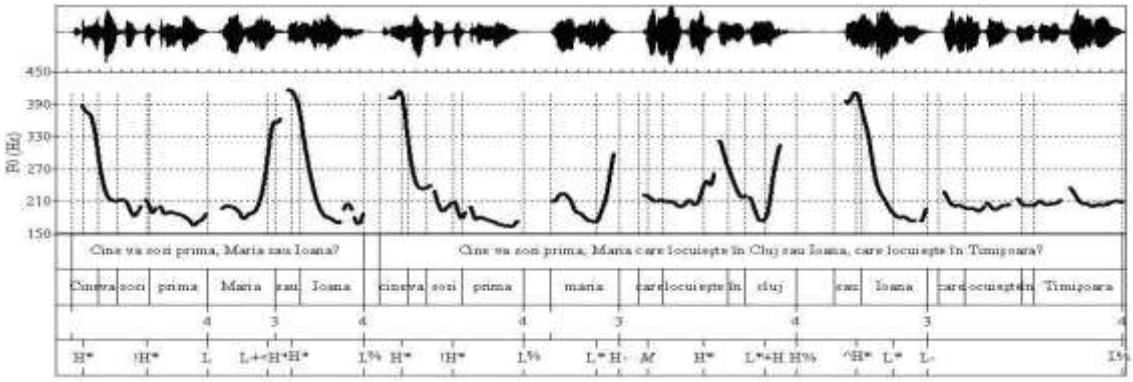


Figure 5: Waveform, spectrogram, F0 contour and prosodic phrasing of the sentences *Cine va sosi prima, Maria sau Ioana?* 'Who will come first, Maria or Ioana?' and *Cine va sosi prima, Maria care locuiește în Cluj sau Ioana, care locuiește în Timișoara?* 'Who will come first, Maria, who lives in Cluj or Ioana, who lives in Timișoara?'

The utterance of each one consists of three IPs. The first IP is specific to a wh-question and corresponds to an interrogative clause (*Cine va sosi prima* 'Who will come first'). The second IP corresponds to the first possible answer and its intonation is specific to a yes-no question composed of one or two MaPs, according to the structure of the noun phrase present in the corresponding clause (Apopei, 2014). When two MaPs are present,

each one ends in an H- boundary tone and has a yes-no question type intonation. The third IP corresponds to the second possible answer and its intonation is specific to a statement. This IP consists of one or two MaPs, depending on the structure of the nominal group (*Ioana* 'Ioana' or *Ioana, care locuiește în Timișoara* 'Ioana, who lives in Timișoara').

c) Echo questions, which are a particular type of total or partial interrogations, used by the speaker to ask his interlocutor to fully or partially repeat his/her last sentence *(− Am cumpărat unt de arahide.− Ce ai cumpărat?* 'I bought peanut butter. - What did you buy?', *Mi-au dat să mănânc ciuperci. − Ce spui că ți-au dat? –* 'They gave me mushrooms to eat. − What did you say they gave you?'). The intonation of the echo questions always expresses surprise. Its pitch track has a low-to-medium level stretch and a right peripheral rising movement starting after the last accented syllable and ending in an H% boundary tone (Jitcă et al., 2014).

d) Rhetoric questions, where the wh-word is uttered with a moderately high tone (***Cum*** *o să te părăsesc tocmai acum? Îți promit că nu o să te părăsesc.* 'How do you think I'll leave you now? I promise I won't leave you.'). Such sentences have a wh-question type intonation, but the tonal level has a smaller variation between the wh-word and the next word. This is why this type of question is perceived rather as a statement.

2.2.3 Multiple interrogations

These include: (a) interrogative sentences with two juxtaposed wh-words/wh-expressions (***Cine ce*** *a spus?* 'Who told what?', ***Cine unde*** *s-a dus?* 'Who went where?', ***Care cui*** *i-a dat invitație?* 'Who gave an invitation to whom?'); (b) copulatively coordinated interrogations (***Unde*** *și* ***când*** *ne întâlnim?* 'Where and when shall we meet?', ***Cine*** *și* ***de ce*** *s-a supărat?* 'Who got upset and c?', ***Unde*** *și* ***cum*** *ne distrăm?* 'Where and how shall we get fun?'; (c) coordinated interrogative clauses (***Unde*** *te duci și* ***când*** *te întorci?* 'Where will you go and when will you come back?', ***Când*** *ai venit și* ***de ce*** *nu m-ai așteptat?* 'When did you arrive and why didn't you wait for me?', ***De la cine*** s-a întors, ***cu cine*** va pleca și ***la cine*** va merge? 'From whom has he come back, with whom is he going to leave and to whom will he go?'); (d) interrogative sentences beginning with *cum, cum adică* or *cum așa* (all meaning 'how is that' or 'what do you mean by'), followed by a comma and a second wh-word/wh-expression (***Cum***, ***cine*** *e cel mai bun?* ***Cum*** *adică,* ***cine*** *e cel mai bun?* 'How is that, who's the best? 'What do you mean by who's the best?'), which corresponds to rhetoric questions. In all these interrogations (a-d), the wh-words/wh-expressions are interrogative, except for those in (d), where the latter wh-word is relative.

Regarding the interrogative sentences with two juxtaposed wh-words/wh-expressions, Dascălu-Jinga (2008) has suggested that the first wh-word has the strongest pitch accent. However, the utterances recorded by us from various speakers have shown that some of them focused the first wh-word, while others focused the second wh-word.

When at least two copulatively coordinated wh-words are present in a sentence, the last interrogative wh-word/wh-expression always carries the strongest focus, regardless of the length of the last interrogation. The previous interrogative wh-words/wh-expressions

are less focused, particularly when they are included in short interrogations. In some cases, their intonation can become similar to that of a relative wh-word/wh-expression. For example, in ***Unde***, ***când*** *și* ***cu cine*** *se va întâlni*? 'Where, when and with whom is he going to meet?' (Figure 6), *unde* 'where' and *când* 'when' are less focused compared with *cu cine* 'with whom'. On the other hand, in ***De la cine*** *s-a întors*, ***cu cine*** *va pleca și* ***la cine*** *va merge?* 'From whom has he come back, with whom is he going to leave and to whom will he go?', the last wh-expression is only slightly more focused than the previous ones.

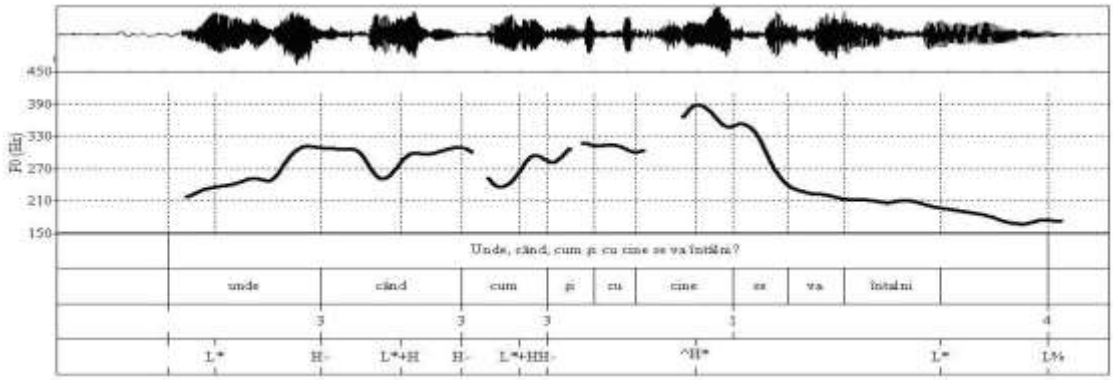


Figure 6: Waveform, spectrogram, F0 contour and prosodic phrasing of the sentence *Unde, când și cu cine se va întâlni?* 'Where, when and with whom is he going to meet?'

Each wh-word/wh-expression is placed in a separate MaP. Except for the first MaP, each MaP containing a wh-word/wh-expression starts after a comma or a coordinating conjunction.

An example of sentence including coordinated interrogative clauses (*Cozonacii pentru cine sunt și pachetele pentru cine le-ați pregătit?* 'For whom are these cakes and for whom have you prepared these packages?') is analysed in Figure 7.

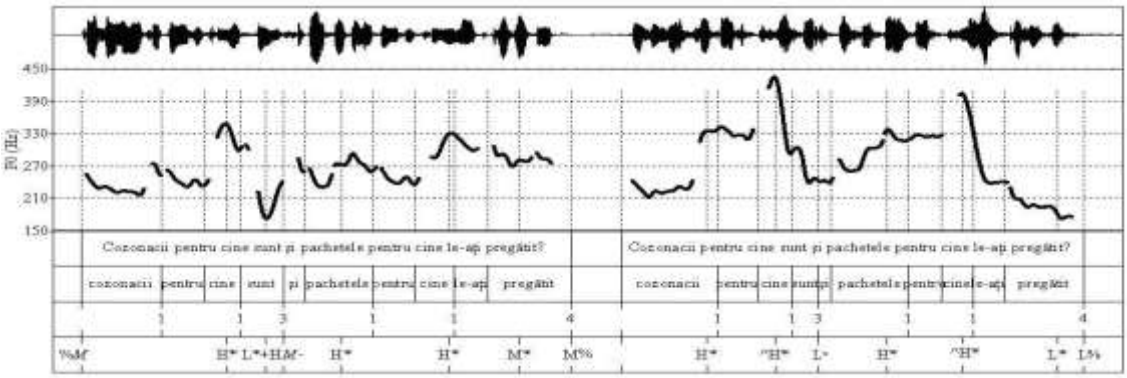


Figure 7. Waveform, spectrogram, F0 contour and prosodic phrasing of the sentence *Cozonacii pentru cine sunt și pachetele pentru cine le-ați pregătit?* 'For whom are these cakes and for whom have you prepared these packages?' uttered by two speakers

In such a sentence, the number of clauses is equal to the number of MaPs. Each MaP has a wh-question type contour which depends on the position of the wh-word in relation to the beginning of the phrase. The IP corresponding to the sentence in Figure 7 is composed of two MaPs: (MaP1) *Cozonacii pentru cine sunt* 'For whom are these cakes' and (MaP2) *și pachetele pentru cine le-ați pregătit* 'and for whom have you prepared the packages'. The analysis of this sentence has revealed the following: for

each MaP, the wh-word has the strongest pitch accent; for each MaP, the word before the wh-word has a pitch accent with a medium tonal level; the boundary tones of each MaP depend on the emotional state of the speaker. These boundary tones are low (L- or L%) for a neutral utterance and medium (M- or M%) for an utterance expressing curiosity.

3. Conclusion

The intonation analysis of questions that include wh-words has revealed that their intonational contours depend on the type of questions (yes/no question or wh-question), on syntactic structure and on the emotional state of the speaker.

The utterances analysed in section 2 of this paper and in (Apopei et al., 2006; Jitcă et al., 2014) have shown various degrees of the following emotional states: surprise, curiosity, persuasion, anger, sadness, fear, joy. These emotions are sometimes caused by previous sentences but there are some cases when they have no connection with the text and depend only on the emotional state of the speaker.

Hirst & Di Cristo (1998) remark that in Romanian "wh-questions are said to be more like emphatic declaratives and rising intonation is said to be rare". Our analysis reveals that the wh-questions with ascending boundary tones (M%, H%) correspond to utterances that express the speaker's surprise or curiosity.

This analysis has revealed that prosodic annotation levels (pitch accent, boundary tones and break indices) of speech corpus must be completed with an annotation level of affectivity/emotion of the utterance and annotation levels for text-to-speech alignment (syllable and word level). These annotation levels are required for more complex analysis: to explain the differences between intonation contours of the utterances of same sentence; to prosody prediction in text-to-speech systems.

Acknowledgments

The research presented in this paper has ben developed within the Institute of Computer Science of the Romanian Academy, Iasi branch. The prosodic hierarchy and functional labels used for describing F0 contours have been developed during the last years together with our colleague Doina Jitcă.

References

Apopei, V. (2014). About prosodic phrasing of the Noun Phrases in speech. In *Proceedings of the Romanian Academy*, 15:2, 2014, 200-207.

Apopei, V., Jitcă, D., Turculeţ, A. (2006). Intonational structures in Romanian Yes-No Questions. *Computer Science Journal of Moldavia Chişinău*, 14:1(40), 113-137.

Boersma, P., Weenink, D. (2014). Praat. www.fon.hum.uva.nl/praat/

Dascălu-Jinga, L. (2008). Organizarea Prozodică a Enunţului. *Gramatica Limbii Române II: Enunţul*, V. Guţu Romalo coord., Bucureşti: Editura Academiei Române, 946-991.

Beckman, M. E., Hirschberg, J., Shattuck-Hufnagel, S. (2005). *The Original ToBI System and the Evolution of the ToBI Framework*', The Phonology of Intonation and Phrasing, S-A. Jun (ed.), Oxford University Press, pp. 9-54.

Guțu Romalo, V. ed. (2008). *Gramatica limbii române*. Bucharest: Romanian Academy Publishing House.

Hirst D., Di Cristo A. (1998). *A survey of intonation systems*. In Intonation systems: a survey of twenty languages, Hirst D., Di Cristo A. (Eds), Cambridge: Cambridge University Press, 1-43

Jitca D., Apopei V., Paduraru O., Marusca S. (2014). *Transcription of the Romanian Intonation*. In Intonational Variation in Romance. Frota S. and Prieto P. (Eds.), Oxford: Oxford University Press, (to be printed in 2014).

Jitca D., Apopei V., Paduraru O. (2012). *A Romanian Prosody Prediction Module Based on a Functional Intonational Model*, Memoirs of the Scientific Sections of the Romanian Academy, Tome XXXV, 109-124.

Jitca D., Apopei V., Jitca M. (2009). The F0 Contour Modelling as Functional Accentual Unit Sequences. *International Journal of Speech Technology*, Volume 12, Issue 2-3, 75-82.

Selkirk E. (2005). *Comments on intonational phrasing in English*. In Prosodies: With Special Reference to Iberian Languages, S. Frota, M. Vigario, and M.J. Freitas (eds.), Berlin: Mouton de Gruyter, 11-58.

YET ANOTHER ROMANIAN READ SPEECH CORPUS

LAURA PISTOL

Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi, Romania
Institute of Computer Science, Iaşi Branch of Romanian Academy, Romania
laura.pistol@info.uaic.ro

Abstract

In this paper we describe the development effort of a Romanian continuous read-speech corpus (VoxRoCorpus), the procedure of voice recording, the (phonemes level) manual segmentation and labelling process. The corpus consists of a total of 6270 words from three female speakers, representing about 45 minutes of speech. The corpus is useful in any type of application where accurate phoneme boundaries are important.

Key words — speech corpus creation, manual segmentation and labelling.

1. Introduction

A speech corpus is a valuable (sometimes indispensable) resource in the spoken language technology field, but comes to a high price, both in time and money. The time requirement is high due to the long and slow process of development. The money cost is high due to the expert knowledge employed. Always welcomed are any kind of new resources development effort that lower the cost of creation for any less resourced languages.

A speech corpus construction requires at least the use of audio recording files (audio books, radio or TV recordings, telephone conversations or recording enquires sessions) and the corresponding orthographic text transcription. Regardless, but depending on, the purpose of development, the corpus has to fulfil requirements such as a minimum number of speakers or a minimum size in minutes.

There is a big interest in speech corpora acquisition for Romanian language (Stănescu et al., 2012, Watts et al., 2013, Stan and Giurgiu, 2010), but to our knowledge, there is only one speech corpus which is manually annotated to phoneme level and freely available (Teodorescu et al., 2014). The corpus is a collection of voice recordings grouped in basic sounds, emotional voices, aspects of double subject, pathological voices, gnathophonics and gnathosonics sounds and a set of labels for Romanian intonation. Although the corpus has more than 50 speakers, the recorded short sentences do not cover all the Romanian language set of phonemes. The corpus comprises no more than 200 minutes, and sentences are lengthened between one and nine words.

The main aim of this paper is to describe the creation of another continuous read speech corpus for Romanian language, along with aspects of phonetics and manual annotation used. Segmentation of speech into phonemes is useful for many spoken language applications like speech analysis, automatic speech recognition or speech synthesis.

The rest of the paper is organized as follows. Section 2 deals with the acquisition and manual annotation of the corpus. Statistical analysis is discussed about VoxRoCorpus in Section 3 and the paper ends with the conclusions drawn in Section 4.

2. Speech database development

The sentences selected to be recorded are extracts, punctuation marks including, from a well know novel, "Amintiri din copilărie" (*Childhood Memories*) by the classical Romanian writer Ion Creangă. The choice towards this classical work was imposed by the necessity for the text corpus to be copyright-free, in order to be able to make the corpus available for research purposes.

For recordings we chose three native Romanian language female speakers, born and educated in the middle area of the county of Moldova; they are all healthy persons, aged 33, 37 and 50 years old. Prior to the recordings, we have informed the speakers of our scientific objectives and they signed a form of consent.

All three speakers read the same set of sentences. The recommendations for the speakers were to read the utterances keeping a steady speaking voice, with consistent volume of sound, and as accurate pronunciation as possible of the older and regional specific Romanian language in the text.

The characteristics for the recordings are as follow:

- Performed in several sessions, with breaks, as often as needed, in a quiet room;
- A 22050 Hz sampling frequency and 16 bit resolution were used, stored as a single channel wave format.

Each utterance has associated three file types:

- .wav, the audio speech file;
- .txt, associate the orthographic transcription of the sentence (including punctuation marks) the speaker said;
- .TextGrid, a text file with labelling data for phonemes and words time aligned tiers in Praat system (Boersma et al., 2013).

2.1. Phonetic considerations

Romanian language is a phonetic language. Depending on the linguist assumptions, the Romanian phonetic system can vary in the number of phonemes (Vasiliu, 1965; Turculeț, 1999).

For our purpose we adopted the phonetic system employed by (Turculeț, 1999), merging seven vowels: /e/, /i/, /a/, /ə/, /ɨ/, /o/, /u/, the short vowel /*i*/, /ʚ/, four semivowels /e̯/, /i̯/, /o̯/, /u̯/ and 20 consonants.

Table 1 lists the standard IPA (International Phonetic Alphabet) symbols used in the annotation process (Section 2.2) along with some words examples (see Table 3 for information about the most frequently used phonemes in VoxRoCorpus).

Table 1: The phoneme set and symbols employed in labelling

IPA symbol	Word example	IPA symbol	Word example
e	soar**e**	tʃ	**c**eva
i	**i**nimă	Z	**j**oc
a	**a**casă	ʃ	u**ş**ă
ə	acas**ă**	ts	**ţ**ărnă
ɨ	**î**n	dZ	me**g**ieş
o	**o**m	d	**d**ar
u	**u**ns	t	**t**ârg
e̯	lum**e**a	f	**f**armec
i̯	**i**arnă	v	**v**esel
o̯	s**o**are	h	**h**az
u̯	dulă**u**	l	**l**acrimă
k	**c**ap	m	**m**are
c	o**ch**i	n	**n**ecaz
g	**g**raniţă	s	**s**ac
ɟ	**gh**em	z	**z**âmbet
b	**b**unica	r	**r**epede
p	**p**om	j	vrem**i**

2.2. Manual annotation

The purpose of the annotation is to mark the time, in the voice signal, of the start and the end boundaries of the phonemes along with the phonetic and orthographic labelling.

The manual annotation task was performed by the author, having an extensive experience in reading spectrograms and labelling phonemes. Using the graphical interface and listening to the audible track in Praat the author identified the acoustic changes in the voice signal in order to determine and mark the phoneme boundaries; using the relationship between the waveform, spectrogram and phonetic data, labelled accordingly. The recordings were annotated at phoneme and word levels. The phonemes are annotated in the International Phonetic Alphabet – IPA as described in Table 1.

A time aligned and labelled recording annotation is a text file in .TextGrid format, structured in tiers; each tier keeping the left and the right time boundaries for each interval. An excerpt of an annotation is listed below:

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 26.946485260770974
tiers? <exists>
size = 2
item []:
    item [1]:
        class = "IntervalTier"
        name = "phoneme"
```

```
        xmin = 0
        xmax = 22.830566893424038
        intervals: size = 257
        intervals [1]:
            xmin = 0
            xmax = 1.4393515166462594
            text = "#"
        intervals [2]:
            xmin = 1.4393515166462594
            xmax = 1.4739161705234223
            text = "s"
        intervals [3]:
            xmin = 1.4739161705234223
            xmax = 1.4885835148517317
            text = "t"
            ...
item [2]:
        class = "IntervalTier"
        name = "word"
        xmin = 0
        xmax = 26.946485260770974
        intervals: size = 71
        intervals [1]:
            xmin = 0
            xmax = 1.4393515166462594
            text = "#"
        intervals [2]:
            xmin = 1.4393515166462594
            xmax = 1.6289872623624797
            text = "stau"
        intervals [3]:
            xmin = 1.6289872623624797
            xmax = 2.3982426303854885
            text = "câteodată"
            ...
```

The Praat screenshot in Figure 1 exemplifies the annotation of a fragment of a sentence; "… (î)mi aduc aminte ce vremi şi ce oa(meni)..." (having the phonetic transcription /mɘ a d u k a m i n t e tʃ e v r e mɘ ʃ i tʃ e u̯ɒ/), in English "…(I) remember what times and what p(eople)...". The first row represents the waveform of the sentence excerpt, the second row is the spectrogram, the third row the phonetic transcription and the last row is the orthographic labelling at the word level.

Boundaries detection sounds like an easy task: using the graphical interface and listening to the audible track in Praat to identify the acoustic changes in the voice signal. The example from Figure 2a however shows that many difficulties can arise. The main difficulty faced during the annotation process was due to the presence of a continuous transition area, making difficult the placing of the boundary between a semivowel (/i̯/) and a vowel (/a/) in "…m i̯ a d u k..."; in this case we placed the boundary closer to the semivowel.

Comparing the two pictures from Figure 2 it can be seen that all phonemes boundaries are clearly defined in (b) spectrogram in comparison to (a) spectrogram. For lowering the chance of an error at the boundaries detection, which may occur due to voice

quality, we annotated all the sentences of each speaker at a time. This way we took advantage of the accumulated experience in "reading" one particular voice.

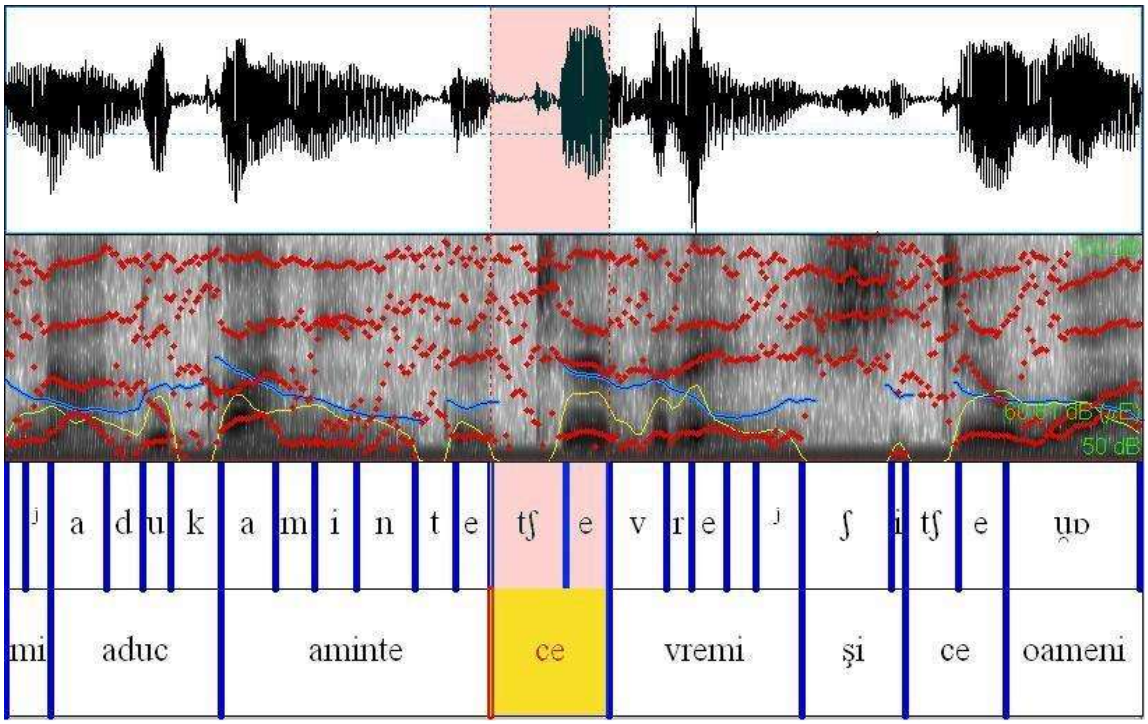
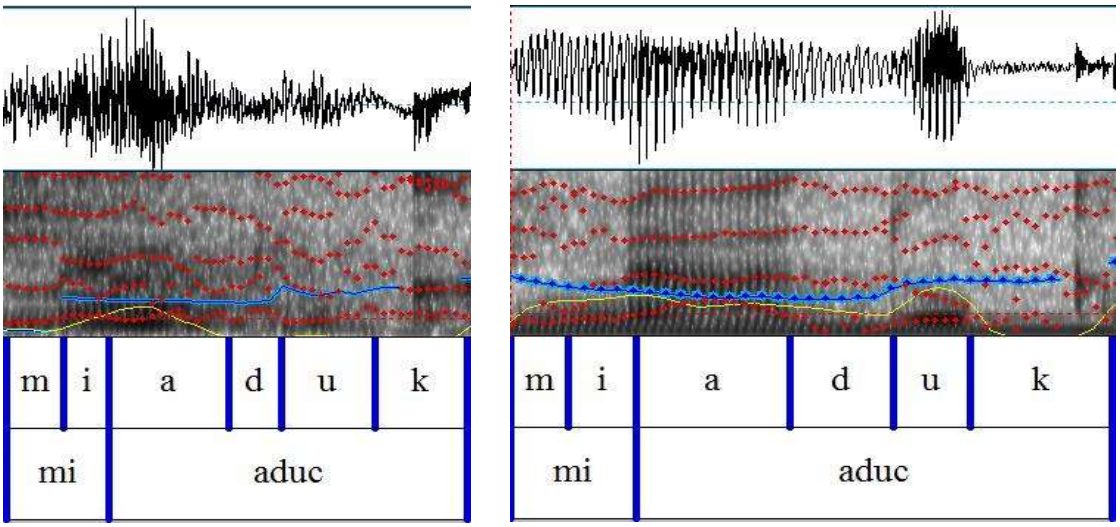


Figure 1: Praat screenshot depicting the annotation of an excerpt of an utterance:

"… (î)mi aduc aminte ce vremi şi ce oa(meni)..."

On average, the time employed for 30 seconds of speech took at least one hour of annotation (just for the annotation process, not including, for example, the time of audio signal gathering).



(a) (b)

Figure 2: Praat screenshot depicting the same utterance for two different speakers

3. Speech database analysis

In Table 2 we outline the distributional properties of the VoxRoCorpus.

Table 2: Distributional properties

#sentences	246
#unique words	819
#words	6270
AVG #words/sentence	25.48
MIN #words/sentence	1
MAX #words/sentence	86
AVG #syllables/word	2.28
AVG #phonemes/word	4.1

As it can be observed, the shortest sentence is one word long (an interjection) and the longest sentence includes 86 words. We mentioned earlier in the paper that the recorded sentences are extracts from a novel keeping the punctuation marks; this was in order to obtain natural read speech recordings in very long sentences.

For the phoneme frequency occurrences a statistical analysis was carried out. Table 3 compares the analysis of phonetic coverage of our corpus and the corpus developed by Horia Cucu (2011), taken as reference distribution. The major difference is that our corpus is manually annotated at the phoneme and word level while the other is not. Although the corpora differ (very much) in the number of phonemes, it might be noticed that eight from the ten most frequent phonemes of our corpus can be found among the ten most frequent phonemes of H. Cucu's corpus.

Table 3: The top ten most frequent phonemes

	VoxRoCorpus (13506 annotated phonemes)		**Cucu, 2011 (865 million phonemes)**	
	Phoneme	**%**	**Phoneme**	**%**
1	i	8.3	e	10.91
2	e	7.1	a	9.52
3	r	7.1	i	7.79
4	n	6.7	r	7.25
5	t	6.6	t	6.46
6	u	6.3	s	6.26
7	a	6.1	n	6.25
8	ă	5.9	u	5.44
9	l	5.5	l	4.61
10	c	4.2	o	4.38

As a note, eight from the most frequent phonemes cover more that 50% of all phonemes occurrences.

4. Conclusions

This paper reports the effort employed in the development of a continuous read-speech corpus for Romanian language. The corpus consists of a total of 6270 words from three female speakers, representing about 45 minutes of speech recordings. Manual annotation has been made; all the sentences of a speaker were annotated at a time.

The corpus is useful in any type of application where accurate phoneme boundaries are important.

An extent in speakers' number will increase the corpus value, as well as annotating an additional set of phonetically rich sentences.

Acknowledgements

This research has been part of an internal theme of the Institute of Computer Science, Romanian Academy, Iaşi Branch, Project: Cognitive Systems and Applications.

References

Boersma, P., Weenink, D. (2013). Praat: doing phonetics by computer. A computer program version 5.3.42, from www.fon.hum.uva.nl/praat/.

Cucu, H. (2011). *Towards a speaker-independent, large-vocabulary continuous speech recognition system for Romanian*. PhD. Thesis, Politehnica University from Bucureşti.

Stan, A., Giurigu M. (2010). Romanian language statistics and resources for text-to-speech systems. *ISETC* 2010, Timişoara, Romania.

Stănescu, M., Cucu, H., Buzo, A., Burileanu, C. (2012). ASR for low-resourced languages: Building a phonetically balanced Romanian speech corpus. In *Signal*

Processing Conference (EUSIPCO), Proceedings of the 20th European, 2060-2064.

Teodorescu, H. N., Feraru, M., Zbancioc, M., Trandabăţ, D., Ganea, R., Verbuţă, A., Hnatiuc, M., Voroneanu, O., Pistol, L., Untu, A., Păvăloi, I., Apopei, V., Jitcă, D., Păduraru, O. (2014). *SRoL – Sunetele limbii române,* www.etc.tuiasi.ro/sibm/romanian_spoken_ language.

Turculeţ, A. (1999). *Introducere în fonetica generală şi românească*. Iaşi: Demiurg Editorial House.

Vasiliu, E. (1965). *Fonologia limbii române*. Bucureşti: Editura Ştiinţifică.

Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi J., King, S. (2013). Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. In *Proceedings of the 8th ISCA Speech Synhesis Workshop*, Barcelona, Spain.

LEXICAL NESTING DICTIONARIES

CĂTĂLINA MĂRĂNDUC[1], CĂTĂLIN MITITELU[2], CENEL AUGUSTO PEREZ[3]

[1] *"Iorgu Iordan – Al. Rosetti" Institute of Linguistics, Bucharest, Romania, catalina_maranduc@yahoo.com*
[2] *Stefanini, Bucharest, Romania, catalinmititelu@yahoo.com*
[3] *"Al. I. Cuza" University, Computer Science Faculty, Iaşi, Romania, augusto.perez@info.uaic.ro*

Abstract

The analysis of the derived or compound words, with the re-motivation of the linguistic sign, is the process by which the lexical information is stored in the speaker's competence. For specialists, the information organized in this way offers a clear perspective on etymology and/or on the meaning filiation. But, along with the simultaneous appearance on the market of the dictionaries organized alphabetically and those organized in nests, we will need, in order to compare them, computer programs that will transform them into a structured format and align them at the word level. We will show what the differences are between words formed by affixes and those by affixoids and why we developed two dictionaries, adapted to the specific problems of these two ways of word formation. The general tendency in the Romanian lexicography will have to follow the tendencies everywhere else in the world, i.e. to simplify the structure of the dictionaries and to include as much information as possible in the smallest space, easily accessible for the user. Therefore, we will need simpler and more flexible parsers than the present ones developed for the academic dictionaries. They should be parsers adapted to some dictionaries with a simpler and more logical structure. We propose here such a parser, and then we show how we have adapted it to the structure of the two dictionaries organized on nests and what difficulties we had in naming some fields with labels belonging to TEI standards, as well as the solutions we could find.

Key words — affixes, affixoids, compound words, derived words, dictionary entry parsing, lexical nesting, speaker's competence.

1. Why Nested Dictionaries?

1.1. The lexical nesting from the language philosophy perspective

The lexical nesting of the dictionaries was frequently used in the 19th century and in the first half of the next century; then it was considered old-fashioned. But, recently, some dictionaries, especially etymologic ones, have used this way of organizing material.

It offers more information about the origin and the structure of the words or about the affinities of the words in different languages. But we think it is more than that, it is the way in which the lexical units are organized in the speaker's competence (Mărănduc, 2012).

Of course, nobody can believe that the lexical organisation of the speaker's competence could be an alphabetical one. Speakers get in contact with words from the first months of their lives and the alphabetical order is a much later acquisition. If we are to believe that the speakers' vocabulary is organized according to the level of generality of the terms, this will mean that it organizes by exclusively semantic criteria, but practical examples show it is not like that.

Theoreticians of the linguistic sign show that it has two sides, a written or sonant form and a meaning, and the relation between them is unmotivated, arbitrary; a form expresses a certain meaning by convention. If we take words with identical meaning from different languages, we notice they can have totally different forms. And yet, the first forms of writing were suggestive pictograms, motivated linguistic signs. We will use Emile Benveniste's (1966: 49-55) work to briefly solve the matter in discussion. He shows that, although the linguistic sign is unmotivated in its essence, the speakers perceive it as motivated. They tend to re-motivate it, to build a new relation between form and meaning, instead of the original one, which, because of time and generalisation of the sign, got lost.

For example, from the words: *lucrător* "worker", *cititor* "reader", *scriitor* "writer", *muncitor* "worker", etc., the speaker notices that the final segment *-tor* is connected to the meaning "person that fulfils the action named by the verb". In the same way, from the words *reface* "redo", *rechema* "recall", *reîncepe* "restart", *redeveni* "rebecome", *reciti* "reread", etc., the speaker draws the conclusion that the initial segment *re-* is connected to the meaning "the repetition of the action named by the verb" and so on.

This type of observations can help the speaker to understand the meaning of a new word which he/she has never met before, combining the meanings of the component parts that he/she recognizes in the structure of the new word. The structure of the new word is not without relation with the meaning, but there is a partial motivation of the way in which it is formed. If the speaker encounters the word *asigurator* "insurer" for the first time, he/she will assign to it the meaning "person that insures; that makes insurances". If the speaker finds the verb *a retehnologiza* "retechnologise" for the first time, he/she will assign the meaning "to technologize again".

We can observe that any linguistic innovation, folkloric etymology or personal creation, is produced by simultaneous analogy at the form and meaning levels. That is why, we consider the study of the natural language should be done, for a better understanding of its functioning, starting from the lexical or family nesting that group words both by form and by meaning.

1.2. Lexical nesting as efficient method of research in present linguistics

Contemporary linguists are tempted to organize the lexicographic material by nests in order to study more easily the etymology of the words, the affiliation of the meanings or both of them. Important conclusions will be reached if we form families from Romanic words evolved from a Latin root or if we align, in a similar way, lexical nesting dictionaries of other related languages. In the Romanian language one cannot find only nests with Latin roots, but also with roots from old Slavic.

Otherwise, lexicographers, for the writing of a new dictionary, need to consult a great number of dictionaries appeared previously, to compare definitions, semantic trees, examples, different etymological solutions, to notice the way in which the semantic scheme of a certain word, done by lexicographers, has evolved in time. But the comparison of two dictionaries, organized differently, one by nesting, the other alphabetically, is not at all trivial. We need to align at the level of entry different organized word corpora, fact that needs the help of computer scientists. Due to this reason, like a helpful instrument for working at a new edition of DTLR (the thesaurus Dictionary of Romanian Language), the CLRE (Reference Electronic Lexical Corpus) project was created at the "A. Philippide" Institute of Romanian Philology and facilitates the work of lexicographers with a corpus of (at least) 100 dictionaries. Besides the extraction of the definitions of the word looked up in all the aligned dictionaries, the program contains also the option to extract the words from the lexical family of the searched word. Thus, the problem of the nests concerned the authors of this project.

The concern for the study of derivation is present in many researchers, philologists and computer scientists. The computerized synthesis of the derivation makes the object of a project which developed at the Mathematics and Computer Science Institute from Chișinău and which automatically generated derived words and then checked on a Romanian web corpus the existence in use of the generated words (Petic, 2010). Barbu Mititelu (2013), from the Research Institute for Artificial Intelligence of the Romanian Academy from Bucharest, semi-automatically found derived words in the Romanian wordnet.

2. Reasons for creating two lexical nesting dictionaries

2.1. DERCU – DER(ived Words Dictionary Organized by) CU(Lexical Nesting)

The dictionary was developed simultaneously with the first volume of the new academic etymological lexical nesting dictionary, DELR AB. We manually took over, based on the etymological information from the end of the articles, all the words of the derivative type relations from MDA, an academic dictionary in 4 volumes, organized alphabetically, intended to sum up the lexical information from DTLR, adding also new words that do not occur there.

Synthesizing these multiple concerns for derivation, we can realize that, unlike the researchers mentioned above, the manuscript dictionary, developed by us, is not based on the electronic synthesis or the analysis of the derived words form, but it is developed by an expert, based on the etymological solutions from MDA. It can be considered a gold corpus of the derivation in the Romanian language, with which the generated derivatives to be compared or electronically extracted from the corpus to improve the performances of this program.

This would be just one of the research areas in which our dictionary plays a part. But such an adventure can be put into practice only after the parsing of the dictionary edited in a word editor and after its turning into a retrievable linguistic resource for computer scientists. This is why in this paper we deal with the conversion into a structured format (XML) of two lexical nesting dictionaries.

2.2. DECO – (Word)D(ictionary formed with) E(lements of) CO(mposition), organized by lexical nests

Another dictionary, in progress, developed on the same lexicographic material, MDA, organizes in nests words formed with elements of composition, named, by some researchers, suffixoides, prefixoides, affixoides. Affixoides are similar to suffixes and prefixes. Like these, they are put at the beginning or the end of the word, forming a new word. Let us see what the difference is.

There is a small number of prefixes for the Romanian language[1] and a bigger number of suffixes, varying from one researcher to another. Working out the clues from the end of the SMFC volumes, we obtain a list of 91 prefixes and 885 suffixes for the Romanian language[2]. They are old in the language, the prefixes coming usually from Latin prepositions and their meaning is very abstract (relational, establishing a rate between the meaning of the derivative and the meaning of its base).

Otherwise, affixoides are recent, they belong to the scientific vocabulary, and they are scholarly borrowings from Latin and Greek, where they were words with full meaning, autosemantic: nouns, verbs, adjectives and adverbs. The same affixoid can be attached sometimes in front or at the end of the root word. Anytime we can borrow from these languages a new affixoid. Their number is potentially equal with the number of old launguage words.

For example: *arhimandrit* "archimandrite" is formed by the same composition element like *ecleziarh* "ecclesiarch", ***arhi-, -arh***. Its meaning is "master, leader", the indirect etymology is < gr. ***archein*** "to lead". But the formant with the superlative meaning, i.e. relational, "very", coming from the modern languages in words like: *arhiplin* "overloaded", *arhiaglomerat* "overcrowded", *arhisuficient* "oversufficient", is positioned only in front of the word on which is attached. This means it is a prefix homonymous with the composition element.

A lexicographic paper dedicated to compound words with composition elements is (Andrei, 2003). From the preface of this dictionary, in its second edition, we find out that it contains 35 000 neologisms created with approximately 4 000 composition elements. The author estimates that, based on these words, where suffixes and prefixes are added, 75 000 derivatives from the technical and scientific language can be analysed.

Therefore, the dictionary of words formed by composition elements is one of neologisms, while the one with derived words is one of general use. For understanding the meaning of derived words, the forming mechanism is important, that is the direct or intuitive etymology, the analysability for the speakers. Whereas for understanding the meaning of composition elements, the indirect etymology is important or more precisely the meaning which the composition element had in the Latin or Greek language (in which it originates). Depending on how much it is known about this dictionary, the neologisms can become analysable for the speakers, like derivatives.

[1] The table with the analysable prefixes in the Romanian language was finished in the second volume of the academic treaty of word formation. It contains 86 prefixes (FC II: 305-308).
[2] *The index of I–V volume*, in SMFC V: 91–99, *The index of VI volume*, in SMFC VI: 161–168.

The two types of word formation pose different problems, so that we decided the developing of two different dictionaries with different structures. The developing of DECO started simultaneously with DERCU, also in word editor, requiring the same effort of transposition in electronic format. So, a parser flexible enough to transpose in electronic format the two dictionaries, whose structure is not so complex, similar but also with important differences, was needed.

3. DEPAR – D(ictionary) E(ntry) PAR(ser)

DEPAR is a framework written in Java for helping computational linguists to parse their dictionary entries. It is made of multiple modules for permitting developing grammars for different kinds of dictionaries. Each such dictionary should have a specific ANTLR4 grammar attached that describes its entries and transforms them to XML format. The adopted XML schema is that developed by TEI (http://www.tei-c.org/index.xml), version TEI P5 (2007). ANTLR4 is a parser generator for LL(*) grammars which are more intuitive and flexible than LR grammars.

3.1. Description

The process consists of successive content transformations from .doc to .xml structure. The transformations mentioned above are the following:

- a .doc to .html conversion for extracting the dictionary content into an easily processing format but keeping the information regarding applied text styles. This step is ensured by a toolkit named Apache Tika (http://tika.apache.org/) which is able to detect and extract text from various document formats: .doc, .pdf, etc.
- a .html to .txt/.feat conversion ensured by an HTML parser written in ANTLR4 which is able to decompose the .html file into content (.txt) and styles or features (.feat).
- a .txt/.feat to .xml conversion ensured by a dictionary specific ANTLR4 grammar able to describe the dictionary entries and to transform them into a TEI P5 structured content.

This approach is different from the traditional one which consists in writing LL grammars that transform directly .html to .xml. A legitimate explanation for the traditional approach is that the style information may be very relevant in deciding which structure a certain sequence of tokens may enter. But we encountered a series of impediments mainly related to human errors that may be partially avoided using this new approach. We will describe them in the following subsections.

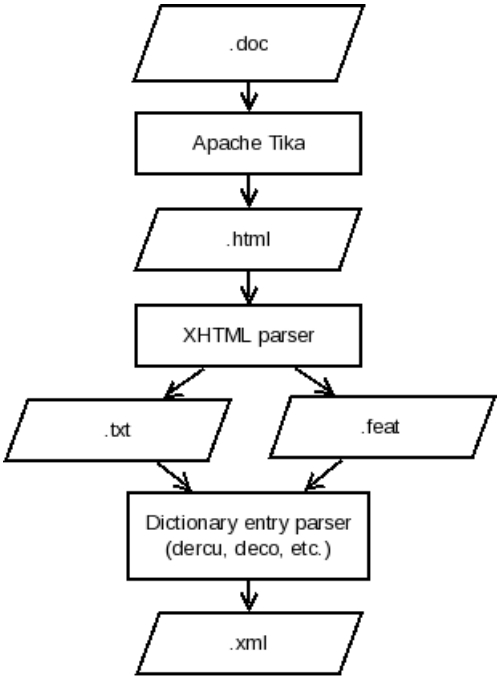


Figure 1: Content transformation diagram

3.2. Why Tika?

In the majority of cases we encountered dictionaries written in Microsoft .doc format, and Microsoft Word is able to export .html files. But the exported files contain a lot of HTML tags, useless for parsing, that should be cleaned. Moreover, the exported files may not be well-formed: either there are no closing tags (the HTML format permits this) or the closing tags are not closed in reverse order. For instance,

```
<b><i>some text</b> other text</i>
```

should become

```
<b><i>some text</i></b><i> other text</i>
```

in a well-formed structure.

Apache Tika is able to export XHTML (a strict HTML) content from different file formats, the resulting file being the minimum necessary for parsing. Moreover, being a Java framework it integrates easily with other Java applications.

3.3. Content versus features

The difficulty of parsing dictionary entries resides in the fact that starting from an unstructured content and using some rules and conventions more or less dictionary specific, the aim is to obtain a structured content. In other words, this is an enrichment process. The parsing rules are written using so called key-words, abbreviations, numbers and symbols: what we generally call tokens. But, in many cases, the style information is very useful for deciding in what structure a certain sequence of tokens should be added. So, starting from an `.html` file and targeting an `.xml` file the ANTLR4 grammar should contain rules about HTML tags and attributes. Actually, the

dictionary entry parser represents an HTML parser enriched with content or dictionary specific rules. Doing the same thing for many dictionaries will be hard to maintain. The solution is to parse the `.html` file and to separate the content (the `.txt` file) from the style (the `.feat` file). This way the parser may avoid a series of human or conversion errors like:

- applying styles (bold, italic) on whitespaces when not necessary, which are not so evident for human eyes.
- the same style information exported/saved in different ways. For example, the bold or italic styles may be applied in HTML in two different ways: with specific tags `<b>` and `<i>` or with style attribute attached to `<span>` tag. This may complicate the grammar by adding many rules describing the same thing.

The `.txt` file represents simply the content of the `.doc` file without style information, while `.feat` file contains a list of all text regions having the same style. We generalized the notion of style to the notion of feature. This way, the ANTLR4 grammar will be reduced to the content parsing and when it needs style information it will interrogate the feature list. The grammar rules are clearer and simpler, having nothing to do with HTML, and when they need to disambiguate based on style they will call a semantic predicate. ANTLR4 is able to activate or deactivate the lexer and parser rules using such semantic predicates.

3.4. Related works

The approaches related to this domain may be classified in two distinct categories:

- The traditional way that adopted the idea of developing rules describing dictionary entries (Neff et al., 1989; Tufiș et al., 1999; Hauser and Storrer, 1993; Lemnitzer and Kunze, 2005; Curteanu et al., 2004; Ion, 2008).
- An alternative way was proposed by the Iași team with the main aim of detecting the sense tree leaving the deep structure of the sense as optional (Curteanu et al., 2008, 2010, 2012).

Both dictionaries (DERCU and DECO) are very simple without complicated sense trees, thus a traditional approach being reasonable.

4. The structure of the two dictionaries

4.1. DERCU

Its structure is different from other previously parsed dictionaries by the fact that it has more types of entries. We distinguished three optional levels in the entries hierarchy (see Figure 2 below):

- "super stub" word (always present);
- stubs of a subfamily (optional), i.e. a derived word from the stub and from which a rich derivative tree is formed;

- simple entries (might be contained by a "super stub" and/or by a stub of a subfamily).

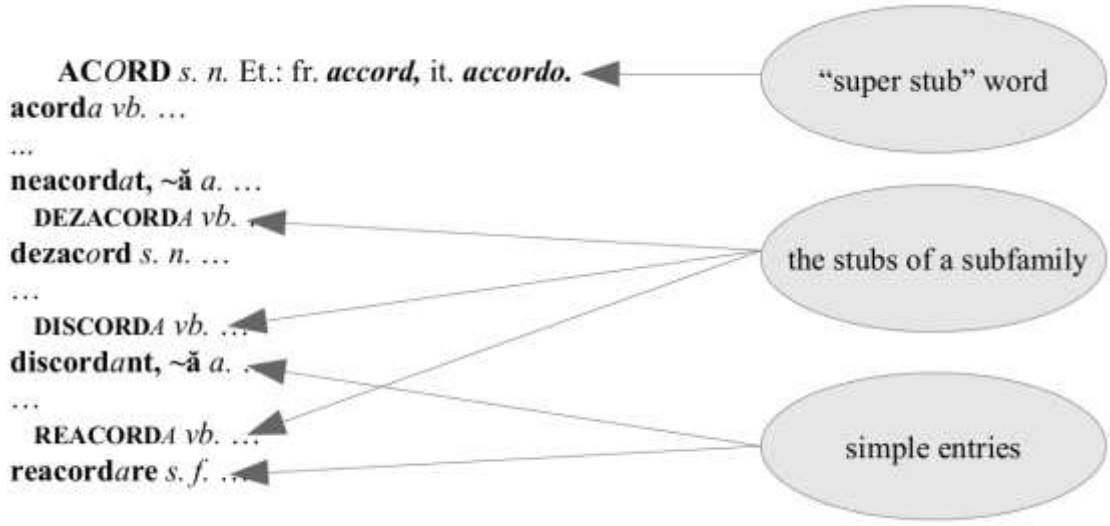


Figure 2: An example of entries hierarchy for the word *acord* "accord"

The "super stub" is not a derived word, so its definition does not constitute the object of this dictionary. In the case of some rare words, a synonym is given, enclosed between converted commas. It has only one etymology, the scientific one.

Both the stub of subfamily and the entry have definitions, with a minimum number of meanings, maximum 10-15, on a single level of generality. Some meanings can contain a phrase unit written with a different body than the proper definition. The definition is of two types: expressed and unexpressed, abbreviated as DD (derivative definition).

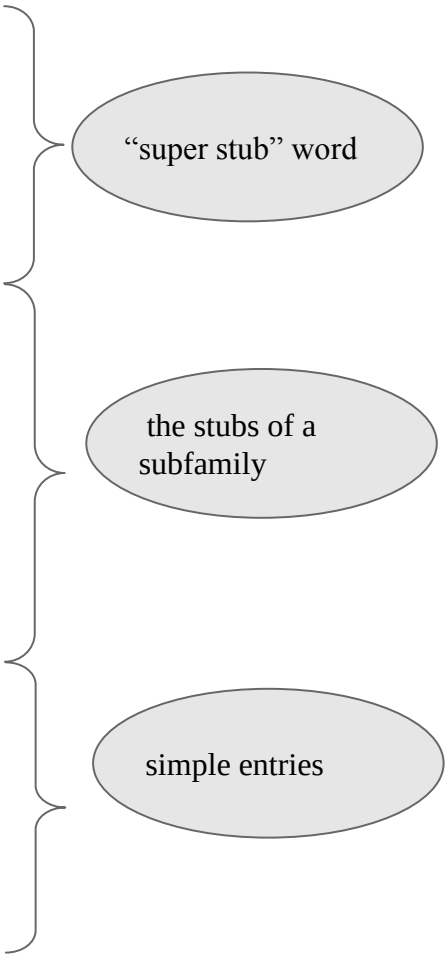
We call derivative definition (Vasiliu, 1981) a definition that does not mention the lexical meaning of the word, but only its semantic relation with the derivation base word, whose semantic explanation it sends to.

The coherence and the economy of the style require that the semantic traits common to the family to be glossed only once, at the first derivative (the order of the simple entries is an alphabetical one). Other derivatives get expressed definitions only if, besides the semantic traits common to the family, they have also specific traits.

Another problem that we have dealt with, unpredicted in TEI standards, is the nonexistence of two or more etymological types. For the lexical families, what matters is the analysable character of the form of the word, in the derivation base and in the affix, by the speaker. Therefore, a first etymology would be the intuitive one, the analysis of the word that points out the relation with the derivation base word, stub or family member. Only when this etymology is not the one accepted by the specialists, it is followed by the scientific etymology (this fact sends us to a foreign language, meaning that there is no derivation in Romanian, but that the derivation was produced in that language).

Therefore, the parsing of the three lexemes from above would look like this:

```
<superEntry type="nest">
<entry type="super stub">
<form><orth>ACORD</orth></form>
<gramGrp><pos>s. n.</pos></gramGrp>
<etym>fr. accord, it. accordo.</etym>
</entry>…
<superEntry type="subfamily">
<entry type="stub">
<form><orth>dezacorda</orth></form>
<gramGrp><pos>vb.</pos></gramGrp>…
<etym type="intuitive">des- + acorda.
</etym>
<etym>fr. désaccorder.</etym>
</entry>…
<entry>
<form><orth>dezacord</orth></form>
<gramGrp><pos>s. n.</pos></gramGrp>…
< etym type ="intuitive"> dezacorda –
-a. </ etym>
<etym>fr. désaccord.</etym></entry>…
</superEntry>
</superEntry>
```



4.2. DECO

A new word entered in a language becomes part of an existing family, it builds itself one, or it disappears, if it cannot be enclosed in a system by the speakers, i.e., if it is not analysable for the users. The idea from which we started is that to make the neologisms easy for decoding and entering the language vocabulary means making them analysable.

That is why we introduced in the dictionary new prefixes also, which are not sufficiently present in the speakers' competence and which combine very often with the composition elements. It is a dictionary with a simpler structure. It has only two hierarchy levels for the word entries. The stub entries are not independent words, but composition elements or new prefixes.

In order for the simple entries (neologisms) to be analysable, since a part of them is a Latin or Greek word, the speaker needs to know its meaning, which contributes to the formation of the new word meaning.

The composition element or the stub prefix is followed by definition and etymology. The traditionalist researchers forbid indirect etymology, which is the only interesting one here. Therefore, after the direct etymology, from dictionaries, we added not only the Greek or Latin word from where it comes, but also its translation.

The simple entry words are followed by definition, in which the meaning of the composition element, the one from the indirect etymon, and of the form analysis, meaning the intuitive etymology, is retrieved. Thus, the structure of the neologism is clear for the user; both the meaning and its form become analysable.

Examples of entries in DECO:

```
<superEntry type="nest">
<entry type="stub">
<form>ACUA-</form><form>ACVA-</form>
<gramGrp><pos>elem. comp. prim.</pos></gramGrp>
<def>„apă”.</def>
<etym>fr. aqua-</etym><etym> it. acqua-</etym><etym type
="source">&lt; lat. aqua. </etym><def>„apă”.</def>
</entry>
<entry>
<form><orth>acvanaut</orth></form>
<gramGrp><pos>s. m.</pos></gramGrp>
<def>Persoană care explorează mediul subacvatic.</def>
<etym type="intuitive">acva- + -naut.</etym>
</entry>…
</superEntry>
```

We should notice that at the development of the categories of the dictionaries, according to the TEI standards, linguists and computer scientists meet serious difficulties when parsing the etymologic section. We are still looking for better solutions for the parsing of this section. We should make aware those that bring these standards up-to-date that we need a bigger number of categories in this section.

5. Conclusions

We offered an improved and complete solution for the traditional approach regarding dictionary entry parsing using exclusively free Java based tools. The improvement consists in decoupling HTML parsing by focusing only on concrete dictionary entry parsing. This way the grammar is easy to maintain and ANTLR4 facilities as supporting a LL(*) grammar, semantic predicates, encouraging parsing actions (Java code) in listeners, etc. have a big impact as far as this aspect is concerned. Moreover, there is no human intervention during the entire process of conversion from `.doc` to `.xml`, grace to Apache Tika, a controllable and predictable toolkit.

Acknowledgements

The authors are grateful to all the people that have read the various preliminary works for the present research and that have made constructive observations: Emanuel Vasiliu, Laura Vasiliu, Theodor Hristea and, lately, Dan Cristea, Ioana Zirra, Verginica Barbu Mititelu.

References

Andrei, N. (2003). *Dicţionar etimologic de termeni ştiinţifici. Elemente de compunere greco-latine, ediţia a doua, revăzută şi adăugită*. Bucureşti: Editura Oscar Print.

Barbu-Mititelu, V. (2013). *Reţea semantico-derivaţională pentru limba română*. Bucureşti : Editura Muzeului Naţional al Literaturii Române.

Benveniste, E. (1966). *Problèmes de linguistique générale*. Paris: Gallimard.

Curteanu, N., Amihăesei, E. (2004). Grammar-based Java Parsers for DEX and DTLR Romanian Dictionaries. *ECIT-2004 Conference*, Iasi, Romania.

Curteanu, N., Moruz, A., Trandabăţ, D. (2008). Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing. In *Proceedings of the workshop on Cognitive Aspects of the Lexicon* (COGALEX 2008), Manchester, 2008, 55-63.

Curteanu, N., Moruz, A. (2012). A Procedural DTD Project for Dictionary Entry Parsing Described with Parameterized Grammars. *COGALEX-III – The Third Workshop on Cognitive Aspects of the Lexicon*, (COLING 2012), Bombay, India, 127-136.

DELR AB (2011). *Dicţionarul etimologic al limbii române,* vol. I, A–B, Bucureşti: Editura Academiei.

DTLR (Dicţionar Tezaur al Limbii Române) = DA (1913–1949) *Dicţionarul limbii române*, Bucureşti, Socec, Universul, 1913–1949, + DLR (1965–2010) *Dicţionarul limbii române.* Serie nouă. Bucureşti: Editura Academiei.

FC (1989). *Formarea cuvintelor în limba română*. Bucureşti: Editura Academiei, vol. I, *Compunerea,* de Fulvia Ciobanu şi Finuţa Hasan, 1970, vol. II, *Prefixele*, 1978, vol. III, *Sufixele,* I, *Derivarea verbală* de Laura Vasiliu, 1989.

Hauser, R., Storrer, A. (1993). Dictionary Entry Parsing Using the LexParse System. *Lexikographica 9*, 174-219.

Ion, R. (2008). Segmentarea în unităţi textuale atomice a intrărilor din dicţionarul limbii române în vederea analizei structural. In *Lucrările atelierului Resurse lingvistice şi instrumente pentru Prelucrarea limbii române,* Editura Universităţii „Alexandru Ioan Cuza" Iaşi, 75-82.

Lemnitzer, L., Kunze, C. (2005). Dictionary Entry Parsing, ESSLLI 2005.

Mattmann, C.A., Zitting J. L. (2011). *Tika in Action*. Manning Publications Co., 2011.

Mărănduc, C. (2008). *Familia de cuvinte.* Bucureşti: Editura Lucman.

Mărănduc, C. (2012). Derivation – a way to organize the vocabulary in lexical nesting. In *Conference held at the Romanian Academy*, 13th of December 2012.

MDA (2003). *Micul dicţionar academic*, vol. I–IV, București: Editura Univers Enciclopedic, volumul I: A-C, 2001, volumul al II-lea: D–H, 2002, volumul al III-lea: I–Pr, 2003; volumul al IV-lea: Pr–Z, 2003.

Neff, M. S., Boguraev, B. K. (1989). Dictionaries, Dictionary Grammars and Dictionary Entry Parsing. In *Proceedings of the 27rd Annual Conference of the Association for Computational Linguistics*, 91-101.

Parr, T. (2012). The Definitive ANTLR 4 Reference. The Pragmatic Programmers.

Petic, M. (2010). Mecanismele generative ale morfologiei derivaționale. În *Lucrările conferinței Resurse lingvistice și instrumente pentru prelucrarea limbii române,* Editura Universității „Alexandru Ioan Cuza" Iași, 195–202.

SMFC (1972). *Studii și materiale privitoare la formarea cuvintelor în limba română*, București: Editura Academiei, vol. I, 1959; vol. II, 1961, vol. III, 1962, vol. IV–V, 1967-1969, vol. VI, 1972.

Tufiș, D., Rotariu, G., Barbu, A. M. (1999). TEI-Encoding of a Core Explanatory Dictionary of Romanian. *Ferenc Kiefer, Gábor Kiss, and Júlia Pajzs (eds.), Proceedings of the 5th International Workshop on Computational Lexicography* (COMPLEX 1999), Linguistics Institute, Hungarian Academy of Sciences, Pecs, Hungary, 219-228.

Vasiliu, L. (1981) Asupra sinonimiei derivatelor sufixale în limba română. In *Semantică și semiotică,* ed. I. Coteanu și Lucia Wald, București: Editura ştiinţifică şi Enciclopedică, 314-344.

DICTIONARY OF FRENCH BORROWINGS – DILF

(DICŢIONAR DE ÎMPRUMUTURI LEXICALE DIN LIMBA FRANCEZĂ)

A SHORT PRESENTATION

DANIELA DINCĂ, MIHAELA POPESCU

University of Craiova, Faculty of Letters, Romania
danadinca@yahoo.fr, cecilia99_ro@yahoo.com

Abstract

The hereby paper aims at presenting the *Dicţionarul de împrumuturi lexicale din limba franceză* (*Dictionary of French Borrowings – DILF*), a first attempt to set up a comprehensive corpus of French borrowings, including origin as a term selection criterion, in order to reveal the deep Latinity, vitality and originality of Romanian language, that was able to successfully integrate foreign elements, assimilating and changing them so as to provide a specific and original physiognomy to vocabulary.

Key words — French Borrowings, Romanian language, multiple etymology.

1. Introduction

It is common knowledge that French borrowings are currently considered a part of the cultural heritage of humanity, a significant element of modern European spiritual identity. Starting from the analysis of French borrowings, the research project *Tipologia împrumuturilor lexicale din limba franceză în limba română. Fundamente teoretice, dinamică şi categorizare semantică (Typology of French borrowings into Romanian. Theoretical Fundaments, Dynamics and Semantic Classification – FROMISEM)*, developed at the University of Craiova during 2009-2011, achieved the following main outcomes: 1. The establishment of the first comprehensive corpus of French borrowings; 2. Terminology-related clarifications regarding specialised language (*borrowing, neologism, neonym, gallicism, multiple etymology*, etc.); 3. The semantic analysis of French borrowings belonging to representative conceptual areas of the Romanian vocabulary (clothes, furniture, gastronomy).

The hereby paper aims at presenting the first comprehensive corpus of French borrowings, published as *Dicţionarul de împrumuturi lexicale din limba franceză* (*Dictionary of French Borrowings – DILF*), which is a useful work tool both for specialists and for anyone interested in words borrowed from French, in their acquired meanings and especially in their position within the vocabulary of Romanian language. In other words, this dictionary is a first attempt to set up a corpus of French borrowings, including origin as a term selection criterion.

2. The Sources for the Creation of DILF

DILF is a large-sized corpus of Romanian words with a French etymon (30,000 entries), providing a global and full image on French origin elements. The starting point was the most frequently used explanatory dictionary of Romanian language (DEX 1998), published by the "Iorgu Iordan – Alexandru Rosetti" Institute of Linguistics of the Romanian Academy. The first edition of this lexicographic work was published in 1975, and subsequent editions (1996, 1998) were considerably enriched, so that the 1998 edition includes 65,000 words. Moreover, it has to be said that etymological indications were enhanced from the 1996 edition of this dictionary, due to the cooperation of the reputed linguist Theodor Hristea. This resulted in etymological definitions that were similar to the ones in DLR, but the advantage was that DEX includes a much higher number of neologisms.

As for the methodological principles of drawing up DILF, we can only underline that words are selected according to the two following etymological criteria: 1. their exclusively French origin; or 2. their multiple etymology, also including French. It has to be stated that, on the whole, the etymological mention in DEX was considered. Though this is not a specialised, but a general dictionary, it does mention certain etymologies or, when the origin of the word is uncertain, it only presents its various sources.

The organisation of the materials in DILF started from entry words, which were subsequently accompanied by their derivatives formed within the Romanian language, thus providing a global image of the lexical creativity of Romanian in terms of French borrowings. Since it is conceived as an auxiliary tool for linguistic research, the dictionary pays due attention to word derivation and composition. For instance, the following words were grouped around the word "a abandona" (Fr. abandonner): abandonare, abandonat:

> **ABANDONA**, *abandonez*, vb. I. **1.** Tranz. A părăsi pe cineva (lăsându-l fără sprijin sau ajutor); a renunţa la ceva. **2.** Tranz. şi intranz. A renunţa la continuarea participării la o întrecere sportivă. Din fr. **abandonner**.
> **ABANDONARE**, *abandonări*, s.f. Acţiunea de *a abandona*; părăsire. V. **abandona**.
> **ABANDONAT, -Ă,** *abandonaţi, -te*, adj. Care a fost părăsit. ♦ Spec. (Despre copii nou-născuţi) Lepădat[2]. V. **abandona**.

Such a structure can reveal several other phenomena: the vitality and perfect adaptation of most Romanian words of French origin, as well as the relative frequency of the various prefixes, suffixes and composition elements.

3. Structuring principles of DILF

Words of French origin were grouped in two volumes, each representing a different section of the work: the first volume includes borrowings with exclusively French etymology (e.g. *abonament, creion, detaliu, frontieră, gri, matineu, obstacol, naiv, a neglija, opinie, a traversa, tren*, etc.), while the second volume includes borrowings with multiple etymology, including French (e.g. *ulterior*, Fr. *ultérieur*, Classical Latin *ulterior* or *pachet*, Fr. *pacquet*, Germ. *Paket*, etc.).

In the first volume, the dictionary provides a list of words having French language as a primary source, in other words loans with single etymology. They mostly exhibit few phonetic or morphological changes from original French word.

A special class includes words "with indirect single etymology", that use a French "model", which is lexicographically indicated by the acronym cf. (Lat. *confer*). For instance, the word *manierat*, despite being a denominative formed within Romanian language from the basic noun *manieră*, is created according to the model of the French adjective *maniéré*, which justifies the mention «cf. Fr. maniéré». The same remark is valid for the following examples: *reabona* from *re* + *abona*, cf. Fr. *réabonner*, *ultraacustic* from *ultra* + *acustic*, cf. Fr. *ultraacoustique*, *alcoolizat*, an adjective proceeding from (*a*) *alcooliza*, cf. Fr. *alcoolisé*, *neangajat* from *ne* + *angajat*, cf. Fr. *non-engagé*, etc.

The second volume includes borrowings with multiple etymology, i.e. those words that have entered Romanian language through several ways, one of which must be French. The concept of multiple etymology, first introduced to Romanian linguistics by Alexandru Graur, is centred on the principle according to which "a word can simultaneously have several possible etymons" (Graur 1950) (v. Popescu, 2013).

The other languages most frequently indicated as possible etymologies are Classical Latin, Modern Greek, Italian, Spanish, English, Russian and German.

Most words with multiple etymology are indicated to be of French and Latin origin (*abac* from Fr. *abaque*, Lat. *abacus*, *abstinent* from Fr. *abstinent*, Lat. *abstinens*, *-ntis*, *absorbţie* from Fr. *absorption*, Lat. *absorptio*, *-onis*). This shows that:

(1) From an etymological point of view, the Romanian language, unlike other Romance languages, did not inherit some words from Latin. They entered the language much later, either from Classical Latin (the so-called *cultisms*) and/or from French language, where such lexems could be words inherited from Latin or neologisms from Classical Latin);

(2) The application of the formal criterion in establishing etymology sometimes makes it impossible to establish whether the source is French or Classical Latin, so that some lexemes are classified as multiple etymology words.

As for the linguistic systems associated to French for etymons with a multiple origin, the most frequently found combinations are presented in Table 1.

Thus, multiple etymology exhibits an overwhelming presence of words or at least lexical models from Latin and Romance languages, proving the re-Romanisation of Romanian language of the 19th century. During this period, the Romanian vocabulary is enriched and structured according to the requisites of a modern society.

In somewhat less cases, French is associated to English (*inconel* < Fr., Engl. *inconel*, *informal* <Engl. *informal*, Fr. *informel*) or German (*intelectualitate* <Fr. *intellectualité*, Germ. *Intellektualitt*, *interzonal* <Fr., Germ. *interzonal*, *ioniu* <Germ. *Ionium*, Fr. *ionium*), or Russian (*istorism* < Fr. *historisme*, Rus. *istorizm*, *macromolecula* < Fr. *macromolécule*, Rus. *makromolekula*).

In conclusion, borrowings with multiple etymology certify that French is still the main path of entrance to Romanian language of words with Classical Latin, Spanish, Italian or even Russian origin.

Table 1: The most frequently found combinations

Combinations	Examples
French + Latin	ABSTINENT (Fr. abstinent, Lat. abstinens, -ntis) ABSTRACŢIE (Fr. Abstraction, Lat. abstractio, -onis)
French + Italian	ACONT (It. acconto, Fr. acompte) ARLECHIN (Fr. arlequin, It. arlecchino) BALET (Fr. ballet, It. balletto)
French + Latin + Italian	ARMA (Fr. armer, It. armare, Lat. armare) CANGRENĂ (Fr. gangrène, Lat. Gangraena, It. cancrena)
French + Spanish	FANDANGO (Fr., Sp. fandango) GHERILĂ (Fr. guérilla, Sp. guerrilla) GITANĂ (Sp. gitana, Fr. gitane) PESCADOR (Sp., Fr. pescador)
French + Modern Greek	EPIBAT (Fr. épibate, MGr. epibátis) LOGOS (MGr. lógos, Fr. logos) MANIE (MGr. mania, Fr. manie)
French + English	SCANER (Eng., Fr. scanner) SCHECI (Fr., Eng. sketch) SCHETING (engl, Fr. skating) ŞUT (Fr., Eng. Shoot)
French + German	SCONCS (Fr. sconse, Germ. Skons, Eng. skunk) SERPENTIN (Fr. serpentine, Germ. Serpentine) ŞILVANIT (Germ. Sylvanit, Fr. Sylvanite) SOCIAL-DEMOCRAŢIE (Germ. Sozialdemokratie, Fr. Social-démocratie)
French + Russian	SOCIOLOGISM (Fr. sociologisme, Rus. soţiologhizm) ŞERARDIZARE (Fr. shérardisage, Germ. Scherardişierung, rus. serardizaţiia) TERMOFOSFAT (Fr. thermophosphate, Rus. termofosfat) TORON (Fr., Eng. Thoron, Germ. Thoron, Rus. toron)

4. The Statistics of French Borrowings into Romanian Language

Due to the creation of *DILF* (2009), we were able to draw up our own statistics on a comprehensive corpus (65,000 words), divided into two large categories: basic words and derivatives.

Table 2: Statistics on basic words and derivatives

Basic words	Words with exclusively French etymology	30.60%
	Words with multiple etymology, including French	9.04%
Total		**39.64%**
Derivatives	Words with French and multiple etymology	7.87%
General total		**47. 51 %**

As it can be seen, most words in *DILF* have single etymology (30%) and, if adding words with a multiple etymology, including French, the percentage raises to 39% of all the words in the dictionary. The large number of words with French origin is due to their massive penetration in the 19th century and the first half of the 20th century, when the vocabulary of modern Romanian was formed. This provides a suggestive image of the discussed phenomenon: almost half of the modern Romanian vocabulary is influenced by French in one way or another.

On the other hand, the general total of 47%, basic words and derivatives, is quite relative and is highly dependent on the etymological indications of the dictionary we resorted to (*DEX*). Obviously, this number reflects the etymological indications of the dictionary used as a basis (such indications are sometimes controversial), which would require, in the future, a rigorous etymological research, with a view to corroborating this information and the data provided by other fundamental, explanatory and etymological dictionaries of Romanian language, especially *Dicţionarul Academiei* and *Dicţionarul limbii române, Serie nouă*.

5. A Word Derived Within Romanian Language or a Borrowed Word?

A significant difficulty encountered when establishing lexical entries refers to the interpretation of derivatives, either as borrowings, or as Romanian formations (in parallel to the evolution of foreign structures). The controversial etymology of this class of words may be explained by the differences between the information sources and the criteria according to which dictionary authors establish etymology.

The most frequent situations can be assigned to three cases:

1. The basic word is borrowed from French, while derivatives are formed within Romanian language:

> **DIRIJA,** *dirijéz,* vb. I. Tranz. A conduce, a îndruma o instituţie, o organizaţie, o activitate etc. ♦ Spec. A conduce o orchestră, un cor (în calitate de dirijor). Din fr. **diriger.**
>
> **DIRIJARE,** *dirijări,* s.f. Acţiunea de *a dirija,* conducere, îndrumare. V. **dirija.**
>
> **DIRIJAT, -Ă,** *dirijaţi, -te,* adj. Care a primit o anumită direcţie sau orientare; condus, îndrumat (de altul). V. **dirija.**
>
> **DIRIJOR, -OARE,** *dirijori, -oare,* s.m. şi f. Persoană care conduce o orchestră sau un cor. - **Dirija** + suf. *-or.*
>
> **DIRIJORAL, -Ă,** *dirijorali, -e,* adj. Care aparţine dirijorului, specific sau necesar dirijorului. - **Dirijor** + suf. *-al.*
>
> **TELEDIRIJA,** *teledirijez,* vb. I. Tranz. A telecomanda. - **Tele-** + **dirija.**
>
> **TELEDIRIJARE,** *teledirijări,* s.f. Acţiunea de a *teledirija.* V. **teledirija.**

2. The word is derived within Romanian language, with a suffix, but indirect single etymology also is indicated with the mention *Cf. fr:*

> **ABONA**, *abonez*, vb. I. Tranz. şi refl. (Cu determinări introduse prin prep. "la") A-şi face un abonament. ♦ Refl. Fig. (Fam.) A veni în mod regulat undeva, a fi un obişnuit al casei. Din fr. **abonner**.
>
> **REABONA,** *reabonez,* vb. I. Refl. A se abona din nou. - **Re-** + **abona.** Cf. fr. r é a b o n n e r .

> **BELIGERANŢĂ** s.f. Situaţia în care se află un beligerant; stare de război. Din fr. **belligérance.**
>
> **NEBELIGERANŢĂ** s.f. Stare a unei naţiuni, care, fără a manifesta o strictă neutralitate, se abţine de a lua parte efectivă la un conflict armat. - **Ne- + beligeranţă**. Cf. fr. n o n - b e l l i g e r a n c e .

3. For several derivatives from a single basis, one of them is a Romanian creation, while the second follows the French model:

> **ACUSTIC, -Ă**, *acustici, -ce*, adj., s.f. **I.** Adj. Care emite, transmite sau recepţionează sunete, care aparţine acusticii (**II 1**), privitor la acustică. ◊ *Nervi acustici* = a opta pereche de nervi cranieni. *Tub acustic* = tub lung care serveşte la transmiterea vocii pe nave, în puţuri minere etc. *Cornet acustic* = dispozitiv cu ajutorul căruia se recepţionează sunete şi se înlesneşte perceperea lor. **II.** S.f. **1.** Parte a fizicii care se ocupă cu studiul producerii, propagării şi recepţionării sunetelor. ◊ *Acustică arhitecturală* = ramură a acusticii care studiază fenomenele legate de propagarea undelor acustice în încăperi. **2.** Calitatea de a înlesni o (bună) audiţie. Din fr. **acoustique**.
>
> **INFRAACUSTIC, -Ă,** *infraacustici, -ce*, adj. (Fiz.; despre vibraţii acustice) A cărei frecvenţă se află sub limita inferioară a domeniului de audibilitate. Din fr. **infra-acoustique.**
>
> **ULTRAACUSTIC, -Ă**, adj., s.f. **1.** Parte a acusticii care se ocupă cu studierea ultrasunetelor. **2.** Referitor la ultraacustică, de ultraacustică. - **Ultra-** + **acustică** - Cf. fr. u l t r a a c o u s t i q u e .

However, *DILF* did not aim at solving etymologically controversial issues regarding derivatives, which are still widely debated in Romanian linguistics (Reinheimer Rîpeanu, 1989).

6. New Research Directions

Borrowings have been thoroughly studied by Romanian linguists, who have showed the importance of French elements and their part in the modernisation of the literary Romanian language. However, one should not overlook the heterogeneity of such studies, which mostly regard phonetic, morphological and syntactic adaptation, the fields of penetration, the criteria for a correct establishment of etymology, frequently neglecting semantic issues.

In this context, *DILF* is a corpus that may be exploited in several directions, such as:

- Deeper etymological analysis

DILF is a corpus outlining etymologically controversial issues for all types of lexical borrowings, with a single or multiple etymology, with indirect single etymology, simple words or derivatives. An attentive and detailed study would be needed for their clarification, which should compare several lexicographic, explanatory and etymological sources.

- Breakdown of lexical compartments such as Common Vocabulary vs. Specialised Vocabulary

The inventory of Romanian words of French origin represents about half of the *DEX*, proving the overwhelming influence of French language in the establishment of the Romanian vocabulary. This influence was practically seen in all fields of the vocabulary. Therefore, words of French origin are found both in fundamental vocabulary, and in the lexis of scientific and technical fields: medicine, botany, zoology, mechanics, sociology, history, psychology, etc. Based on the corpus provided by *DILF*, a statistics can be drawn up regarding the share of words of French origin in some of these specialised fields.

Upon a transversal analysis of the corpus, one can find out that medical terms take the first position, which can be explained by the French influence on the creation of Romanian medical system. Besides, medicine was very well represented in Transylvania in the 18th century. The first medical texts of the 18th century are translations of Hungarian and German studies. In terms of adaptation to Romanian language, the words reproduce the written version of the French etymon.

Another field where the share of words of French origin is important is legal terminology. Therefore, after 1830, the main source of modernisation of legal language is represented by neological Latin borrowings (*cod, dosar, ordonanţă, sentenţă*), progressively replacing former borrowings. The Civil Code (1865) was based on the Napoleon Code (1804), and the Romanian Constitution was inspired by the French one. The number of legal terms of Latin origin has vertiginously increased so far, so that Romanian legal and administrative language has a modern and contemporary appearance. The approximately 118 terms of French origin of *DILF* certify the level of adaptation of these terms in Romanian language, as most of them penetrated our language together with their related concept, which had to be named somehow.

- Adjustment of borrowings to the phonetic, morphological and syntactic system of Romanian language

The borrowed words must be integrated in the system of the recipient language, complying with its morphological, syntactic and phonological rules. The establishment of a typology of the morphological and syntactic changes undergone by French borrowings can be a possible clue for a contrastive approach of the two languages, with immediate utility in French language teaching practice for Romanian speakers.

- Extension of semantic analysis

If Romanian language performed few innovations in term of phonetics and morphology, compared to French, the originality of Romanian language can be fully seen in the semantic richness of Romanian words, compared to their French etymons. Some words have fully maintained their meaning from French, and this is seen, as it was expected, in words belonging to scientific and legal language.

In the approach of borrowings, the sociolinguistic study of the two languages in contact is a significant tool that might outline the particularities of the cultural development of the recipient society as a whole, along with the development of Romanian vocabulary. Seen from this point of view, neology, as an integral part of applied linguistics, reveals the originality and specificity of each language, the relations between languages, as well as the interdependence between language and the society it evolves and operates within.

Linguistic comparativism is thus doubled by a cultural comparativism, since borrowings reflect the ideology of both a language and the people that speaks it.

By comparing the meanings of etymons, their transfer to the recipient language, along with the semantic changes characterising such borrowings, the role of the extralinguistic factor in the selection of the sememes of the French etymon, reflecting a certain step in the evolution of Romanian society, as well as the semantic evolutions of borrowings within the Romanian language can be established.

7. Conclusions

In conclusion, the French language was more than a cultural influence for the Romanian language, actually triggering the deep restructuring of its entire lexical structure, a phenomenon that has been coined as *re-Romanisation*. The existence of a percentage of 39% words with a French origin of the 65,000 words included in *DEX* (1998) proves the full European integration of Romania and Romanian language in the 19th century and the beginning of the 20th century.

The starting point of the corpus in *DILF* is the largest general dictionary of Romanian language (*DEX*), including 65,000 entries. Thus, *DILF* is a representative corpus of French borrowings, providing a global and complete image on Romanian elements with a French origin and outlining the importance of such elements, their derivative force, etymology-related issues and semantic particularities. At the same time, it reveals the deep Latinity, vitality and originality of the Romanian language, which was able to successfully integrate foreign elements, assimilating and changing them so as to provide a specific and original physiognomy to vocabulary.

Another individual feature of *DILF* is that it represents the first large corpus of French borrowings to Romanian, also considering words with multiple etymology. Consequently, one of the specificities of the Romanian language is the existence of a wide amount of linguistic solutions for the adaptation of newly acquired words, inherent to a highly permissive and creative language.

One may state, hence, that the vitality and adaptation of a significant number of words of French origin to the orthographical, phonetic, morphological and syntactic system of Romanian language proves the hospitality of Romanian language and its permanent enrichment under the influence of the languages it comes in contact with.

Acknowledgements

This work was partially supported by the grant number 34C/2014, awarded in the internal grant competition of the University of Craiova.

References

Avram, M. (1982). Contacte între română şi alte limbi romanice. *Studii şi cercetări de lingvistică* 3, 253-259.

Dimitrescu, F. (1994). *Dinamica lexicului limbii române*. Bucharest: Logos.

Goldiş-Poalelungi, A. (1973). *L'influence du français sur le roumain. Vocabulaire et syntaxe*. Paris: Les Belles Lettres.

Graur, A. (1950). Etimologia multiplă. *SCL* 1, 22-34.

— (1954). *Încercare asupra fondului principal lexical al limbii române*. Bucharest: Editura Academiei.

Hristea, Th. (1968). *Probleme de etimologie*. Bucharest: Editura Ştiinţifică.

Iliescu, M. (2003-2004). Din soarta împrumuturilor româneşti din franceză. *AUI*, XLIX-L, 277-280.

Lombard, A. (1967). Latinets oden i oster. *Filologiskt arkiv* 12, Lund.

Popescu, M. (2013). Une notion-clé dans la lexicologie roumaine : l'étymologie multiple. *Actes del 26é Congrés de Lingüística i Filologia Romàniques* (València, 6-11 de setembre de 2010), Vol. IV, Berlin: W. de Gruyter, 337-347.

Popovici, V. (1992). Derivat sau moştenit? O problemă a lingvisticii romanice. *Studii şi cercetări de lingvistică* 43, 71-79.

— (1996). Mots hérités ou dérivés en roumain. Un problème d'étymologie roumaine en perspective romane. *Rumänisch: Typologie, Klassifikation, Sprachkarakteristik*. München: Südosteuropa-Gesellschaft, 265-275.

Reinheimer-Rîpeanu, S. (1989). Derivat sau împrumut. *Sudii şi cercetări de lingvistică* 4, 373-379.

Sala, M. (coord.) (1988). *Vocabularul reprezentativ al limbilor romanice* (VRLR). Bucharest: Editura Ştiinţifică şi Enciclopedică.

Şora, S. (2006). Contacts linguistiques intraromans: roman et roumain. *RDG*, tome 2, 1726-1736.

Dictionaries

CADE = Candrea, I. A. / Densuşianu, Ov. (1907-1914). *Dicţionarul etimologic al limbii române*. Bucharest.

CDER = Ciorănescu, Alexandru (1958-1966). *Dicţionarul etimologic al limbii române*: Bucharest: Editura Saeculum I.O.

DA = *Dicţionarul limbii române* (1913-1949). Bucharest: Academia Română.

DCR = Dimitrescu, Florica (1982). *Dicţionar de cuvinte recente*. Bucharest: Editura Albatros.

DEX = *Dicţionarul explicativ al limbii române* (1998, 2nd edition). Bucharest: Univers Enciclopedic.

DILF = Costăchescu, Adriana / Dincă, Daniela / Dragoste, Ramona / Popescu, Mihaela / Scurtu, Gabriela (2009). *Dicţionar de împrumuturi lexicale din limba franceză*. Craiova: Editura Universitaria.

DLR = *Dicţionarul limbii române*. Serie nouă (1958-2010). Bucharest: Editura Academiei.

DLRC=Academia Română (1955-1957). *Dicţionarul limbii române literare contemporane*. Bucharest: Editura Academiei Române.

DLRM = *Dicţionarul limbii române moderne* (1958). Bucharest: Editura Academiei.

DOOM = Academia Română (2005, II-ème édition). *Dicţionarul ortografic, ortoepic şi morfologic al limbii române*. Bucharest: Univers Enciclopedic.

SENSE-TAGGING OF ROMANIAN GLOSSES

CORINA HOLBAN, FELICIA CODIRLAȘU,
ANDREI MINCĂ, ȘTEFAN DIACONESCU

Research and Development Department, SOFTWIN, Bucharest – Romania
{cholban, fcodirlasu, aminca, sdiaconescu}@softwin.ro

Abstract

This work describes a research phase carried out in the project SenDiS ("General Word Sense Disambiguation System applied to Romanian and English Languages") that is the creation of a specific lexicon network obtained by manually tagging the tokens in a lexicon's glosses with their contextual meaning. Therefore, we present here: the annotation model, the tool used for manually tagging the glosses and some statistics following the annotation effort. Also, we give here the set of 'sense-tagging' principles derived from the preliminary analysis of glosses and enriched afterwards by the tagging experience.

Key words — knowledge-based Natural Language Processing, lexicon networks, Lesk algorithm, semantically annotated glosses, WordNet tagged glosses.

1. Introduction

Supervised and especially unsupervised paradigms are the nowadays popular approaches to Natural Language Processing (NLP). Both exploit the increasing computing power of machines and also the exponential growth of digital information.

Sometimes, this process is accelerated by private initiatives of great effort. Consider only the Google Print Library Project, started in 2004, which was to digitize and make available approximately 15 million volumes within a decade. With this undertaking, Google also made the switch from rule-based NLP to statistical NLP and machine learning.

However, at the same time, carrying the Google Print Library Project sent an important message: the need of qualitative or structured data as input for better NLP. Whether the choice should be statistical NLP or rule-based NLP or even hybrid NLP, there is a strong belief that Linguistic Knowledge Bases (LKB) is the stepping stone of all.

Digital Linguistic Resources (LR) is now pursued in multiple laboratories in various standards, volumes and for a large number of languages. The success of WordNet led to a revolution in the creation of LRs with the advent of other WordNet like-type LRs for most of the European languages and with the inclusion of richer linguistic information.

An important aspect in creating LKBs is the model used and the granularity level of the linguistic information. We find LKBs that describe several aspects of a language or of language pairs using shallow or deep structures, the case of WordNet, EuroWordNet, BalkaNet (Lenci, 2000; Calzolari, 2001; Barbu, 2008). Also, we find initiatives that deal with all aspects of a language and provide deep description of linguistic information, the case of (Diaconescu, 2007). Automatic creation of LKBs is preferred but, at best, ends

in obtaining large corpuses using shallow parsing. Building specialized tools that linguists operate easily is required for the management of complex language descriptions (Diaconescu et al., 2009; Simionescu, 2012).

Lexicon networks are a particular field in the creation of LKBs which gained, in time, a great interest as a linguistic resource showing improved outcomes for various NLP tasks. As a result, if one is to construct an LKB model this usually involves at least one lexicon network with various types of relations (synonyms, antonyms, hyponyms, etc.).

A great part of lexicon networks is in fact the semantic networks. Semantic networks are a particular type of lexicon networks that implies interconnecting the meanings of lexicon entries using one or more specific types of semantic relations. However, semantic relations that make up these networks have a significant disadvantage: the directed graph obtained from a single type of semantic relation is not a connected graph, but is in fact a large set of isolated connected components. To overcome this situation, the graphs are mixed together in order to obtain a large network of interconnecting meanings. Having a large network with various types of relations can be a serious challenge when mining for knowledge in NLP tasks.

Taking into consideration the above, our work focuses on describing the process of creating a particular semantic network obtained by tagging the tokens in a lexicon's glosses with their contextual meaning. The paper is organized as follows: In Section 2, we present related research and work that make the basis of this research report. In Section 3, we further describe the annotation process and its challenges. In section 4, some measures of the annotation process are provided. Section 5 concludes the current research activity.

2. Related Work

As of 2008, the WordNet incorporates the "Princeton Annotated Gloss Corpus" – a corpus of manually annotated WordNet synset definitions ("glosses"). WordNet 3.0 is the sense inventory against which the annotation was conducted. This type of annotation implies the tokenisation of each gloss (words or collocations) and manually selecting a contextual meaning for each token. In other words, a gloss is linked with other glosses by a semantic relation called "sense-tagged glosses". This type of semantic network can be very dense and it links senses belonging to different morphological lexicon entries.

A great deal of interest is shown for this type of network, as it is a natural extension of a classical concept, the 'dictionary'. While a dictionary is a collection of words explaining themselves in plain text, the "sense-tagged glosses" network is a collection of senses explaining one another in a structured way.

While in the case of WordNet it was mostly a manual undertaking, others are pursuing this using automatic or semi-automatic approaches either for improving WordNet, the case of eXtended WordNet, (Castillo et al, 2004; Litkowski, 2004; Moldovan and Novischi, 2004; Navigli, 2009 for other languages).

As mentioned above, semantically annotated glosses show promising outcomes in various NLP tasks, especially in word sense disambiguation (Banerjee and Pedersen, 2002; Navigli and Velardi, 2005; Mincă and Diaconescu, 2013).

3. Manual disambiguation of glosses

The consideration of glosses as a potential linguistic resource for NLP tasks started relatively late in the field (Lesk, 1986). Glosses were first mined for the purpose of word sense disambiguation. Two words that share the same context are disambiguated by overlapping their glosses' definitions. The meanings with a maximum overlapping score would have been selected for each of the two words. Another version of the Lesk algorithm, less complex, would overlap the glosses for each word only to the context. Soon enough, this approach outlined the importance of glosses and the Lesk reasoning was extended with the need to disambiguate the glosses (as in Figure 1).

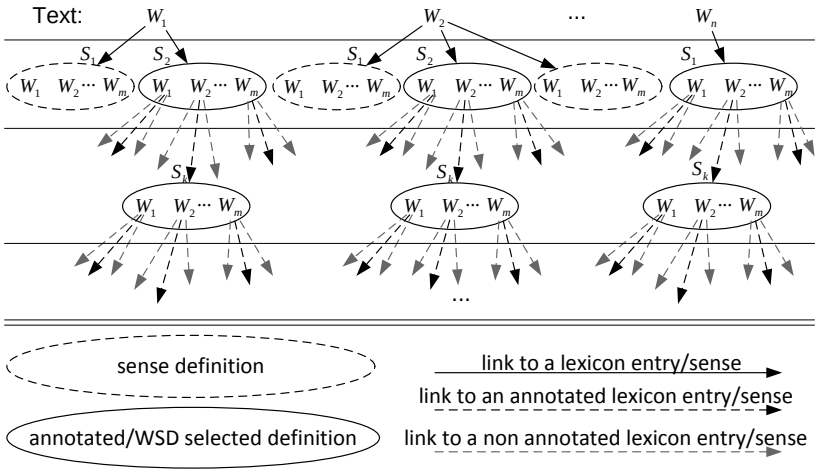


Figure 1: Lesk algorithm reasoning extended. Every annotated sense is extended with its definition that also has words with disambiguated senses and so on.

As in the case of WordNet sense-tagged glosses, we considered to manually disambiguate the glosses for Romanian language. In order to do that, a thorough study of the Romanian glosses was conducted and a specialized tool was developed for the task.

3.1. Annotation model

As seen in Figure 2, the annotation model is very similar to the one used in the case of WordNet.

However, we considered that it will be useful not only to tag the words in a gloss with their contextual meaning, but also tag each word with a contextual degree of relevance for its contribution in building up the meaning of the gloss. For this purpose, we devised three levels of relevance that a linguist can choose for a word: *Weak*, *Medium* and *Strong*. The levels of relevance have the following parity: (5, 5/4, 1).

Additional information can be added in the annotation process (lemma, morphological and syntactic categories, etc.) but it is preferable to automate the process as much as possible.

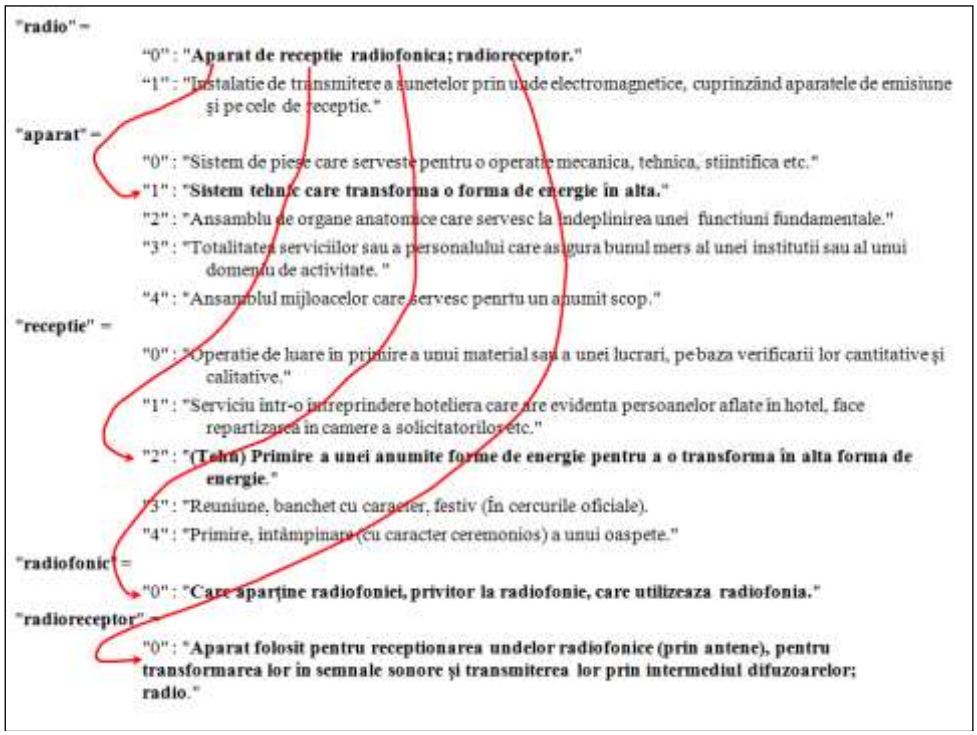


Figure 2: Annotation example for the Romanian word "radio" (first meaning)

In the case of large lexicons (like Romanian) the estimated effort is at a maximum of 300,000 glosses and a maximum of 3,000,000 tokens that need to be operated.

3.2. Sense Inventory

Our lexicon is based on the Explanatory Dictionary of the Romanian Language, 1998 (DEX – Dicţionarul explicativ al limbii române). Mostly, the glosses were imported from the digital resource "DEX online" (found at http://dexonline.ro). To fill certain gaps (missing words, missing definitions) other sources were used, but in a limited degree, like the Small Academy Dictionary (Micul Dicţionar al Academiei – 4 volumes), and the Thesaurus Dictionary of the Romanian Language (Dicţionarul Tezaur al Limbii Române – 19 volumes). Moreover, the text of the definitions was updated according to the morphological, orthoepic and orthographical rules form DOOM2. The structure of a dictionary, DEX '98 implicitly, does not fully match the needs for our LKB model. Therefore, we used other criteria in structuring the lexicon, in deciding what entry words the lexicon should have, and, implicitly, how to divide the definitions.

Generally, dictionaries group their definitions for entry words taking into account the etymology of the word (the etymology is their main criterion in deciding whether some words with the same form are homonyms or polysemantic words), regardless of the morphological category or aspects like animation, defective of plural/singular form, gender, transitivity, reflexivity, etc. Therefore, we may find, gathered under a single entry word, definitions for different morphological categories that word may have, as for example:

> ***FRUMÓS, -OÁSĂ,*** *frumoşi, -oase,* ***adj., adv., s. n. I.*** *Adj.* ***1.*** *(Adesea substantivat; despre fiinţe şi părţi ale lor, despre lucruri din natură, obiecte, opere de artă etc.)*

Care place pentru armonia liniilor, mișcărilor, culorilor etc.; care are valoare estetică; estetic. ◊ *Arte frumoase = pictură, sculptură, gravură (în trecut și arhitectură, poezie, muzică, dans)...*

In contrast, our framework, GRAALAN (Diaconescu and Dumitrașcu, 2007), takes into consideration the aspects mentioned above. Our lexicon has for each morphological category (considering also gender, transitivity, reflexivity, etc.) an entry word in the dictionary, thus dividing the definitions from DEX '98.

In order to build a coherent and uniform semantic network, some steps had to be followed, some preceding the actual annotation process. A first step was to evaluate and to improve the definitions from the lexicon in order to comply with the annotation method and to extract some rules of good practices (see 3.3) for the annotation process. Overall, definitions suffered changes like additions, deletions, separations, updates, etc.

3.3. Good practices

Each word from a definition has a certain semantic degree of relevance depending on its contribution in defining the respective word. As it was described before, three degrees of relevance were used: ***weak, medium*** and ***strong***. There is also the possibility to ***Ignore*** various words from a definition if they do not bring any semantic contribution in defining a word. It is very important to assign the degrees of relevance in a uniform way, complying with some general rules in order to obtain a coherent semantic network.

Further on we will present the main rules observed in the annotation process, however, without going into details. In addition to the rules described below, there are also some particular rules that apply to a limited number of definitions, or to patterns of definitions that we will not mention here.

3.3.1. Words ignored in the process of annotation

As a general rule, the functional elements from a definition like certain conjunctions and prepositions, certain relative or indefinite pronouns (*care, cine, cineva, vreunul*) with no semantic relevance, certain adverbs (*unde, când*) and pronominal determiners when they do not have anaphorical value (*acest, acesta*), indefinite and genitive articles (*un, o, a, al*), the auxiliary and copulative verb *a fi* (when it has a positive form) are ignored in the process of annotation.

3.3.2. Degree of relevance

Generally, within a definition, the highest degree of relevance (***Strong***) is given to the word that has the biggest contribution in building the semantic of the word defined. Most of the times, this word is of the same morphological category as the word defined (a synonym, a word from the semantic/lexical paradigm of the word, etc.).

For example, in *CASĂ* **–** ***Clădire*** *care servește drept locuință,* and *PREOȚIE - Calitatea, demnitatea, funcţia de* ***preot,*** the words *clădire* and *preot* will have the degree of relevance ***Strong.***

The degree of relevance ***Medium*** is assigned mainly to hyperonyms or to generic words when we have abstract or verbal nouns. For example, in *BUNĂTATE* **-** ***Însușirea*** *de a fi*

*bun, **înclinarea** de a face bine; PLECARE - **Acțiunea** de a pleca,* the words *însușirea, înclinarea* and *acțiunea* will be assigned ***Medium***.

The words that are assigned ***Weak*** are those with a minimum impact on the overall meaning of the word defined. Such words are, for example, from the metalanguage of the definition as *reprezintă*, *marchează*, *exprimă*, *aparţine*, etc., or the copulative verb *a fi* when it has a negative form.

3.3.3. Identified problems during the annotation process and solutions

Two of the main problems encountered in the annotation process were the missing words or idioms from the lexicon (negative participles, composed words with prefixoids, neologisms, etc.) and the incomplete meanings. These words/idioms have been introduced in the lexicon and the definitions were enriched as they were found.

Another problem was that of the anaphoric words (a word that refers one or more previous words) as in: *TÂNĂR - Care a fost plantat sau a răsărit de puțină vreme, care n-a ajuns încă la maturitate; care este format din **asemenea** plante* where *asemenea* refers to the type of plants described before. In this case, the rule adopted was to introduce the definition *anaphoric value* to each word that may be anaphoric.

3.4. Annotation Tool

In order to facilitate the manual annotation process a dedicated tool was devised considering the following principles: to be a collaborative tool, to ensure an ergonomic but still a fast tagging task, to allow the browsing of the current network, to provide a quick feedback in the case of any changes in the network, to allow different sense mappings for a certain word in a gloss than those given by the annotator.

3.4.1. Workflow

A linguist will start working with this tool by first choosing a lexicon entry of interest. From the list of meanings corresponding to this entry, a certain meaning is chosen to have its gloss tagged. The selected gloss is tokenized and the resulting tokens are annotated with a list of meanings. In case a token is not recognized by the annotator, the linguist can provide another word form and thus recall the annotator. Then, the linguist selects groups of tokens for which will provide a certain degree of relevance or can ignore a specific group. For each valid token, the linguist will choose the contextual meaning. Finally, the linguist will save the modifications and the tool will mark the current annotated gloss as in progress or completed.

4. Statistics

The following tables show the annotation results in contrast with the measurements found in WordNet. A greater tagging density can be observed in the case of Romanian glosses.

Table 1: Romanian glosses vs. WordNet

LexNets	Glosses	Tagged Glosses	Targeted Glosses	Tags Density
Romanian	130,087	118,536	58,976	0.5757
WordNet	206,941	206,938	59,251	0.3486

Table 2: Annotation results and WordNet

LexNets	All Tokens	Operated Tokens	Operated Tokens Valid	Op. Tokens Related	Op. Tokens Valid & Related
Romanian	1,528,819	1,191,942	691,010	720,420	686,210
WordNet	2,394,190	2,394,190	2,394,189	834,803	834,803

The number of valid and related operated tokens is distributed as follows: **299,098** ***Strong*** relations, **209,879** ***Medium*** relations and **177,233** ***Weak*** relations.

The annotation process can be covered by an effort of a total of 40 man-months.

5. Conclusions

We presented here the experience gathered from manually building a lexicon network over a dictionary glosses, for Romanian language. A dedicated tool drastically reduced the manual labour of linguists. We concluded a set of specific rules for sense-tagging of Romanian glosses that can also be considered for other languages. Finally, the effort ended with a large network of over 130,000 meanings interconnected by over 600,000 "sense-tagged glosses" relations[1].

Acknowledgements

This work was carried out in the project SenDiS, co-funded under Romanian grant no. 207/20.07.2010.

References

"Princeton Annotated Gloss Corpus". http://wordnet.princeton.edu/glosstag.shtml

Banerjee, S., Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Barbu, A. M. (2008). Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries. In *The 6th edition of the Language Resources and Evaluation Conference.*

Calzolari, N., Lenci, A., Zampolli, A. (2001). International Standards for Multilingual Resource Sharing: The ISLE Computational Lexicon Working Group.

[1] Please contact the authors for information on availability of the tagged glosses and the annotation tool.

Proceedings of the Workshop on Sharing Tools and Resources, 39th ACL, 7 July 2001, Toulouse, France, 71-78.

Castillo, M., Real, F., Asterias, J., Rigau, G. (2004). The talp systems for disambiguating wordnet glosses. *In Proceedings of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain, 93-96.

DEX – Dicţionar explicativ al limbii române, ed. a 2-a, coord. Ion Coteanu, Luiza şi Mircea Seche, Ed. Univers Enciclopedic, Bucureşti, 1998.

Diaconescu, S., Dumitrascu, I. (2007). Complex Natural Language Processing System Architecture. Advances in Spoken Language Technology, *The Publishing House of the Romanian Academy*, Bucharest. 228-240

Diaconescu, S., Ingineru, C., Codirlașu, F., Rizea, M., Bulibașa, O. (2009). General System for Normal and Phonetic Inflection – *SpeD 2009 Conference on Speech Technology and Human-Computer Dialogue*, 18-21 June, Constanta.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13:4, 249-263.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of SIGDOC '86*.

Litkowski, K. C. (2004). Senseval-3 task: Word-sense disambiguation of wordnet glosses. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain, 13-16.

Mincă, A., Diaconescu, S. (2013). An Approach to Reduce Part of Speech Ambiguity Using Semantically Annotated Lexicon Definitions. *MSIE 2013*, September 28-29. Jinan, Shandong, China.

Moldovan, D., Novischi, A. (2004). Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, 18, 301-317.

Navigli, R. (2009). Using Cycles and Quasi-Cycles to Disambiguate Dictionary Glosses. *Proc. of 12th Conference of the European Association for Computational Linguistics (EACL 2009)*, Athens, Greece, 594-602.

Navigli, R., Velardi, P. (2005). Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27:7, July, 1063-1074.

Simionescu, R. (2012). Graphical Grammar Studio as a Constraint Grammar Solution for Part of Speech, *ConsILR2012*.

CHAPTER 3
APPLICATIONS
IN LANGUAGE PROCESSING

SENTIMENT ANALYSIS OF TOURIST REVIEWS: DATA PREPARATION AND PRELIMINARY RESULTS

COSTIN BĂDICĂ[1], MIHAELA COLHON[2], ALEXANDRA ŞENDRE[1]

[1]*Department of Computer and Information Technology, University of Craiova*
[2] *Department of Computer Science, University of Craiova*
cbadica@software.ucv.ro, mghindeanu@inf.ucv.ro, sendre.alexandra@yahoo.com

Abstract

In this paper we propose a sentiment classification method for the categorisation of tourist reviews according to the sentiment expressed in three categories: positive, negative or neutral. The method is applied on a real data set extracted from the popular Romanian web site AmFostAcolo dedicated to tourists' impressions. Based on experimental results we concluded that our algorithm for sentiment analysis works very well.

Key words — natural language processing, sentiment analysis, text mining.

1. Introduction

Tourism is a highly information-based industry, while the tourism product is a confidence good. At the moment of the customer's choice only the product information, not the product itself, is available. This explains the bias of the tourism industry sector towards IT systems. In particular, a great interest was shown during the last 15 years in the application of intelligent information technologies for the development of intelligent or smart tourism business (Stabb et al., 2002). A lot of efforts are currently invested in the application of text mining and natural language processing techniques for improving the quality of tourism information services (Kaur and Gupta, 2013). Tourists will benefit of advanced IT systems for knowledge and information management to assist them in making decisions with less effort and in a shorter time.

We have recently set up a research project focused on improving the management of information and knowledge extracted from tourists' reviews and opinions that can be publicly found on the web. There are so many information sources containing tourists' reviews and opinions about tourist destinations like: post-visit experiences, tourist advertisements, descriptions of tourist attractions, tourist highlights and advice, recommendations, photos, etc., addressing various aspects like accommodation, trips, historical places, landscape, sightseeing, food, shopping, entertainment, local attractions, a.o. Tourist information is most often presented as textual reviews or comments expressed in natural language that describe the customer's opinions or experiences about various tourist destinations. This information is usually poorly structured, can be more or less focused on a tourist entity or aspect, and can be multi-lingual. Therefore, the task of collecting, aggregating and presenting it in a meaningful way can be very difficult by posing cognitive challenges to the users, as well as technical challenges to the computational methods employed.

In this study we were interested in the application of sentiment analysis (also known as opinion mining) methods to tourist information extracted from the web.

Previous works on mining opinions can be divided into two directions: sentiment classification and sentiment related information extraction (Gînscă et al., 2012, 2011). Following (Liu and Zhang, 2012), "sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes". Typical usage scenarios of opining mining are: (1) detection of good or bad aspects about a certain target object or service that must be evaluated by the users or the owners of that object, (2) detection of sudden sentiment changes about the object, or (3) automatic interpretation of large amounts of opinionated data that are impossible to achieve by manual inspection.

Our task is to extract and classify tourists' reviews according to the sentiment expressed in three categories: positive, negative or neutral. We employed a sentiment classification method and we present the outcome of its application results to tourists' reviews. We focused on a real data set that we extracted from AmFostAcolo[1] - a Romanian Web site having a similar purpose as IveBeenThere[2].

Our proposed method is based on an enhanced term-counting method built primarily on exploiting parsing techniques from natural language processing for detecting dependency links between the words of a text, as well as considering the contextual valence shifters (Polanyi and Zaenen, 2006). This method has the advantage that it does not require training, so it can be applied to reviews where training data is not available.

The main contributions of our paper are:

(1) The proposal of an evaluation framework of sentiment orientation in tourists' reviews that can help hotels managers to improve their business, as well as tourist web site developers by providing valuable feedback to their customers;

(2) The development of a new sentiment analysis algorithm that benefits from syntactic parsing techniques, as well as initial experimental results obtained by its application on tourists' reviews about hotels (in Romanian).

Based on our knowledge, we could not find other works in the literature targeting sentiment analysis on real tourists' reviews expressed in Romanian language. Moreover, our approach is language independent.

The paper is organized as follows. We start in Section 2 with a presentation of data set preparation and pre-processing. We follow in Section 3 with the introduction of our sentiment analysis method. Then we present and discuss the experimental results of the application of our method on the real data set extracted from AmFostAcolo web site. In the last section we conclude and point to future research and developments.

2. Data Set: Preparation and Pre-processing

2.1. Data Set Preparation

The source that we chose as basis for building our data set was the AmFostAcolo web site. It provides a large semi-structured database with information describing post-visit

[1] http://amfostacolo.ro
[2] http://www.ive-been-there.com

tourists' reviews about a large variety of tourist destinations covering specific aspects of accommodation units, as well as general impressions about tourist geographical places and regions.

The information of the data source can be conceptualized as a tree-structured index organized hierarchically according to the destination, region, section and location. Most often a destination represents a country, for example Romania, while sometimes it can also be a continental region including several countries. Each destination contains several regions. For example, a region of Romania is Oltenia. A section most often represents a locality (for example Craiova). A location can represent an accommodation unit or general impressions about the locality and surroundings. Finally, each location is a container of tourist impressions or reviews written by users registered at AmFostAcolo. Each user has associated a trust score that is calculated based on his or her activity and feedback received on the site.

The data set was extracted by crawling the AmFostAcolo web site. For the extraction we have employed a set of heuristic rules, based on identifying the HTML context and style of each extracted item. Our data set contains: 423 male users and 662 female users; 2521 reviews; 45 countries destinations; 161 regions among which there are 16 country subregions and 145 other regions (not country subregions); 529 sections among which 489 localities (cities, towns or villages) and 40 other sections (not localities); 1420 tourist locations among which 534 accommodation units (i.e. sections representing cottages, pensions, hotels, houses or villas) and 886 sections (i.e. that do not represent accommodation units, but rather general impressions about a tourist location).

Figure 1 presents the partitioning of the reviews from our data set depending on their size with the help of a histogram. The size is specified as an interval defined by lower and upper bounds of the number of review words. By analysing the figure it can be easily noticed that most of the reviews in our data set fall between 300 and 400 words.

There are basically two different formats in which reviews are given on tourist sites. The format „pro and cons" describes separately the positive („pro") and negative („cons") opinions (for example, Booking.com[3] uses this format). The other „free format" does not separate pros and cons, and the reviewer can write freely his or her comments. This format is used by several sites, including AmFostAcolo. Additionally, most review sites use a method for the evaluation of the review as a score or mark that is represented by a real value ranging from 1 to 10.

2.2. Data Set Pre-processing

The textual content of the reviews extracted from the web site was preprocessed and organized as a corpus by adopting a simplified form of the XCES standard (Ide et al., 2000). All the words of the resulted corpus were annotated with syntactic data as it is further detailed.

[3] http://www.booking.com/

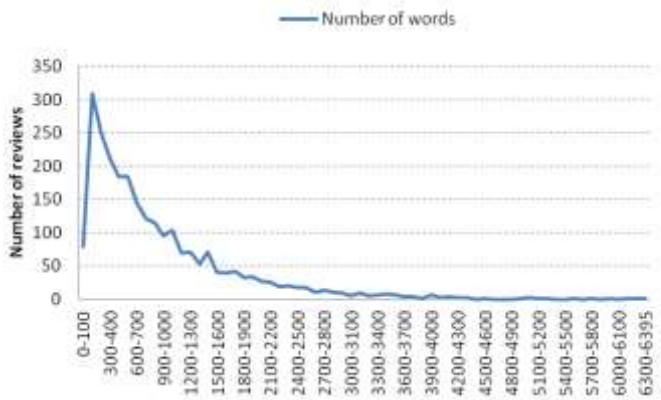


Figure 1: Number of reviews depending on their size counted as number of words

By running an automatic processing chain that includes sentence segmentation, tokenisation, POS-tagging and lemmatisation (Simionescu, 2011), the boundaries of the reviews sentences were marked and each word got attached its part of speech data and lemma. We need lemmas as the method of counting positive and negative terms requires the mapping of comments words into their base forms. The word tokens of the reviews sentences were automatically annotated for their head-words and the corresponding dependency relations by running a Dependency Parser[4].

The corpus resulted from preprocessing of the text extracted from the tourist reviews contains 91,966 sentences and 2,242,841 word tokens. The corpus elements are expressed in XML, as shown in what follows. The basic layers of annotation include: borders of each sentence (marked as `<S></S>` elements and identified by unique identifiers – attribute `id`) and words (marked as `<W></W>`) and including unique IDs, part of speech (attribute `POS`), lemma (attribute `LEMMA`) and dependency information (dependency relation - attributes `DEPREL` and `HEAD`).

```
<S id="1">
  <W DEPREL="det." HEAD="2" ID="1" LEMMA="un" POS="ARTICLE">un</W>
  <W DEPREL="ROOT" HEAD="0" ID="2" LEMMA="hotel" POS="NOUN">hotel</W>
  <W DEPREL="a.subst." HEAD="2" ID="3" LEMMA="cu" POS="ADPOSITION">cu</W>
  <W DEPREL="det." HEAD="5" ID="4" LEMMA="un" POS="ARTICLE">o</W>
  <W DEPREL="prep." HEAD="3" ID="5" LEMMA="gradină" POS="NOUN">gradină</W>
  <W DEPREL="a.adj." HEAD="5" ID="6" LEMMA="superbă"
POS="NOUN">superbă</W>
</S>
```

3. Sentiment Analysis Experiment

3.1. Lexical Resources

In our approach we rely on lexical information combined with the syntactical information obtained by running a Romanian automatic pre-processing chain available

[4] At the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iași, a Dependency Treebank for the Romanian language was built by the Natural Language Processing Group (Perez, 2012). We have used this resource for the training and evaluation stages of the Dependency Parser.

as a Web Service[5] and a Dependency Parser on the textual reviews contents. The lexical information assumes a list of keywords that must be looked up in the reviews and based on which, the analysis can be implemented, as well as a list of negative and positive words upon which the sentiment value of a review can be "calculated".

The features that contribute to increase the accuracy of sentiment analysis are the words in the lists of positive and negative terms (Kennedy and Inkpen, 2006). The idea of counting positive and negative terms or expressions was firstly proposed by (Turney, 2002). Note that the term-counting method does not require training, so it can be applied when training data is not available.

We used an English lexicon of positive and negative terms[6] that was firstly reported in (Hu and Liu, 2004). In order to be used on Romanian texts we translated it with the help of an English-Romanian dictionary[7]. By removing duplicates and by taking care that the intersection of the two lists of terms must be empty (otherwise the senses cancel each-other), we obtained 1943 terms with positive senses and 4678 terms with negative senses. Most of the terms are adverbs and adjectives in base form with no plurals and other inflected forms. We augmented the considered term-counting method by taking into account contextual valence shifters. Valence shifters are terms that can change the semantic orientation of another term, by increasing or decreasing its qualification or even by changing the sentiment of positive or negative terms in the sentence into their opposite value. As in (Polanyi and Zaenen, 2006), we considered as valence shifters the negative and intensifier terms.

Negations are terms that reverse the sentiment value of the word to which they apply. Examples of negation terms: "nu" (En. "not"), "nici" (En. "nor", "neither"), "niciodată" (En. "never"), "nicidecum" (En. "noway"), "deloc" (En. "at all"), etc. Intensifiers increase the intensity of a positive/negative term. Examples of intensifiers: "super", as in "super frumos" (En. "super nice"), "extrem de" (En. "highly"), "extraordinar de" (En. "extraordinarily"), etc.

3.2. The Sentiment Analysis Method

Our proposed approach is based on counting the positive and negative terms in a review that are related to *aspects* or *facets* of the object under discussion. In this approach, a review is considered positive if it contains more positive than negative terms, and negative if there are more negative than positive terms. A review is neutral if it contains an (approximately) equal number of positive and negative terms.

Following (Liu, 2012) in our proposed sentiment analysis method, the entities are conceptualized as triples (*entity*, *facet*, *sentiment*) where:

[5] The WSDL specifications are available at http://nlptools.infoiasi.ro/WebPosRo/.
[6] The lexicon is available at http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.
[7] The dictionary can be found at http://www.mcolhon.ro.

(1) An *entity* can be an accommodation unit (hotel, apartment, villa, etc.), as well as a tourist place or region. According to the *entity* type, some facets cannot be used to properly describe an entity aspect[8].

(2) The *facets* under discussion on AmFostAcolo are *services*, *accommodation*, *kitchen*, *landscape*, *entertainment*. The reviewers can grade these facets separately.

(3) The *sentiment* can be *positive*, *negative* or *neutral*.

The evaluation of the opinion sentiment towards a specific facet rather than the whole review is not easy. For example, a negative review does not necessarily mean that everything mentioned in the comment is negative. There can be some positiveness regarding a particular aspect. Likewise for a positive review. To obtain such detailed results we deepened our analysis at the sentence level to extract the relevant features.

To determine a sentiment towards a specific facet of an entity, we must find correlations between the corresponding facet and the textual content of the review. We manually built accordingly five sets of words so-called *seeds*. Each set is defined for determining the sentiment value of each specific facet. We relate the positive and negative terms with the seeds under investigation by means of the grammatical relations generated by the Dependency Parser on the reviews sentences, as well as by investigating the terms present around seeds using the following context window-based approach[9].

Our algorithm lookups to match the reviews words with the selected seeds. For each occurrence of seed *s* in the text:

(1) We select all positive and negative terms that are in a dependency relation with *s*.

(2) By applying the *bag-of-words* principle, we consider a fixed-size context-window around the current word matching a seed. We select the positive and negative terms within the window that are not in a dependency-based relation with other seeds.

(3) The positive/negative score for *s* is set to the number of selected positive/negative terms. Then we map the obtained sentiment scores of *s* to the corresponding facet.

One argument to support the correctness of the scores resulted by the application of this proposed Sentiment Analysis algorithm (*SA* in what follows) is that the positiveness and negativeness of the terms found in reviews are considered only if they affect the facets' seeds of the evaluated entities.

4. Results, Discussions and Conclusions

On the AmFostAcolo web site, the tourist entities are described by five facets which receive scores ranging from 0 up to 100.

We evaluated the resulted sentiment analysis scores by dividing the data set into positive, negative and neutral rankings. For a scale from 0 to 100, we consider that:

[8] For example, hotels as customer-based units are usually described by the *services* scores. This is not the case for tourist regions for which, usually, the *services* scores have 0 as value.
[9] The window is considered to include maximum 8 words immediately before and after the seed, being properly resized not to exceed the seed's sentence.

(1) A negative comment on a certain topic means that the score given by the user is between [0, 40]. The *SA* topic scores match the user score only if: negative score > positive score;

(2) A positive comment on a certain topic means that the score given by the user is between [60, 100]. The *SA* topic scores match the user score only if: positive score > negative score;

(3) A comment is considered neutral if the user's score is between [40,60]. The *SA* topic scores match the user score only if: $|positive\ score - negative\ score| \leq threshold$ (we considered $threshold = 1$).

Evaluation is performed by comparing the initial user reviews scores with the scores calculated by our *SA* algorithm. We evaluate all the reviews in terms of the standard measures of *Precision*, *Recall* and *F-score*.

Table 1: Evaluation Scores

Reviews	**Precision Corrects/Total/Score**	**Recall Corrects/Total/Score**	**F-score**
Positive	1981/2280/**0.87**	1981/2280/**0.87**	0.87
Negative	11/18/**0.61**	43/234/**0.18**	0.27
Neutral	2/7/**0.29**	2/7/**0.29**	0.29
ALL	1994/2305/**0.87**	2026/2521/**0.80**	0.83

Generally, we achieved an overall sentiment analysis *Precision* of 87% which is a good result. Our algorithm performs slightly worse for negative reviews (*Precision* = 60% and *Recall* = 18%) and neutral reviews (*Precision* and *Recall* around 30%). We consider that the weak results for negative comments come from what we call the *0-scores:* the 0 scores of reviews that do not match with 0 scores obtained by the application of our *SA* algorithm (i.e. our algorithm obtains nonzero positive and/or negative values). Indeed, even though it is not explicitly specified on the site, we suspect that sometimes the 0 scores assigned to a facet in a review means that its text does not describe anything related to that facet rather than assigning the maximum negative opinion to the facet. But there is another issue related to this score: we suspect that the default value for a score is 0 and if the user neglects or forgets to explicitly update it, the score remains assigned with the erroneous 0 value.

Following (McCallum, 1999), human classification has around 70% correctness because human raters typically agree about 70% of the time. For example, in (Bjørkelund et al., 2012) the authors evaluated a system to have around 70% accuracy and they concluded that it is "as good as human raters". The fact that the overall evaluation scores obtained with our proposed Sentiment Analysis algorithm are far above this score, enables us to conclude that our algorithm works well on the data set employed in our experiments.

References

Bjørkelund, E., Burnett, T.H., Nørvåg, K. (2012). A Study of Opinion Mining and Visualization of Hotel Reviews. In *Proceedings of the 14th International*

Conference on Information Integration and Web-based Applications and Services (iiWAS 2012), ACM, 229-238.

Gînscă, A. L., Iftene, A., Corîci, M. (2012). Building a Romanian Corpus for Sentiment Analysis. In *Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language"*, "Al. I. Cuza" University of Iaşi Publishing House, 63-71.

Gînscă, A. L., Boroş, E., Iftene, A., Trandabăţ, D., Toader, M., Corîci, M., Perez, C. A., Cristea, D. (2011). Sentimatrix – Multilingual Sentiment Analysis Service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (*ACL-WASSA2011*), Portland, Oregon, USA, June 19-24.

Hu, M., Liu, B. (2004). Mining Opinion Features in Customer Reviews. *Proceedings of Nineteeth National Conference on Artificial Intellgience* (AAAI-2004).

Ide, N., Bonhomme, P., Romary, L. (2000). Xces: An XML-based Encoding Standard for Linguistic Corpora. *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC 2000) European Language Resources Association.

Kaur, A., Gupta, V. (2013). A Survey on Sentiment Analysis and Opinion Mining Techniques. *Journal of Emerging Technologies in Web Intelligence*, 5:4, 367-371.

Kennedy, A., Inkpen, D. (2006). Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22:2, 110-125.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Liu, B., Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis: *Mining Text Data*. Springer US (Aggarwal, Charu C. and Zhai, ChengXian, eds.), 415-463.

McCallum, A. (1999). Text Classification by Bootstrapping with Keywords, EM and Shrinkage. *Proceedings of ACL99 - Workshop for Unsupervised Learning in Natural Language Processing*.

Perez, C.A. (2012). Casuistry of Romanian Functional Dependency Grammar. In *Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language"*, "Al. I. Cuza" University of Iaşi Publishing House, 19–28.

Polanyi, L., Zaenen, A. (2006). Contextual Valence Shifters. *Computing Attitude and Affect in Text: Theory and Applications* (Shanahan, J.G., Qu, Y., Wiebe, J., eds) 20. Springer, 1–10.

Simionescu, R. (2011). Hybrid POS Tagger. *Proceedings of "Language Resources and Tools with Industrial Applications" Workshop* (Eurolan 2011 summerschool)

Stabb, S., Werther, H., Ricci, F., Zipf, A., Gretzel, U., Fesenmaier, D.R., Paris, C., Knoblock, C. (2002). Intelligent Systems for Tourism: *IEEE Intelligent Systems*, 17:6, 53-66.

Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL 2002), 417-424.

AUTOMATIC IMAGE ANNOTATION

ANDREEA-ALICE LAIC, ADRIAN IFTENE

"Alexandru Ioan Cuza" University of Iaşi, Faculty of Computer Science
{andreea.laic, adiftene}@info.uaic.ro

Abstract

In the recent years, multimedia content has grown increasingly over the Internet, especially in social networks, where users often post images using their mobile devices. In these networks such as Flickr, the content is later used in search operations when some users want to find something using a specific query. Nowadays, searching into these networks is primarily made using the title and the keywords associated to resources added by users that have posted the content. The problem we face comes from the fact that in many cases, the title or the related keywords are not relevant to the resource and only after we analyse the image, can we conclude what it contains in reality. The project that we want to present in this article proposes that each image is connected to relevant keywords according to its content. In order to do this, the first step was to create a collection of images that was annotated by human annotators, while the second step was to expand this collection of images performing search on the Internet using keywords associated to the initial collection of annotated images. Currently, for a new picture, we can identify similar images in our collection of images and based on the keywords associated with them, we can determine what keywords characterize this new image. The evaluation of this system has demonstrated that our approach works efficiently for images for which we can find similar images in our collection.

Key words — Image retrieval, automatic image annotation, Flickr.

1. Introduction

The domain of image retrieval is dedicated to systems which deal with browsing, indexing and searching for images in a large context (Datta et al., 2008). Typically, this search is done by keywords, metadata and descriptions of images. The volume of data has significantly increased during the past years, which has led to the development of algorithms performing image processing, the *Image Retrieval* domain being in continuous expanding. Big companies like Google[1], Bing[2], Yahoo[3] have developed tools in time and they have optimized algorithms to be efficient while searching for images, a proof of this is the option "*Image Search*" that they offer.

Content-Based Image Retrieval is preferable because usual keyword search depends on the quality and accuracy of annotations (Eakins et al., 1999). Until now, Google has had the most complex system for automatic recognition of image elements.

[1] https://support.google.com/websearch/answer/1325808?hl=en
[2] https://www.bing.com/?scope=images&nr=1&FORM=NOFORM
[3] http://images.search.yahoo.com/

Google proposes a new type of image search, the one through similar images[4] (images that have similar content, both in color and texture, and the components of the image) to user data. This option is available only in the browser, allowing the user to drag-and-drop an image, enter the URL of the image or make a simple image upload. The advantage of this option (compared to what is now on the market) comes from the fact that the image database from Google is impressive (~100,000,000 gigabytes[5] of indexed pages). The disadvantage of this option regarding programmers is that Google still does not provide an API for application developers.

Similar to what Google offers, TinEye[6] developed a framework that allows you to perform a *reverse search* by image. There is a Web application where the user can enter an URL, drag-and-drop or upload an image and get similar results with the image inserted by him. Different from Google, TinEye offers an API for application developers, but the process of integration into an application development is chargeable.

RevIMG[7] is a search engine of images through other images. It provides a library for JavaScript and one for Android mobile applications. This engine is intended only to certain image categories like pictures, monuments, famous people, flags, etc.

Besides these applications, there are a series of platforms (*Lire*[8], *pHash*[9]) which are able to extract the content items (color, texture, etc.) of the image. The application that we developed uses *Lire*, a library corresponding to the application requirements in terms of type of search (which is done by image content) and speed of rendering the results (which is small).

2. System architecture

The first development step was to build a collection of images that were manually annotated with keywords, collection which was expanded, using the Internet, by searching the keywords associated to the initial collection of annotated images. Next, for a new picture offered by the user, we can identify similar images in our collection and based on the lists of keywords associated with them, we can determine what list of keywords characterizes this new image.

2.1. Creation of gold collection with annotated images

The initial collection of images consisted of 100 images, from different areas. The images were categorized in the following proportions: 30% images with peoples, 15% images from nature, 20% images with animals and the remaining images were from various categories (art, furniture, sport, other, etc.).

The images were selected by six human experts and then were manually annotated by human annotators. Some of the images have words in their visual content to see how

[4] https://support.google.com/websearch/answer/1325808?hl=en
[5] http://www.google.com/insidesearch/howsearchworks/crawling-indexing.html
[6] https://www.tineye.com/
[7] http://www.revimg.net/
[8] http://www.semanticmetadata.net/lire/
[9] http://www.phash.org/

this can influence the process of annotation. In Figure 1, you can see how a logged user can annotate an image.

Figure1: Application interface where users can annotate images

In the section "*Ce părere ai despre imagine*?" (English: *What is your opinion about this image*?), the user can select how much she/he liked the image shown. We record these opinions in our database and this action allows us to build profiles for users who have annotated images and also to build a recommendation system for them.

In the section "*Ce etichete ai asocia imaginii*?" (English: *What keywords would you associate to the image*?), the user can indicate a series of Romanian keywords that she/he considers suitable for the image. Besides the simple words, they can also write expressions which they consider appropriate for the image.

2.1.1. Experiments

In the process of annotating, there were 28 volunteers in third-year and master students of the Faculty of Computer Science from Iaşi. They had to annotate 100 images; the only criterion was to write keywords in the Romanian language, criterion that was established from the beginning.

Comparing the keywords entered by users for the same picture, it was seen that there were small differences among the words entered, most of them were from the same lexical family or they were synonyms. Each user was able to annotate how many pictures she/he wanted, but in the analysis entered only keywords by 21 users who have annotated all 100 images.

Performing an analysis on what users annotated over a period of two weeks, it can be said that their tendency was to introduce, on average, 3.41 keywords per image, with a minimum of 2 keywords for an image and a maximum of 12 keywords for an image. Looking further into the keywords that they have entered, it can be said that most users have opted for simple words and not phrases. As a general rule, they have chosen to

annotate the content of the image that quickly appears in sight. In the end, the 21 users have entered a total of 1,514 keywords for 100 images.

For example, for Figure 2, the users have chosen keywords such as *câine, cățel, copil, pat, cerceaf, puritate* (in English: "dog", "puppy", "baby", "bed", "bed sheet", "purity"), elements that can be easily seen in the image, and not keywords like *lemn* (in English: "wood"), which can hardly be seen in the background.

Figure 2: One of the images annotated by the users

Furthermore, we have implemented an algorithm which, for each image, counts the frequency of lemmas of the keywords associated by users and keeps those with a frequency of at least 4. Besides frequency, we considered the relation of synonymy using the Romanian WordNet (Tufiș et al., 2004). Among all synonyms, we kept the keyword which appears more often at the users who have annotated the image.

For expressions, we used the division into component words, and then we calculated the frequency of word components based on lemma and synonymy. If all components of the expression had an occurrence frequency over 4, we decided to keep the expression and give up the words which appeared in the expression. In the end, we considered for every image a list of keywords in a descending order of frequency (of course, for frequencies over 4).

In addition to the score calculated for each keyword based on frequency, we decided to calculate a score for each user who annotated all images. Furthermore, regarding the way we calculated the score for keywords, for the user's score we took into account the order of the entered keywords. For example, the user's score was calculated as a product

between the number of users who entered that keyword and its quota, given from the formula (1).Thus, we could identify the reliable and the less reliable annotators.

$$(1)\ \frac{1}{abs(index\_of\_keyword\_in\_final\_list - index\_of\_keyword\_in\_user\_list)+1}$$

Each image initially contained around 30-40 different keywords from all users, and afterwards, we applied the algorithm, the number of keywords was reduced to approximately 3-4 keywords per image. The average remained of 3.32 keywords per image, with a minimum of 1 and a maximum of 7. It can be seen that filtering was done quite rigorously.

After we performed the steps explained above for the image from Figure 2, we were left with the following keywords: *cățel*, *copil*, *pat* (in English: "dog", "child", "bed").

After completing this step, we increased the initial collection with 100 images as it follows. For each image from the initial collection, we added 10 new images to our collection, thus increasing the image collection to 1,000 images. For this, we searched for Google images using lists of keywords associated with each image. For the first 10 results, we initially associated the list of keywords used in the search process, followed by a process of verification, corrections, additions to this list; this process was done with human annotators.

2.2. Reverse Image Search

This module uses the 1,000 collection of images with related keyword lists obtained at the previous step. Regarding this collection, we know that the list contains relevant keywords associated with images.

Figure 3: Reverse Image Search application

The main purpose of this module is to generate a list of keywords that characterize an image given by the user. This is done as follows:

- The user introduces an URL and presses the **Proceed** button (See Figure 3).
- Behind the applications, we use the LIRE[10] library (Lucene Image REtrieval) (Lux and Marques, 2013), which compares the new image with the images from our collection. It establishes a set of 20 images most closely to the new inserted image in terms of texture and color. LIRE uses many low-level characteristics in the

[10] LIRE: http://www.semanticmetadata.net/lire/

indexing processing such as Color Layout[11], Edge Histogram (Park et al., 2000), CEDD[12] (Color and Edge Directivity Descriptor), FCTH (Fuzzy Color and Texture Histogram) (Chatzichristofis and Boutalis, 2008), etc. and then it uses the Euclidian distance for finding similar images. We used, in the first instance, the FCTH characteristic, but we also made experiments, with the other values.

- To establish the list of keywords that we associate with an image and their order in this list, we apply the algorithm from section 2.1.1. For that, the input that we use is represented by lists of keywords from 20 similar images, and then we use lemmatisation, the synonymy relation and the processing of expressions.
- An exception to the above is the case when all Euclidean distances between the new image and all images from the collection are below 0.2. This value was found experimentally and it tells us that the new image is too different in comparison with the existing images from our collection. In this case, we cannot associate keywords to the new image.

2.2.1. Use-cases

To illustrate the two cases described above, we carried out two searches to see how the application behaves.

2.2.1.1. There are similar images in the collection

This is the case when the application works as we wanted and it is able to build a list of relevant keywords to be associated to a new image.

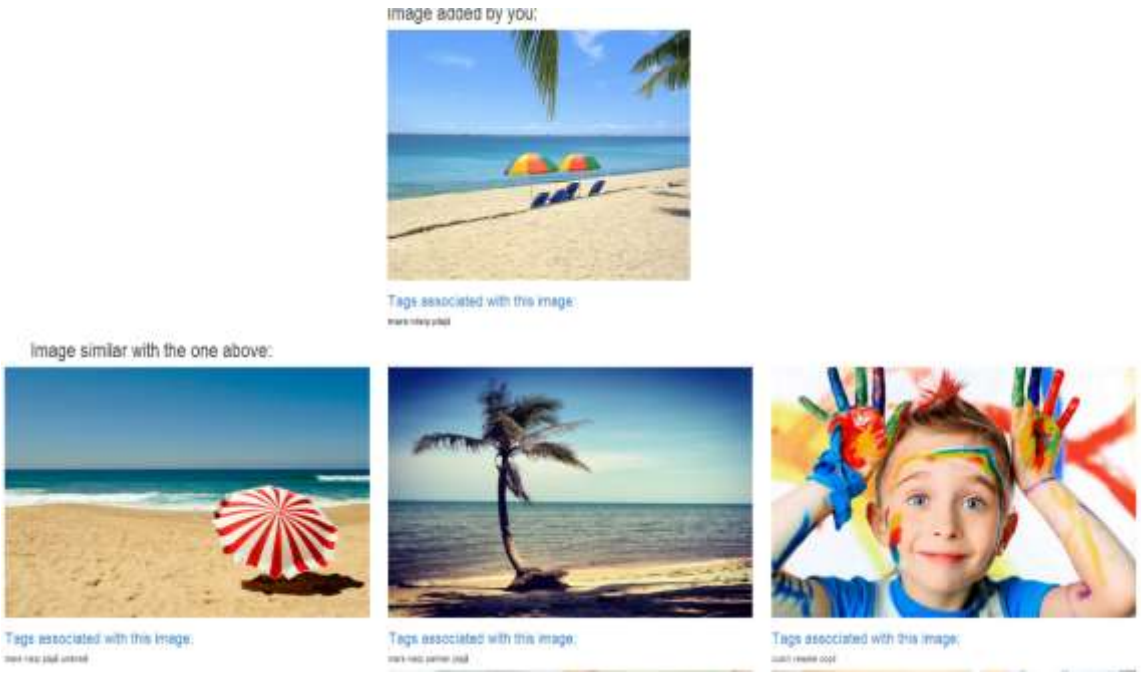


Figure 4: Reverse Image Search when there are similar images in the collection

[11] ColorLayout: http://en.wikipedia.org/wiki/Color_layout_descriptor
[12] CEDD: http://www.itec.uni-klu.ac.at/lire/nightly/api/net/semanticmetadata/lire/imageanalysis/CEDD.html

In Figure 4, it can be seen that there are similar images in the database with the one introduced by the user, and the list of keywords contains relevant keywords for this image.

2.2.1.2.There are no similar images in the collection

In this case, because of the limit imposed by the Euclidean distance, the user will receive a negative response. This means that the algorithm didn't find similar images with the user image in the collection of images, and thus it is unable to create a list of keywords that characterize it.

Figure 5: Reverse Image Search when there are not similar images in the collection

In Figure 5, it can be seen that in the collection of images there are no similar images with the one introduced by the user (the images shown in the second row have the Euclidian distance below the limit imposed by us). We also note that these images do not contain common keywords, which may characterize the new image inserted by the user.

The conclusion that can be drawn from the analysis of the two cases: the more images we have in our collection of annotated images, the more chances of finding similar images. This conclusion is also strengthened by the experiments that we perform in the next section.

2.2.2. Evaluation

To see how accurate the above system is, we conducted a series of experiments on two different sized collections of images with related keyword lists. The first collection has 100 images and the second collection has 200 images (100 from the first collection plus 100 similar to those). In the following pages, we present the experiments that we have done to evaluate the system.

We considered 20 new images taken from the Internet and then we used the application separately on two collections of images. For each of the 20 new images taken from the Internet, we have made processing using the system created and we monitored the following values:

- How many keywords are added, on average, to an image;
- How long the processing of an image takes;
- How many keywords added to an image are incorrect.

In Table 1, one can see the obtained results: the number of keywords added to a picture is, on average, 1.89 keywords for a collection with 100 images and 2.74 for a collection with 200 images. Of course, these values depend on the number of keywords associated to the images from our collections (where the average number was around 3.32).

Table 1: System evaluation

Image from database	**How many keywords are added, on average, to an image**		**The average length of the processing (seconds) per imagine**		**How many keywords added to an image are incorrect**	
Number of images from database	**100**	**200**	**100**	**200**	**100**	**200**
	1	1	52	101	0	0
	3	3	50	115	0	0
	2	4	27	240	0	1
	2	3	26	302	0	0
The average of those 20 images	*1.89*	*2.74*	*40*	*212*	*0.31*	*0.45*

The average duration for application was around 40 seconds for the collection of 100 images, and around 212 seconds for the 200 image collection. This duration varies due to the feature histogram FCTH.

The number of incorrect keywords for a picture is quite small. Wrong keywords appear when the terms of texture and color of an image are very similar to another image from the image collection, showing different elements in the picture.

As a conclusion, after we analysed the results from Table 1, we can say that the system created is a stable one and it offers good results to the user.

3. Conclusions

The application presented in this article can be very useful when you have a new image and you want to know what elements it contains. In order to do this, firstly, we need a large collection of annotated images with relevant lists of keywords. Secondly, we need performance algorithms which provide the distance between images in order to find images which are similar to the new one. Thirdly, we assign a list of keywords to the new image by making the intersection of keywords from similar images.

After evaluating the created system, we can say that the system works effectively as long as the requested image finds similar images in our collection. Consequently, it is very important that this collection be very large.

One problem that arises comes from the fact that the application responds slowly when the collection of used images is large (as we can see in Table 1).

Therefore, future directions for improving this application are the following: (1) the first direction is related to optimal and accurate algorithms that can identify specific elements in the new image (such as buildings, trees, people, sky, sea, etc.). These algorithms do not depend on the size of the collection of images and the offered results can be very fast and very accurate. (2) A second direction concerns the increase of the collection of annotated images, but it needs to be combined with the use of cloud platforms in order to have a low response time.

Acknowledgments

The research presented in this paper was funded by the project MUCKE (Multimedia and User Credibility Knowledge Extraction), number 2, CHIST-ERA/01.10.2012.

References

Chatzichristofis, S., Boutalis, Y. S. (2008). FCTH: fuzzy color and texture histogram, a low level feature for accurate image retrieval. In *Proceedings of WIAMIS'08 Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, IEEE Computer Society Washington, 191-196.

Datta, R., Joshi, D., Li, J., Wang, J. Z. (2008). Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys (CSUR)*, 40:2(5), 1-60.

Eakins, J., Graham, M. (1999). Content-based Image Retrieval. *Library and Information Briefings*, 85, 1-15.

Lux, M., Marques, O. (2013). Visual Information Retrieval using Java and LIRE. *Synthesis Lectures on Information Concepts, Retrieval, and S*, Morgan & Claypool Publishers.

Park, D. K., Jeon, Y. S., Won, C. S. (2000). Efficient use of local edge histogram descriptor.In Proceedings of *the 2000 ACM workshops on Multimedia (MULTIMEDIA '00)*. ACM, New York, NY, USA, 51-54.

Tufiş, D., Barbu, E., BarbuMititelu, V., Ion, R., Bozianu, L. (2004). The Romanian Wordnet. *Romanian Journal of Information Science and Technology*, 7:1-2, 107-124.

HOW TO DO DIVERSIFICATION IN AN IMAGE RETRIEVAL SYSTEM

ADRIAN IFTENE[1], ALEXANDRA-MIHAELA SIRIȚEANU[1], MIRCEA PETIC[1, 2]

[1] *"Alexandru Ioan Cuza" University, Faculty of Computer Science, Iasi, Romania*

[2] *"Alecu Russo" Balti State University, Balti, Republic of Moldova*

{adiftene, alexandra.siriteanu, mircea.petic}@info.uaic.ro

Abstract

MUCKE (Multimedia and User Credibility Knowledge Extraction) is a CHIST-ERA research project whose goal is to create an image retrieval system that takes into account available information from social networks. In this paper, we give a short overview of the MUCKE project**,** and we present the work done by the UAIC group. MUCKE incorporates modules for processing multimedia content in different modes and languages and UAIC is responsible with text processing tasks. One of the problems addressed by our work is related to search results diversification. In order to solve this problem, we first process the user queries in both languages and secondly, we create clusters of similar images.

Key words — image retrieval, search diversification, YAGO.

1. Introduction

In the last years, social networks have been used not only for sharing multimedia data, but also as the main method to fulfil their information needs. MUCKE project addresses this stream of multimedia social data with new and reliable knowledge extraction models designed for multilingual and multimodal data shared on social networks. One of the aims of this project is to give a high importance to the quality of the processed data by protecting the user from an avalanche of equally topically relevant data. The project comes with two central innovations: automatic user credibility estimation for multimedia streams and adaptive multimedia concept similarity. Credibility models for multimedia streams are a highly novel topic, which will be cast as a multimedia information fusion task and will constitute the main scientific contribution of the project. In this context, we build a novel image retrieval framework that performs a semantic interpretation of the user queries and returns a diversified and accurate result set (Iftene and Alboaie, 2014).

Over time, various theories involving search results diversification have been developed, theories that have taken into consideration (Drosou and Pitoura, 2010): (i) content (Gollapudi and Sharma, 2009), i.e. how different the results are from each other, (ii) novelty (Carbonell and Goldstein, 1998; Clarke et al., 2008), i.e. what the new result offers in addition to the previous ones, and (iii) semantic coverage (Zheng et al., 2012), i.e. how well covered the different interpretations of the user query are. In the MUCKE project, we work with a collection of approximately 80 million images and their associated metadata that have been downloaded mainly from the Flickr database. Over this collection, we perform several processing tasks at both textual (on associated metadata) and image level and retrieve the results in a diversified way.

2. *MUCKE project*

The purpose of the system is to enable users to create and retrieve multimedia content. However, for evaluation purposes, it also supports the possibility to extract and index various image collections, such as the ImageCLEF[1] collection.

Figure 1 shows an overview of the MUCKE framework, covering how documents are processed, concepts extracted and indexed, similarity computed based on concepts, text and images, and how credibility is estimated and fed into the re-ranking process to improve the final set of results.

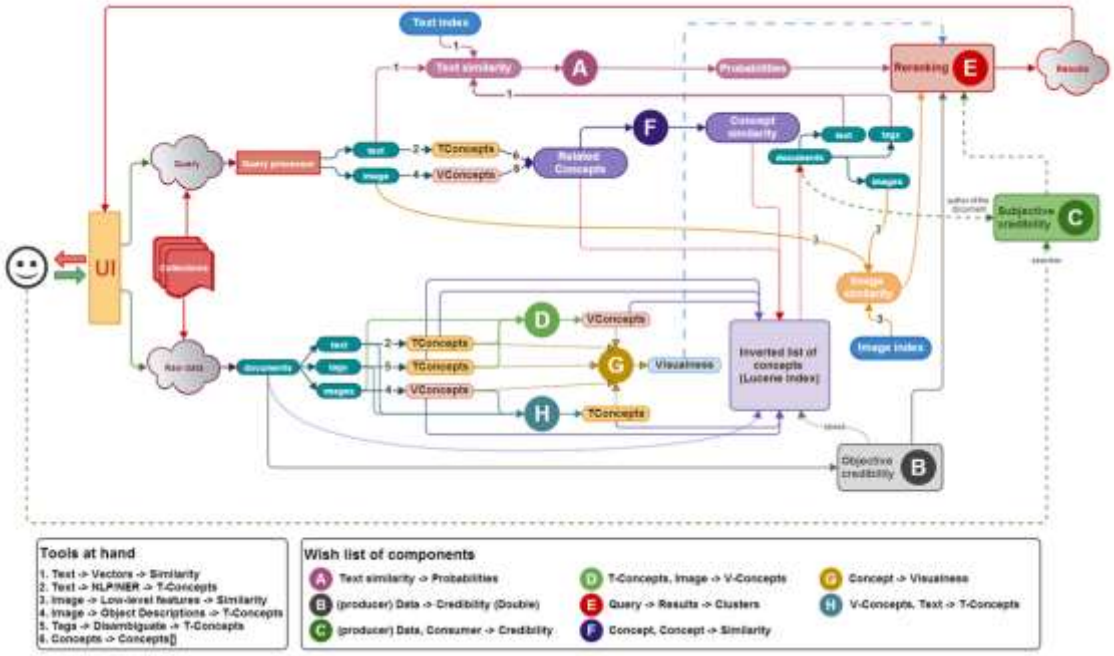


Figure 1: MUCKE Framework (Bierig et al., 2014)

One of the project goals is to prove the feasibility of the models and methods over large-scale multimodal data. The data collected for the MUCKE project is highly dynamic and complex, two characteristics required for an extraction framework to be implemented in a flexible manner so as to cope with new data whenever needed. The project works with approximately 80 million images belonging to almost one million Flickr[2] users, including their metadata, and approximately 10 million Wikipedia articles in 4 languages (English, Romanian, German and French), with their associated multimedia elements from Wikimedia[3]. The concept similarity resources include already around 100,000 multimedia concepts, roughly five times more than ImageNet[4], a widely used resource in computer vision.

[1] ImageCLEF: http://www.imageclef.org/
[2] Flickr: https://www.flickr.com/
[3] Wikimedia: https://www.wikimedia.org/
[4] ImageNet: https://www.image.net/

3. Text processing module

The text processing module is used to process on one hand, the images associated metadata and, on the other hand, the user queries. For the text processing tasks, standard tools are used for POS-tagging (Simionescu, 2011), lemma identification (Simionescu, 2011) and named entity identification (Gînscă et al., 2011). After the images associated metadata are processed, the image collection is indexed with Lucene[5]. In order to achieve diversification in the results set, the system incorporates a query expansion module that makes use of the YAGO[6] ontology.

YAGO ontology comprises well known knowledge about the world (Hoffart et al., 2013). It contains information extracted from Wikipedia[7] and other sources like WordNet[8] and GeoNames[9] and it is structured in elements called entities (*persons*, *cities*, etc.) and facts about these entities (which *person* worked in which *domain*, etc.). For example, with Yago we are able to replace in a query like "*tennis player on court*", the entity "*tennis player*" with instances like "*Roger Federer*", "*Rafael Nadal*", etc. Thus, instead of performing a single search with the initial query, we perform several searches with the new queries, and in the end we combine the obtained partial results in a final result set. Because of its structure, YAGO will be used only when the text queries will match WordNet concepts that are linked by a hypernymy relationship to other Wikipedia entities, such as person, location or organisation.

Wikipedia: to decide when to use YAGO, we created a resource based on hierarchies of Wikipedia categories. For this, we started with Romanian Wikipedia which has 8 groups of categories: culture, geography, history, mathematics, society, science, technology, privacy. In turn, these categories have subcategories or links to pages directly, as follows: Culture (30) (among which we mention *photo*, *architecture*, *art*, *sports*, *tourism*, etc.) Geography (15) (among which we mention *Romania*, *Africa*, *Europe Countries*, *maps*, etc.), History (6) (among which we mention *After the recall*, *By region*, etc.), Mathematics (11) (among which we mention *Algebra*, *Arithmetic*, *Economics*, *Geometry*, *Logic*, etc.), Society (22) (among which we mention *Anthropology*, *Archaeology*, *Business*, *Communications*, *Philosophy* , *Politics*, etc.), Science (23) (among which we mention *Anthropology*, *Archaeology*, *Astronomy*, *Biology*, etc.), Technology (19) (among which we mention *Agriculture*, *Architecture*, *Biotechnology*, *Computer*, etc.), Private life (8) (among which we mention the *Fireplace*, *Fun*, *People*, *Health*, etc.). In the end, we obtained 8 big groups with 134 categories, which are subdivided into several subcategories and pages (hierarchical depth depends on each category and subcategory). In general, this hierarchy covers most of the concepts available for Romanian. For example, for Sport, we obtained 70 subcategories containing other subcategories and 9 pages. Going through these categories and subcategories, we built specific resources with words that signal concepts of type *person*, *location* and *organisation*.

[5] Lucene: http://lucene.apache.org/
[6] Yago: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
[7] Wikipedia: http://en.wikipedia.org/wiki/Main_Page
[8] WordNet: http://wordnet.princeton.edu/
[9] GeoNames: http://www.geonames.org/

Some examples of signal words from these categories are:

- For Person: *acordeonist*, *actor*, *inginer*, *antropologist*, *arheolog*, *arhitect*, *femeie*, *arhivist*, *asasin*, *astronaut*, *astronom*, *astrofizician*, etc. (En: "accordionist", "actor", "engineer", "anthropologist", "archaeologist", "architect", "woman", "archivist", "assassin", "astronaut", "astronomer", "astrophysicist"). This is the biggest resource with over 391 signal words.
- For Location: *continent, ţară, oraş, comună, sat, regiune, munte, râu, fluviu, piaţă, stradă, bulevard, târg, instituţie, universitate, spital, teatru,* etc. (En: "continent", "country", "city", "township", "village", "region", "mountain", "river", "market", "street", "avenue", "fair", "institution", "university", "hospital", "theatre").
- For Organisation: *companie*, *SRL*, *partid*, *grupare*, etc. (En: "company", "Ltd", "party", "group").

Examples:

(1) starting from a query that includes the word *actor* (En: "actor"), it decides to use YAGO because our system identifies this word in the list with signal words for type *person*, and it calls a Sparql query with the following form:

```
PREFIX yago:<http://yago-knowledge.org/resource/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

select ?instance ?category ?length where
{
   {select distinct ?instance
     where{
        ?class rdfs:label "actor"@ron.
        ?category rdfs:subClassOf ?class.
        ?instance rdf:type ?category.
     }
    LIMIT 5000
   } .
   ?instance yago:hasWikipediaArticleLength ?length.
   ?instance rdf:type ?category.
   ?class rdfs:label "actor"@ron.
   ?category rdfs:subClassOf ?class.
}
order by desc(?length) LIMIT 2000
```

The results retrieved by YAGO are ordered by their article length and include entities like: *Ronald Reagan, Jennifer Lopez, Elvis Presley, Madonna, Hulk Hogan, Clint Eastwood, Linda Ronstadt, Steven Spielberg, Orson Welles, Britney Spears, Eminem, Paul Robeson, John Cena, Lindsay Lohan, Cher*, etc. It is noted that not all entities are of type actor (for example *Steven Spielberg*), but most are.

After performing a search on Google with the word *actor* (En: "actor") we obtain the results from Figure 2. After performing the same search in our application, we obtain the results from Figure 3.

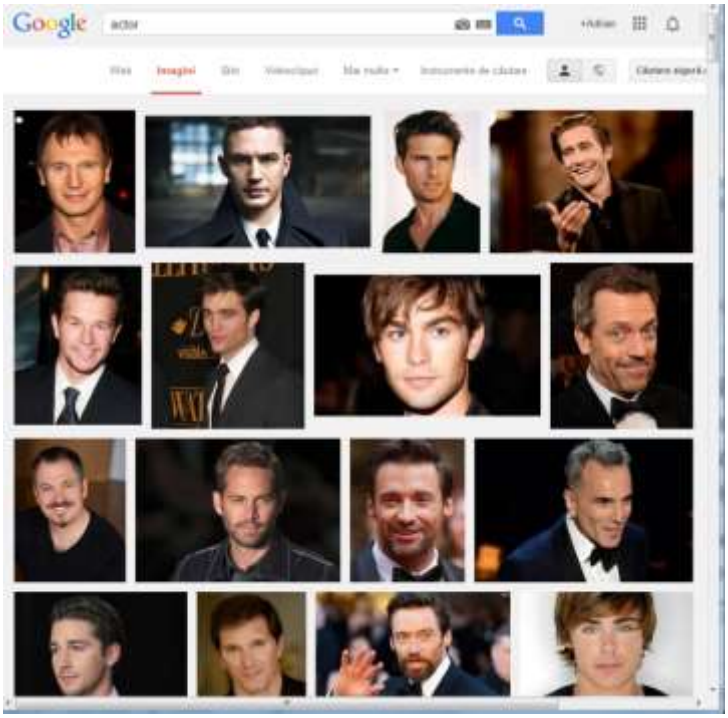


Figure 2: Results offered by Google Image Search for the query "actor"

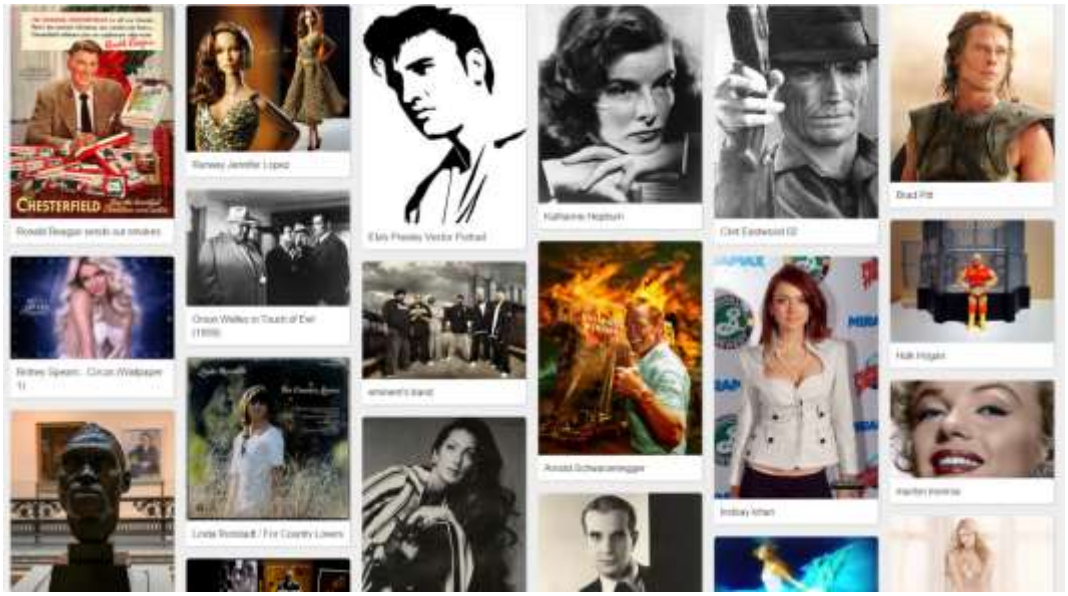


Figure 3: Results offered by our application for the query "actor"

(2) starting from a query that includes the word *companie* (En: "company"), it decides to use YAGO because our system identifies this word in the list with signal words for type *organisation*, and it calls a Sparql query with the following form:

```
PREFIX yago:<http://yago-knowledge.org/resource/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
```

```
select ?instance ?category ?length where
{
   {select distinct ?instance
     where{
        ?class rdfs:label "companie"@ron.
        ?category rdfs:subClassOf ?class.
        ?instance rdf:type ?category.
     }
    LIMIT 5000
   } .
   ?instance yago:hasWikipediaArticleLength ?length.
   ?instance rdf:type ?category.
   ?class rdfs:label "companie"@ron.
   ?category rdfs:subClassOf ?class.
}
order by desc(?length) LIMIT 2000
```

The results retrieved by YAGO include entities like: *Cirque du Soleil, English National Opera, Théâtre Lyrique, American Ballet Theatre, The Royal Ballet, New York City Opera, Tulsa Ballet, San Francisco Opera, Pacific Northwest Ballet, The Second City*, etc. It is noted that the majority are *operas* and *ballet companies*.

After performing a search on Google with the word *companie* (En: "company") we obtain the results from Figure 4. After performing the same search in our application, we obtain the results from Figure 5.

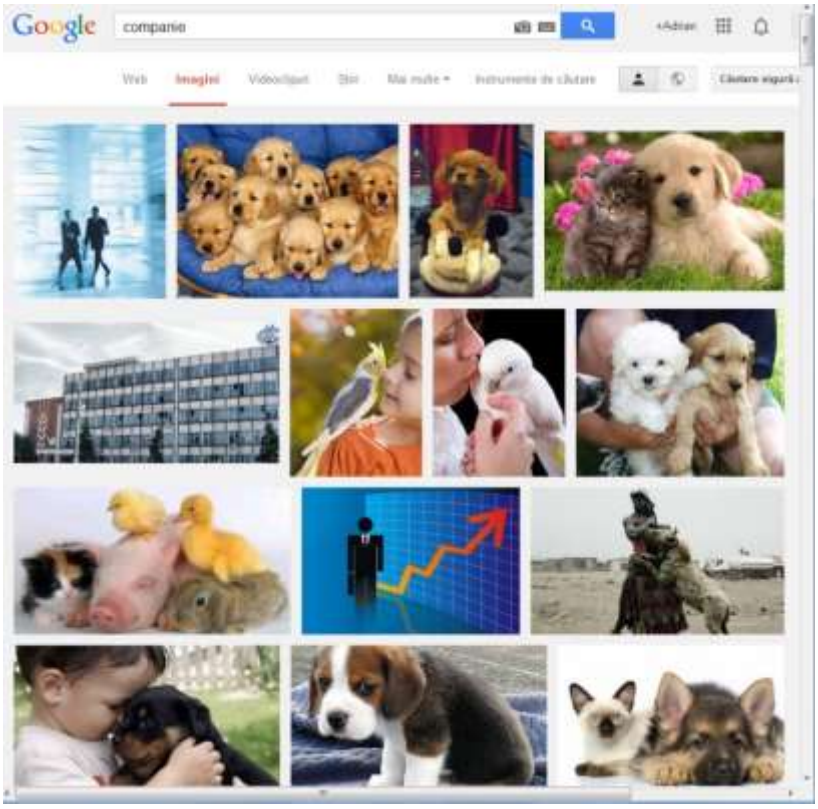


Figure 4: Results offered by Google Image Search for query "companie"

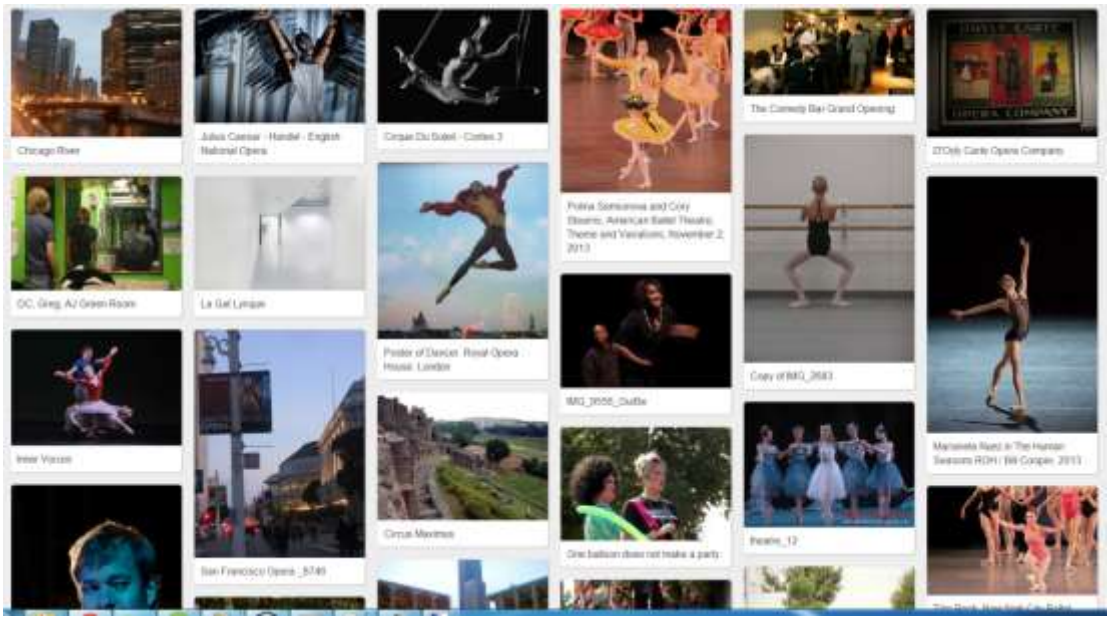


Figure 5: Results offered by our application for query "companie"

(3) starting from a query that includes the word *munte* (En: "mountain"), it decides to use YAGO because our system identifies this word in the list with signal words for type *location*, and it calls a Sparql query with the following form:

```
PREFIX yago:<http://yago-knowledge.org/resource/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

select ?instance ?category ?length where
{
   {select distinct ?instance
     where{
        ?class rdfs:label "munte"@ron.
        ?category rdfs:subClassOf ?class.
        ?instance rdf:type ?category.
     }
    LIMIT 5000
   } .
   ?instance yago:hasWikipediaArticleLength ?length.
   ?instance rdf:type ?category.
   ?class rdfs:label "munte"@ron.
   ?category rdfs:subClassOf ?class.
}
order by desc(?length) LIMIT 2000
```

The results retrieved by YAGO include entities, such as, *Rogue River (Oregon), Aliso Creek (Orange County), Ore Mountain passes, Santa Ana River, Matterhorn, Klamath River, Lō'ihi Seamount, Mount St. Helens, Mount Pinatubo, Mount Edziza volcanic complex, Mount Garibaldi, Metacomet Ridge, San Juan Creek, Mount Rainier, Mount Baker*, etc. It is noted that many of the entities are of type *creek* or *river*, but in this case this kind of entities can be easily eliminated with simple rules from our list.

After performing a search on Google with the word *munte* (En: "mountain") we obtain the results from Figure 6. After performing the same search in our application, we obtain the results from Figure 7.

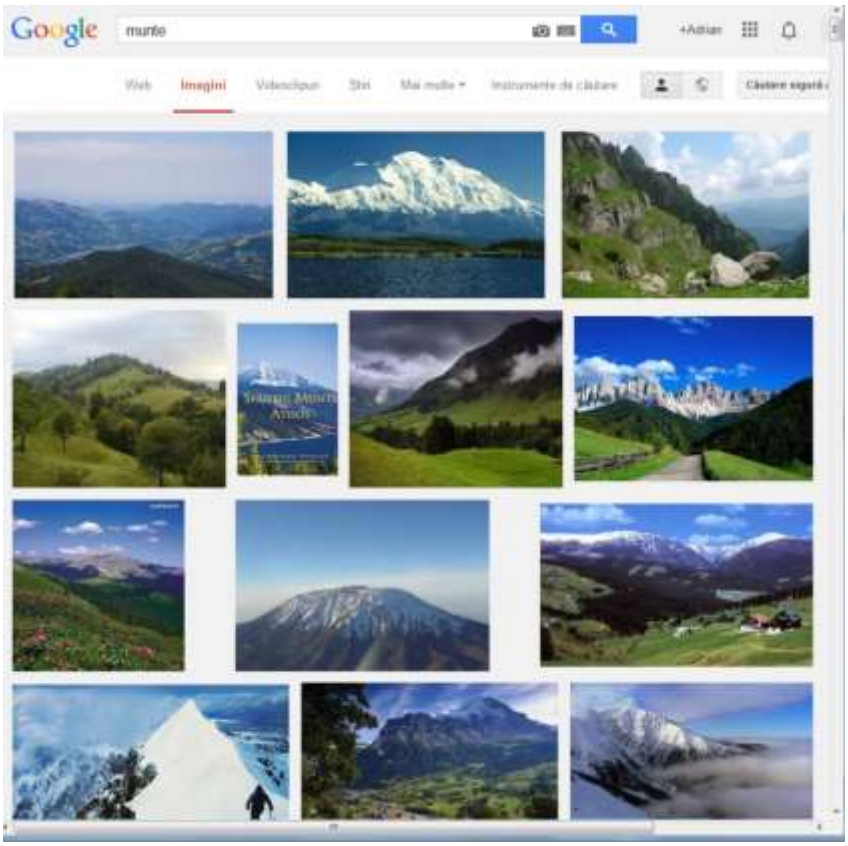


Figure 6: Results offered by Google Image Search for query "munte"

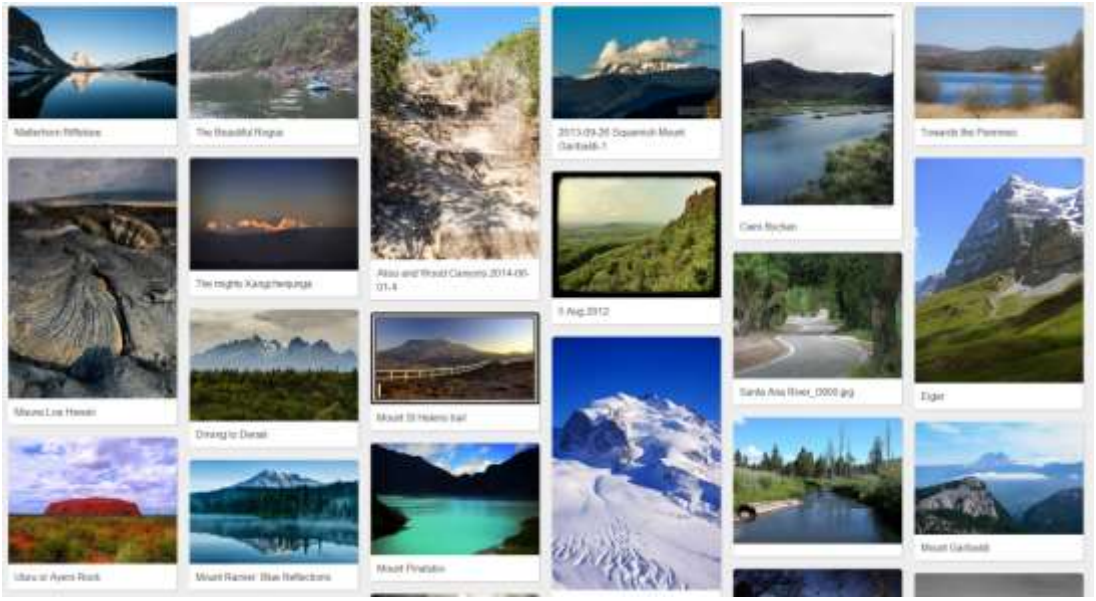


Figure 7: Results offered by our application for query "munte"

In all three cases the results offered by Google are similar from the point of view of concepts presented in images returned. In the case of our application there are more "colours" and more concepts in comparison with the results offered by Google.

Query reformulation module provides a technique of processing a given query by obtaining new concepts that are both efficient and relevant in the context of information retrieval operations. This module is very similar to the module responsible with question analysis in a question answering system (Iftene et al., 2010). In this case, we face two major issues that occur when an end user entered a query: *it is not precise enough*, meaning that there are too many results returned, most of them being irrelevant or *it is not abstract enough*, meaning that the search does not return any results at all. Here, we apply two approaches: (1) *a global technique*, which analyses the body of the query in order to discover word relationships (synonyms, homonyms or other morphological forms from WordNet), to remove stop words (*un*, *la*, *pentru*, (English: "a", "at", "for"), etc.), to remove wh- words (*cine*, *ce*, *de ce*, *unde*, (English: "who", "what", "why", "where"), etc.) and to correct any spelling errors; (2) *local feedback* which implies the analysis of the results returned by the initial query, leading to re-weighting the terms of the query and relating it with entities and relationships originating from the target ontology.

4. Image processing module

The information retrieval system queries the image collection against the newly obtained queries (when we are in one of the cases presented in section 3) or against the initial query (in other cases) and returns a collection of images with their associated metadata. In both cases, the image processing module performs a diversification task on the returned results. The aim of this module is to create clusters with similar images and instead of offering to user all the results, the system retrieves only one representative image from every cluster. In this way similar images are hidden and the user is able to see all the pictures on his request.

In order to accomplish this, we use Matlab[10] and its predefined functions to extract visual characteristics such as shape, colour, texture, etc. Also, we implement a naive algorithm that calculates the Euclidean distance between the average colours of the two images. Using these features, we organize the files in clusters using the DBSCAN algorithm (Ester et al., 1996) and we display the resulting clusters to the user.

5. Conclusions

In this paper we present our current work in MUCKE project. The paper addresses the diversification problem that is very important in an image retrieval system. For that, we perform several text processing tasks on user queries and we identify signal words related to entities of type *person*, *location* or *organisation*. If such words are identified, we use YAGO to expand the user query and we perform several searches with the new queries in our image collection. Finally, we perform image processing tasks in order to create clusters of similar images.

[10] Matlab: http://www.mathworks.com/products/matlab/

From what we have seen so far, the results are promising, and as future work we want to develop a module that allows us to evaluate the created system.

Acknowledgments

The research presented in this paper was funded by the project MUCKE (Multimedia and User Credibility Knowledge Extraction), number 2 CHIST-ERA/01.10.2012. We would also want to thank to the students from Faculty of Computer Science.

References

Bierig, R., Șerban, C., Sirițeanu, A., Lupu, M., Hanbury, A. (2014). A System Framework for Concept- and Credibility-Based Multimedia Retrieval. In *ICMR '14 Proceedings of International Conference on Multimedia Retrieval*, Glasgow, Scotland, April 2014, 543-550.

Carbonell, J. G., Goldstein, J. (1998). The use of mmr, diversity-based re-ranking for reordering documents and producing summaries. *SIGIR*, 335-336.

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Bttcher, S., MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. *SIGIR* (2008), 659-666.

Drosou, M., Pitoura, A. (2010). Search result diversification. *SIGMOD* (2010), 41-47.

Ester, M., Kriegel, H. P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD-96, AAAI (1996)*, 226-231.

Hoffart, J., Suchanek, F., Berberich, K., Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Elsevier, Artificial Intelligence*, 194 (2013), 28-61.

Gînscă, A. L., Boroș, E., Iftene, A., Trandabăț, D., Toader, M., Corîci, M., Perez, C. A., Cristea, D. (2011). Sentimatrix - Multilingual Sentiment Analysis Service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-WASSA2011)*, Portland, Oregon, USA.

Gollapudi, S., Sharma, A. (2009). An axiomatic approach for result diversification. *WWW*, 381-390.

Iftene, A., Alboaie, L. (2014). Diversification in an image retrieval system. *IMCS-50. The Third Conference of Mathematical Society of the Republic of Moldova dedicated to the 50th anniversary of the foundation of the Institute of Mathematics and Computer Science*. August 19-23, 2014, Chisinau, Republic of Moldova.

Iftene, A., Trandabăț, D., Moruz, A., Pistol, I., Husarciuc, M., Cristea, D. (2010). Question Answering on English and Romanian Languages. In C. Peters et al. (Eds.): *CLEF 2009, LNCS 6241, Part I (Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments)*, Springer, Heidelberg, 229-236.

Simionescu, R. (2011). Hybrid POS Tagger. In *Proceedings of "Language Resources and Tools with Industrial Applications" Workshop* (Eurolan 2011 summerschool).

Zheng, W., Wang, X., Fang, H., Cheng, H. (2012). Coverage-based search result diversification. *Journal IR (2012)*, 433-457.

AN AUTOMATIC SYSTEM FOR IMPROVING BOILERPLATE REMOVAL FOR ROMANIAN TEXTS

ALEX MORUZ[1, 2], ANDREI SCUTELNICU[1, 2]

[1] *"Alexandru Ioan Cuza" University, Faculty of Computer Science, Iaşi – România;*
[2] *Romanian Academy, Iaşi Branch, Institute for Computer Science*

{mmoruz, andrei.scutelnicu}@info.uaic.ro

Abstract

Within the current context of the information society, the importance of representative corpora for languages is becoming more and more important. The CoRoLa (COntemporary ROmanian LAnguage) project aims to create such a corpus for the Romanian language and intends to gather and annotate several hundred million words. Most of the texts that are to be processed are given in a formatted medium (such as PDF, HTML or DOC), so they need to be subjected to boilerplate removal in order to extract the relevant content. This paper describes our proposed solutions to a set of problems specific to boilerplate removal for Romanian language documents, namely diacritic encoding and paragraph detection.

Key words — boilerplate removal, glossary based diacritic extraction, CoRoLa.

1. Introduction

The Computational Reference Corpus of the Contemporary Romanian Language (CoRoLa) is a project of the Romanian Academy, which will take place between the years 2014 and 2017, in collaboration between the Romanian Academy Research Institute for Artificial Intelligence[1] of Bucharest and the Institute of Computer Science of Iasi[2]. The corpus will[3]:

- be representative, as it will contain all functional styles, as well as all (of possible) domains of activities;
- represent the contemporary form of the Romanian language, as it will include only those texts and voice from 1945 up to present times;
- be composite, as it will incorporate a collection of distinct texts and recordings, and each component will be a part or an entire document (printed or recorded);
- be open, as after the ending of the project other documents can be added;
- possibly be continued after 2017, with constant addition of documents and recordings, the corpus could become a monitoring corpus for the Romanian language.

[1] http://www.racai.ro/en
[2] http://iit.academiaromana-is.ro/
[3] As described in (Barbu Mititelu et al., 2014) and in the IIT technical report for the CoRoLa project, Natural Language Processing research group, June 2014

In this paper we will describe some proposals for the automatic acquisition and cleaning of raw textual resources in order to prepare the basic texts for automatic processing and inclusion in the corpus.

2. Extracting Relevant Text from Target Documents

One of the most important issues in corpus creation is that of extracting the relevant base text from formatted documents by removing unwanted artefacts such as figures, tables, headers, etc. This section describes some of the more relevant methods for identifying text content. Extracting information from a formatted document is based on a) visual representation, b) text structure and c) content.

Extracting information on the basis of visual representation is performed by identifying important blocks of text on the basis of visual features. Extraction is carried out on the basis of features such as text length (word count or character count), capitalisation, punctuation, etc. More advanced techniques can determine the document topic and then extract information relevant to that topic.

The most prevalent methods for text content extraction are based on heuristics. Some of the most used heuristics for text content identification are given below:

- The relevance of any given text can be computed on the basis of word length or character number;
- Titles or footnotes can sometimes be recognized by comparing their length to that of neighbouring paragraphs (titles, for example, are generally shorter than the average paragraph);
- Lists can be identified by recognizing tags used, the length of text for each tag, capitalisation and the probability of continuing the list;
- Paragraph identification can be carried out by calculating the ratio of punctuation marks to the number of words and the ratio of function words to the total number of words.

In order to reduce the error rate, the heuristic based methods are complemented with machine learning techniques using decision trees (Quinlan, 1993) and genetic algorithms (Hofmann et al., 2007). Generally, heuristic based methods are significantly more time-efficient than manual correction and extraction, even though the error rate for automatic extraction is higher.

In addition to heuristic based text extraction, some methods for relevant text identification are inspired from web based approaches. One of the more important approaches of this type is that of visual page segmentation, which identifies those elements in a page which have the same semantic structure and use annotated graphs for extracting the required information (Buneman et al., 1997). This type of approach can be used on texts which have no previously defined structure and for which the identification of even a basic structure can greatly improve the systems performance.

If the schema used by the text is not known, then the semantics of the text can be determined and used for optimizing the extraction process (Suciu, 1996). Another method for determining the semantics of a page is that of using human visual perception

(Deng et al., 2003). Usually, human visual structures are different from those used by the structures which represent documents (such as HTML structures), and therefore independent methods of text structuring are necessary (bottom-up approaches).

Databases have also been used for storing text information, either in a relational manner or an object oriented one (Papakonstantinou et al., 1995).

In the following sections we will describe two algorithms which we have used for extracting raw text from formatted documents (mostly PDF format). The algorithms focus on identifying paragraph boundaries and diacritic correction.

3. Boilerplate Removal for Romanian Texts

Because most of the documents that will be included in the corpus are in PDF format, we had to make use of a web based application that transforms a given PDF file into an editable text file. This process is carried out using the software package Apache PDFBox, which is a boilerplate removal program freely available from the Apache foundation[4]. This software package is responsible for extracting the text content of any PDF file, including headers, footers, footnotes, etc., but it does not keep graphical information such as tables, figures or graphs.

Although the accuracy of the transformation is good, it is completely dependent on the encoding of the base PDF file, and, as such, many of the extracted text files have content problems such as inserting new lines after each row of text, replacing non ASCII diacritics with other ASCII characters, mixing of text from different columns, etc. The subsections given below describe solutions for some of the most common problems encountered, inserting new lines after each row of text and replacing non ASCII diacritics with other ASCII characters

3.1. Automatic Correction of Basic Texts

3.1.1. Extracting Paragraph Boundaries

After extracting the basic text from a PDF file, it is a common occurrence to find the *newline* character (*n*) at the end of each row of text from the original formatted document. These extra characters can create problems for automatic processing tools and, as such, need to be removed. Removing these extra characters, however, is not trivial, as the boundaries of the original paragraphs need to be estimated in order not to change the nature of the acquired text. The excerpt below gives an example of such a case, where each row of text is followed by the *newline* character; the paragraph boundary that needs to be kept is highlighted in grey.

Tanti Lenuța, fusese cãsãtoritã cu un "nemernic"
(cum ziceau mama și bunica mea), profesor pe
undeva, într-o comunã peste dealuri, care a pãrãsit-
o pentru altã femeie, iar mama mea a divorțat de
tatăl meu când eu aveam numai un an. Deci …două

[4] http://pdfbox.apache.org/

femei fără bărbați, speriate de viitorul sumbru care se prevedea la orizontul roșu și trei copii (ba chiar patru, dacã o punem la socotealã și pe verișoara mea Rodica Popescu) puși numai pe joacă și năz-bâtii, în ciuda sărăciei de care nu eram conștienți - cam acesta era mediul în care trãiam și creșteam mari zi cu zi.
Din când în când mama mea si cu tanti Lenuța îșsi găseau timp să stea de vorbă, uneori ore întregi. Ce aveau să-și spunã? Curios din fire cum eram, trã-geam cu urechea. Vorbeau în șoaptã și nu prea

In order to remove these characters, we have performed a case study over a set of 20 documents from different sources (totalling up to over 500 page of original documents) in order to observe the cases in which new lines are added. The case study showed the following:

- This issue is to be found only in the case of acquiring text from PDF formats due to the different styles of coding the final PDF file;
- In some cases, new paragraphs are indented using white spaces (tabs, multiple spaces, etc.);
- For those cases where indenting is not kept, we have proposed a heuristic for determining the start of each paragraph.

On the basis of the observations stated above, we have created an algorithm for extracting paragraph boundaries, which is described below. The algorithm receives as input the text file extracted from the source document and returns a text file which contains the text with correct paragraphs.

```
1. foreach line in input, lineList.add(line)
2. foreach line in lineList
   a. if(line starts with whiteSpace) then add paragraph
   b. else if(line starts with "-" or capital letter and
      previousLine ends with ".?!") then add paragraph
   c. else if(previousLine ends with "-") then add line without
      white spaces and check if the word exists in the dictionary
      without "-"
   d. else add line
3.  return file
```

After testing the algorithm on several texts, we have estimated an error rate below 3% for paragraph boundary detection. For the example given above, the output of the algorithm is given below.

Tanti Lenuța, fusese cãsãtoritã cu un "nemernic" (cum ziceau mama și bunica mea), profesor pe undeva, într-o comunã peste dealuri, care a pãrãsit-o pentru altã femeie, iar mama mea a divorțat de tatăl meu când eu aveam numai un an. Deci …două femei fără bărbați, speriate de viitorul sumbru care se prevedea la orizontul roșu și

trei copii (ba chiar patru, dacã o punem la socotealã și pe verișoara mea Rodica Popescu) puși numai pe joacă și năzbâtii, în ciuda sărăciei de care nu eram conștienți - cam acesta era mediul în care trãiam și creșteam mari zi cu zi. Din când în când mama mea si cu tanti Lenuța își găseau timp să stea de vorbă, uneori ore întregi. Ce aveau să-și spunã? Curios din fire cum eram, trăgeam cu urechea. Vorbeau în șoaptã și nu prea

4. Diacritic Recovery

Another issue which is specific to text importing from PDF files is the difference in coding for diacritics. In the case of certain types of PDF files, diacritics are not represented in UTF-8 format, and are instead replaced with ASCII characters painted with a specific font in order to represent the correct glyph. Because of this, when we extract the base text from the PDF file, diacritics will be represented as non-literal ASCII symbols. An example of this is given below:

La Sfântul Evanghelist Ioan, în Prolog, Dumnezeu Tat\l [i Dumnezeu Fiul sunt numi]i simplu Dumnezeu [i Cuvântul (Logosul).Tat\l, prima persoan\ a Treimii, este Cel care are în sine plin\tateavie]ii. Plin\tatea fiin]ei dumnezeie[ti a Tat\lui se roste[te viu de c\treEl ca adev\r [i Cuvânt. „Tat\l este astfel Dumnezeu care roste[te,iar Fiul este Cel rostit. Fiul este ideea infinit\ a lui Dumnezeu, plin\tatea fiin]ial\, plin\tatea de valoare [i de ordine a lui Dumnezeu,[i anume ca fa]\ deschis\, descoperit\ de Logos. Acest Logos, acestal doilea El în Dumnezeu, S-a f\cut om în Hristos”

Doamna Cãlinescu avea douã surori. Virginica, tot profesoarã de ªtiinþele Naturale cãsãtoritã cu profesorul de fizico-matematici Bãlan-un om calm, domol, avea sã ne predea pentru o scurtã perioadã în timpul rãzboiului, ºi Rica, învãþãtoare în comuna Pãuºeºti Mãglaºi. O comunã în drum spre Olãneºti, la 14 kilometri de Râmnic. ªtiu pentru cã am fost pe acolo ºi cu aceastã ocazie doamna Rica mi-a povestit o întâmplare din copilãria lor de fete, întâmplare confirmatã de doamna Cãlinescu. Virginica nemaiputând confirma, fiind rãpusã de un cancer spre sfârºitul rãzboiului.

In the second example above, the letter “ă” is encoded as the symbol “ã”, the letter “ţ” as the symbol “þ”, etc. The replacements are unique, in the sense that once a translation is established (an association between a desired letter and a symbol), it will be used throughout the entire text. This means that an automatic replacement for each symbol will determine the correct text. A single translation table is not enough, however, as each individual document can have its own specific translations, which are not necessarily used in any other document. Because of this we have proposed a solution for automatically extracting translation tables for each document on the basis of a glossary of word forms for Romanian. We have also proposed an algorithm for replacing candidate symbols with the corresponding translation so that each replacement is indeed a valid one in the case of such symbols as “[” (which in one of the examples above is used to denote the letter “ş” but also the parenthesis symbol). Moreover, we have chosen to use a glossary of word forms as opposed to a table of frequencies for bigrams

and trigrams because, in some cases, some of the diacritics are correctly depicted (mostly in the case of those characters which can be found in ASCII).

The algorithm for extracting the translation tables, together with the algorithm for replacing the diacritic characters are given below. For creating the glossary of word form for Romanian we have used the ROMBAC corpus (Ion et al., 2012).

1. foreach file in input directory //create glossary
   a. for each word in file
      i. add lowercase word to HashSet // the search for diacritics will be performed on lowercase letters only, in order to reduce the search space
2. foreach word in inputFile //extraction of the translation table
   a. if(word has only one candidate) then find all possible matches for the candidate in the HashSet// for extracting the translation table, we prefer those candidate words which have exactly one diacritic symbol to reduce the search space. The candidate symbol is then replaced with all the diacritics of the Romanian language, and the newly created words are searched for in the glossary. For each wod that can be matched to a word in the glossary, the corresponding diacritic symbol is mapped into the translation table.
   b. while TranslationMap has multiple matches // this step aims to reduce the number of translations for each special character to exactly one
      i. find more words with candidate and search for matchs using the known solutions. If at least one unmatched word is found, eliminate the correspondence from the TranslationMap.
3. For all symbols in TranslationMap that have the same translation // litere mari şi mici vor fi rezolvate ca acelşi caracter. Pentru a face diferenţa dintre dintre majuscule i minuscule facem presupunerea că literele mici vor fi mai frecvente
   a. count the number of occurrences of symbolA and symbolB. The symbol with less occurrences is the capitalisation
4. Return translationMap for document
5. For each word in document
   a. Use translationMap to replace candidates
   b. If word worth replacements is in HashSet, replace in document
   c. Else do not replace

After running the algorithms described above for the last two examples above, the result is given below:

La Sfântul Evanghelist Ioan, în Prolog, Dumnezeu Tatăl şi Dumnezeu Fiul sunt numiţi simplu Dumnezeu şi Cuvântul (Logosul).Tatăl, prima persoană a Treimii, este Cel care are în sine plinătateavieţii. Plinătatea fiinţei dumnezeieşti a Tatălui se rosteşte viu de cătreEl ca adevăr şi Cuvânt. „Tatăl este astfel Dumnezeu care rosteşte,iar Fiul este Cel rostit. Fiul este ideea infinită a lui Dumnezeu, plinătatea

fiinţială, plinătatea de valoare şi de ordine a lui Dumnezeu,şi anume ca faţă deschisă, descoperită de Logos. Acest Logos, acestal doilea El în Dumnezeu, S-a făcut om în Hristos"

Doamna Călinescu avea două surori. Virginica, tot profesoară de Ştiinţele Naturale căsătorită cu profesorul de fizico-matematici Bălan-un om calm, domol, avea să ne predea pentru o scurtă perioadă în timpul războiului, şi Rica, învăţătoare în comuna Păuşeşti Măglaşi. O comună în drum spre Olăneşti, la 14 kilometri de Râmnic. Ştiu pentru că am fost pe acolo şi cu această ocazie doamna Rica mi-a povestit o întâmplare din copilăria lor de fete, întâmplare confirmată de doamna Călinescu. Virginica nemaiputând confirma, fiind răpusă de un cancer spre sfârşitul războiului.

Evaluation showed a rate of success of more than 97%, tested on 20 documents. In the future we intend to improve the running time of the algorithm by using previously extracted translation tables, which can reduce the search space by reducing the number of initial candidates offered for each symbol according to previous results.

5. Conclusions

In this paper we have described a method for automatically correcting Romanian text extracted from PDF documents which have been collected and processed within the CoRoLa corpus. The algorithms proposed are accurate, as shown in section 3, and novel, as, to our knowledge, the problem of recuperating diacritics from non-alphanumeric ASCII characters has not been solved before.

In the future, we intend to further test and improve the solutions proposed, especially in terms of time complexity, in order to manage the very large volume of texts that will have to be processed.

Acknowledgements

The research presented in this paper was founded by the Romanian Academy, Iasi Branch.

References

Barbu Mititelu, V., Irimia, E., Tufiş, D. (2014). CoRoLa – The Reference Corpus of Contemporary Romanian Language. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation - LREC*, Reykjavik, Iceland, 1235-1239.

Buneman, P., Davidson, S., Ferdandez, M., Suciu, D. (1997). Adding Structure to Unstructured Data. In *Proceedings of the 6th International Conference on Database Theory* (ICDT'97), 1997, 336-350.

Cai, D., Yu, S., Wen, J. R., Ma, W. Y. (2003). Extracting Content Structure for Web Pages based on Visual Representation. In *Proceeding APWeb03 Proceedings of the 5th Asia-Pacific web conference on Web technologies and applications*, 2003, 406-417.

Hofmann, K., Weerkamp, W. (2007). *Web Corpus Cleaning using content and Structure*.

Ion, R., Irimia, E., Ştefănescu, D., Tufiş. D. (2012). ROMBAC: The Romanian Balanced Annotated Corpus. In *Proceedings of the 8th LREC*, Istanbul, Turkey, 339-344.

Papakonstantinou, Y., Garcia-Molina, H., Widom, J. (1995). Object exchange across heterogeneous information sources. In *IEEE International conference on Data Engineering*, March 1995.

Quinlan, J. R. (1993). *Programs for Machine Learning*. Morgan Kaufmann.

Suciu, D. (1996). Query decomposition for unstructured query languages. In *VLDB*, September 1996.

USING ARGUMENTATION FOR IDENTIFYING NON-AMBIGUOUS INTERPRETATIONS OF NATURAL LANGUAGE

MATEI POPOVICI

POLITEHNICA University of Bucharest, Romania, matei.popovici@cs.pub.ro

Abstract

We introduce Argumentation Theory as a setting for identifying non-conflicting interpretations of natural language discourses. Following existing work on textual entailment, our solution relies on extracting entailments and contradictions from texts. We cast the former and latter into arguments and attacks from Argumentation Theory, respectively. Finally, we examine different types of semantics and select the preferred semantics as the appropriate characterisation which yields un-ambiguous interpretations of a discourse.

Key words — argumentation, natural language processing, entailment.

1. Introduction

Formal logic is the language of mathematical proofs: logical sentences are built using well-defined formation rules, and interpreted by non-ambiguous semantics as either true or false, in some particular model. Going beyond its original starting point, logic was established as the de-facto language of Artificial Intelligence: knowledge representation, planning, multi-agent systems and their verification all rely on fragments of First or Second-Order Logic.

In contrast, the semantics of natural language is variable and context-dependent, often equivocal and uncertain. However, humans exhibit the amazing ability to express and understand each other in natural language, despite its apparent hurdles.

The development of Pervasive Computing (Weiser, 1999) and Ambient Intelligence (Chong and Mastrogiovanni, 2011), has blended technology with our day-to-day life. In this setting, the languages of man and machine must find some common ground. This achievement would allow us to empower (software) agents with the ability to recognize and understand human intentions, based on natural language.

In this paper, we restrict our attention to a small subset of this problem: we look at texts describing possibly ambiguous or contradictory scenarios, and try to identify non-conflicting interpretations. We assume information from texts can be seen as a set of textual entailment (Dagan et al., 2010) and textual contradiction (de Marneffe et al., 2008) relationships. For instance, the text “John enters the shop in order to buy wine” textually entails “John buys wine” as well as “John is in the shop”. More generally, a statement C(conclusion) is entailed by P(premise), if a human reading P would infer C to be true (Dagan et al., 2010). In the rest of the paper, we restrict our attention to texts satisfying the above assumption.

Further on, we transform textual entailment into arguments and textual contradiction into attacks between arguments. Finally, we use Argumentation Theory (henceforth called AT) in order to reason about the possible interpretations of the original text. We believe AT is suitable for this task because: (i) it disregards the structure of the arguments and the process of deriving them; (ii) it offers grounded semantics (i.e. interpretations) to a possibly conflicting set of arguments.

Argumentation has been already been employed with Natural Language Processing in order to analyse disagreement in debates (Cabrio and Villata, 2012) as well as for disambiguating knowledge bases extracted from natural language (Wyner et al., 2012). Also, in (Palau and Moens, 2009), argumentation is used in order to explore the structure of natural language arguments. To the author's knowledge, Argumentation has yet to be used in identifying unambiguous interpretations of texts in unrestricted natural language. The contribution of this paper consists of a method for achieving this objective.

We would like to underline that many details regarding our proposal do not receive the attention they would otherwise rightfully deserve. For instance, when discussing argument and contradiction extraction, we leave out the actual methodology and refer to (Dagan et al., 2010) and (de Marneffe et al., 2008) for details. Thus, our paper should be understood as an "argument in favour of Argumentation" for the disambiguation of natural language texts.

The rest of the paper is structured as follows: in Section 2 we look at some existing body of work which allows for the extraction of entailments and contradictions from a discourse. In Section 3 we introduce the machinery of Argumentation Theory, and illustrate how the preferred semantics can be used to capture non-ambiguous text interpretations. Finally, in Section 4, we conclude.

2. From natural language to argumentation

We consider three basic situations which, in our view, capture the nature of ambiguous information found in natural language: (i) contradictions, (ii) texts which are subject to multiple interpretations, (iii) texts whose interpretation is unique, but is implicit in the text itself. We illustrate each of these situations by examples.

2.1. Extracting assertions

As an initial step, we consider the identification of assertions. An assertion contains (simplified) factual information related to the text. Therefore, summarisation can be applied in order to identify assertions. However, we opt for techniques such as (Duffield et al., 2010). According to (Duffield et al., 2010), traditional summarisation techniques uniformly remove background information which is considered not directly relevant. In doing so, useful pieces of information are lost. (Duffield et al., 2010) identifies the Presentational Relative Clause (PRC), as a phrase construction which more than often carries factual information, and which is removed during summarisation. For instance, in:

Bob and Alice are in the elevator, ***which arrives on the ground floor,*** *when Bob asks Alice a question.*

the emphasized fragment is a relative clause, which contains key information about the text. (Duffield et al., 2010) presents three properties which help identify PRC in texts. We defer the interested reader to the respective paper. Further on, we illustrate situations (i)-(iii) by examples.

Example 1 (*Contradiction*)

"*This sentence is false.*"

In Example 1, the sentence refutes itself. We consider the sentence (denoted S) as an assertion itself.

Example 2 (*multiple interpretations*)

"*John, would you buy a bottle of wine from the store?*

If they have fresh eggs, get at least six of them."

The example paraphrases the well-known anecdote whose subject is John, an engineer. Following the instructions, he asks if fresh eggs are available at the store. Following a positive answer he subsequently buys six bottles of wine.

One assertion, denoted H, is that "fresh eggs are at the store". Note that this assertion is not entailed from the text. Other assertions are: "John buys a bottle of wine", "John buys six bottles of wine", "John buys six eggs". We denote these as: W_1, W_6 and E, respectively.

Example 3 (*unique implicit interpretation*)

"*Alice and Bob are in the elevator.*

The elevator stops at the ground floor, and Alice asks Bob:

Are you coming down?"[1]

In Example 3, the possible interpretation in which Bob remains in the elevator in order to descend is ruled out by the implicit information provided by the text, namely that the elevator is on the ground floor. The possible assertions are "Bob does not come down from the elevator" (I), "Bob comes down the elevator" (O), "Bob descends" (D), "Bob does not descend" (ND), "The elevator is on the ground floor" (G).

2.2. Extracting arguments

Extracting arguments is a special case of argument mining in natural language (Palau and Moens, 2009). Argument mining aims at detecting arguments from unrestricted text, as well as their underlying structure and relations. One of the underlying challenges, as emphasized in (Palau and Moens, 2009), consists in choosing the appropriate representation for arguments. In papers pertaining to argumentation in Law such as (Prakken et al., 2008; Wyner, 2010), the structure of choice is a restricted form of First-Order Logic:

[1] We have opted for the term "coming down" instead of "getting off" in order to underline the apparent ambiguity. In the Romanian language, the translation of both expressions would coincide.

$\bigwedge_i premise_i \rightarrow conclusion$

where

$premise_i \equiv P(x_1,\ldots,x_n)$

and each x_i is seen as universally quantified over the domain of discourse.

We propose a rather simpler approach, similar to that from (Wyner et al., 2012), where arguments are seen as entailments. An entailment is a relation between two assertions α and β, denoted $\alpha \vdash \beta$, which denotes that β is true as a consequence of α being true. We shall henceforth call these entailments arguments. For instance, in Example 1, S is an argument in itself. We can view the argument as $S \vdash \neg S$.

In Example 1, we identify the following arguments: (i) "if the eggs are fresh, John buys one bottle of wine ($H \vdash W_1$), (ii) "if the eggs are fresh John buys six bottles of wine" ($H \vdash W_6$), (iii) "if the eggs are fresh, John buys six eggs" ($H \vdash E$).

Finally, in Example 2, the arguments are: "Bob stays in the elevator, hence he descends" ($I \vdash D$), "Bob comes out of the elevator hence he does not descend" ($O \vdash ND$), "The elevator is on the ground floor" ($\top \vdash G$).

2.3. Extracting contradictions (attacks)

A contradiction can be defined as a (directed) relation between two entailments of the following forms:

	First argument	Second argument
(a)	$\alpha \vdash \neg\beta$	$\beta \vdash \gamma$
(b)	$\alpha \vdash \beta$ where	$\beta \vdash \gamma$ $\neg\gamma \vdash \delta$
(c)	$\alpha \vdash \beta$	$\gamma \vdash \neg\beta$

In case (a), the premise of the latter entailment is defeated by the conclusion of the former. In (b), the conclusion of the former argument in turn entails γ, which defeats the premise of the latter argument. It is the case of an indirect defeat. In case (c), the entailments yield contradictory conclusions.

Identifying contradictory assertions of the form α and ¬α has already been tackled in paper such as (de Marneffe et al., 2008), and which can be employed in our setting.

When a contradiction between two arguments A and B has been identified, we say A attacks B. In Example 1, the argument $S \vdash \neg S$ naturally attacks itself, according to case (a). In Example 2, $H \vdash W_1$ and $H \vdash W_6$ mutually attack each other, since, $W_1 \vdash \neg W_6$ and $W_6 \vdash \neg W_1$: one cannot buy precisely one bottle of wine, as well as six bottles of wine, in the same time. We are in contradiction case (c). Finally, in Example 3, we have that arguments $I \vdash D$ and $O \vdash ND$ mutually attack each other, contradictions which can be seen as both of type (b) (since $I \vdash \neg O$: it cannot be that Bob descends as well as Bob comes out of the elevator) as well as of type (c) ($D \vdash \neg ND$ and $ND \vdash \neg D$: Bob cannot stay and go out of the elevator in the same time). Finally, $\top \vdash G$ attacks $I \vdash D$, which spawns from a contradiction of type (c).

3. Argumentation Theory

3.1. Preliminaries

The field of abstract argumentation theory spawns from that of non-monotonic logic. In formal logic, monotonicity refers to the fact that, by applying some rule for deriving knowledge, we cannot have less information about the world than we would have prior to rule application. Thus, in non-monotonic logic, one has to accommodate (new) facts which may be inconsistent with (and thus rule out) old ones.

AT, introduced with Dung's paper (1995), provides with an abstract framework for dealing with such inconsistencies. Starting from a domain description D=(R,K) where R is a set of possibly defensible rules, and K is a knowledge base, one can build an argumentation framework (to be discussed in detail in what follows), which precisely captures the conflictual relations from D. The strength of the approach resides in the fact that the actual representation of D (i.e. the chosen logical language), and the AT-related mechanisms for reasoning about conflicts, are completely independent.

The abstraction provided by AT has attracted its use in areas outside non-monotonic reasoning, ones which include applications for natural language.

3.2. Argumentation frameworks

Definition. An argumentation framework (Dung, 1995; Caminada and Gabbay, 2009) is a pair AF=(A,att) where A is a finite set of arguments and att⊆A×A is a an attack relation between arguments. We say argument a∈A attacks b∈B iff (a,b)∈att.
The argumentation frameworks corresponding to Examples 1, 2 and 3 are shown in Figures 1, 2 and 3, respectively. Each node represents an argument, while each directed arrow represents a directed attack between arguments.

In the rest of the paper, we assume AF=(A,att) is an argumentation framework.

Extension semantics. An extension-based semantics Swith A a subset $S(A)\subseteq 2^A$ of arguments from AF. Intuitively, an extension-based semantics gives a possible interpretation of a discourse.
For instance, in Figure 1, {⊤⊢G,I⊢D} is an extension-based semantics. It corresponds to the interpretation of Example 3 in which Bob remains in the elevator and descends. Obviously, this interpretation is erroneous, as is intuitive from the attack from ⊤⊢G to T⊢D.

Admissibility and conflict-freeness. For the above-mentioned reason, we introduce the notion of admissibility and conflict-freeness (Caminada and Gabbay, 2009). First, for any argument a∈A we define the set ⟩a⟨={b|(b,a)∈att} of arguments which attack a.
For instance, in Example 3, the set of arguments which attack I⊢D, namely ⟩I⊢D⟨, is {⊤⊢G,O⊢ND}.

Let S⊆A be a set of arguments and a∈A be an argument. S defends a, iff, for any b∈⟩a⟨, there exists some c∈S such that c attacks b. For instance, in Figure 3, S={⊤⊢G} defends

$O \vdash ND$. In other words, the information "*the elevator is on the ground*" defends the argument "*Bob comes out of the elevator hence he does not descend*".

The function $F:2^A \rightarrow 2^A$ such that $F(X)=\{a \in A | X \text{ defends } a\}$ is called the characteristic function of AF. Intuitively, F maps any set X to a set of arguments which X defends.

A set S of arguments is called conflict-free iff for all $a,b \in S$ it is not the case that a attacks b or b attacks a. Furthermore, S is called admissible, iff S is conflict-free and $S \subseteq F(S)$.

Intuitively, an admissible set S is any conflict-free set which defends all its arguments. In our setting, admissible sets correspond precisely to those un-ambiguous interpretations of the original text.

Figure 1: The argumentation framework AF_1 resulting from Example 1

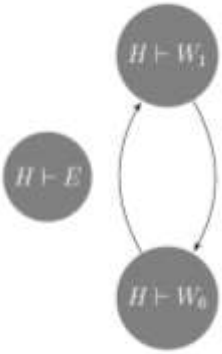


Figure 2: The argumentation framework resulting from Example 2

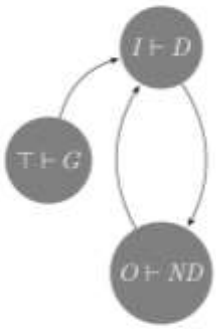


Figure 3: The argumentation framework resulting from Example 3

Returning to AF_1 from Figure 1, we note that the unique admissible set is ∅. In AF_2 (Figure 2), we have 5 admissible sets. We enumerate only two: {H⊢E,H⊢W1} and {H⊢E,H⊢W8}. The two sets correspond to the two possible interpretations of Example 2 in which John either buys one or six bottles of wine.

Finally, in AF_3, the admissible sets are: ∅, {⊤⊢G} and {⊤⊢G,O⊢ND}.

3.3. Selecting appropriate semantics

Note that all admissible sets correspond to components of the same un-ambiguous interpretation of Example 3. Also, it can be easily seen that not all admissible sets characterize complete interpretations of a discourse. In order to capture a unique (and un-ambiguous) semantics, such a characterisation is necessary.

Well-known semantics

In the literature (Caminada and Gabbay, 2009), different possible semantics are defined. We review only a few:

- *grounded semantics* – they capture the minimal (non-empty) admissible sets, w.r.t. set inclusion. Thus, the grounded semantics selects the smallest number of arguments which contain no interior attacks, and defend themselves from all outside attacks. The intuition corresponds to the "the position which is least questionable" (Caminada and Gabbay, 2009);
- *preferred semantics* – they capture maximal admissible sets, and take a more optimistic view on the discourse; in this setting, the greatest number of admissible arguments is taken;
- *stable semantics* – they follow the principle "you are either with us or against us" and rule out non-attacking arguments. This can also be interpreted as selecting useful arguments only. Thus, an argument is useful if it provides some defence for the other arguments in the semantics: hence, it attacks an argument outside the semantics.

A suitable semantics for NLP. In what follows, we opt for the preferred semantics, as an appropriate characterisation of non-ambiguous discourse interpretations. The choice is governed by the intuition that our semantics should provide an un-ambiguous interpretation of a discourse. Hence, there is no reason to exclude an argument from a semantics, as long as it does not attack other members of the semantics, thus spoiling conflict-freeness.
We designate as preferred extension(s), the maximal admissible set(s) of an argumentation framework (w.r.t. set inclusion).

Revising our examples, we note that the preferred extension of AF_1 is $\emptyset$, which confirms the observation that Example 1 is contradictory. The preferred extensions of AF_2 are {H⊢E,H⊢W1} and {H⊢E,H⊢W8} and finally, the preferred extension of AF_3 is {⊤⊢G,O⊢ND}.

Computing the preferred extension. According to (Caminada, 2006; Caminada and Gabbay, 2009) we have the following result, which yields a computational method for identifying preferred semantics: a set S is a preferred semantics iff it is a maximal fixed-point of the function F. We recall that X is a fixed-point of F iff F(X)=X. Also, if F is monotone (X⊆Y implies F(X)⊆F(Y)) then, there exists an unique maximal fixed-point which can be computed as: $\bigcup_n F^n$, due (Tarski, 1955), where $F^n=F(F^{n-1})$ and $F^0=\emptyset$.
Thus, it is sufficient to compute the maximal fixed-points of F in order to identify preferred extensions. We also note that fix-point computation can be computed in time O(|A|), where A is the set of arguments of the underlying AF.

3.4. Applications

We envision a wide range of applications which are naturally accommodated by our approach: finding unambiguous interpretations in official documents such as codes of law, as well as in text from the media – a line of research which is also pursued in (Prakken et al., 2008; Wyner, 2010; Wyner et al., 2012; Cabrio and Villata, 2012); facilitating the interaction between humans and software. In order to achieve the latter, multiple preferred extensions can be ruled out by adding new arguments as well as attacks. In Example 3, we can envision John as a software agent able to compute preferred extensions. When more than one extension is identified, the agent is able to formulate additional questions, whose responses can be incorporated into the discourse as new arguments and attacks. In doing so, the agent is able to isolate a single preferred extension, and thus precisely identify the intentions of the speaker.

4. Conclusion and future work

We have illustrated how Argumentation Theory can be employed as a methodology for extracting un-ambiguous information from natural language text. Our work should only be understood as a preliminary road-map. The essential ingredients of our approach, (i) the identification of entailments and (ii) contradictions from text, have been explored in the literature. It remains to be seen to what extent these methods can accommodate our endeavour.

Annotated text (where assertions, arguments and contradictions are highlighted manually) can be used as a starting point. Once the machinery of AT is put to practice, automated means for (i) and (ii) should be developed, and different languages for representing more structure in arguments can be deployed.

As seen by our examples, the interpretation of each text can only be achieved in the presence of an ontology (which must specify, for instance, that buying both one and six bottles of wine at the same time is impossible). Thus, ontologies may be a key for any such implementations.

We also note that, apart from certain extraction techniques specific to (i) and (ii), our method is language independent, and should be easily applicable to the Romanian language.

Another important issue which deserves a future discussion is the identification of the minimal set of necessary set of arguments and/or attacks which ensure the existence of a unique argumentation framework. To our knowledge, this topic has not yet been addressed in Argumentation Theory.

References

Baroni, P., Caminada, M., Giacomin, M. (2011). Review: An Introduction to Argumentation Semantics. *Knowl. Eng. Rev.*, 26:4, 365-410.

Bos, J. (2009). Applying automated deduction to natural language understanding. *Journal of Applied Logic*, 7:1, 100-112.

Cabrio, E., & Villata, S. (2012). Natural Language Arguments: A Combined Approach. In L. D. Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, și alții (Ed.), *ECAI. 242*, IOS Press, 205-210.

Caminada, M. (2006). On the Issue of Reinstatement in Argumentation. *Proceedings of the 10th European Conference on Logics in Artificial Intelligence*. Berlin, Heidelberg: Springer-Verlag, 111-123.

Caminada, M., Gabbay, D. (2009). A Logical Account of Formal Argumentation. *Studia Logica*, 93(2-3), 109-145.

Chong, N., & Mastrogiovanni, F. (2011). *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspective.* Information Science Reference.

Dagan, I., Dolan, B., Magnini, B., Roth, D. (2010). Recognizing textual entailment: Rational, evaluation and approaches – Erratum. *Natural Language Engineering*, 16, 105-105.

de Marneffe, M. C., Rafferty, A. N., Manning, C. D. (2008). Finding Contradictions in Text. In K. McKeown, J. D. Moore, S. Teufel, J. Allan, & S. Furui (Ed.), *ACL*, The Association for Computer Linguistics, 1039-1047.

Duffield, C. J., Hwang, J. D., Michaelis, L. A. (2010). Identifying Assertions in Text and Discourse: The Presentational Relative Clause Construction. *Proceedings of*

the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 17-24.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:2, 321-357.

Palau, R. M., Moens, M. F. (2009). Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. New York, NY, USA: ACM, 98-107.

Prakken, H., Issn, W. C., Prakken, H. (2008). Formalising ordinary legal disputes: a case study. *Artificial Intelligence and Law*, 333-359.

Tarski, A. (1955). A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics* 5 (1955), no. 2, 285-309.

Weiser, M. (1999). The Computer for the 21st Century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3:3, 3-11.

Wyner, A. A. F. (2010). Approaches to Text Mining Arguments from Legal Cases. *Semantic Processing of Legal Texts*, 60-79.

Wyner, A., van Engers, T. (2010). Web-based mass argumentation in natural language. *Proceedings of Knowledge Engineering and Knowledge Management.*

Wyner, A., Schneider, J., Atkinson, K., Bench-Capon, T. J. (2012). Semi-Automated Argumentative Analysis of Online Product Reviews. In B. Verheij, S. Szeider, & S. Woltran (Ed.), *COMMA. 245*, IOS Press, 43-50.

A LANGUAGE INDEPENDENT NAMED ENTITY RECOGNITION SYSTEM

DANIELA GÎFU[1, 2], GABRIELA VASILACHE[1]

[1]*Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași*
[2]*Center for Advanced Research in Applied Informatics, University of Craiova*

{daniela.gifu, gabriela.vasilache}@info.uaic.ro

Abstract

The paper presents a hybrid language independent Name Entity Recognition system consisting of two parts: a handcrafted component and an automatic one. Extraction of named entities from the text is an important operation in many natural language processing applications like information extraction, question answering, machine translation, etc. In order to develop the automatic part of the system, namely semi-supervised recognition of entities, we created a gold corpus for the Romanian language, consisting of different textual genres like journalistic, literary, administrative, and religious. This study demonstrates statistical significance in terms of precision over the previous systems.

Key words — information extraction, entities categorisation, natural language processing, gold corpus, statistics.

1. Introduction

Name Entity Recognition (NER) is a special information extraction task that aims to recognise the proper names and assign semantic markers from a list of categories established a priori (persons, organisations, locations, etc.). As specified in the abstract, we used different types of texts (journalistic, literary, religious and administrative), as resources for developing the gazetteers and for defining other categories.

The paper is structured in 5 sections. After a brief introduction about this topic, in section 2 the most important written papers on NER are retained. Section 3 presents a methodology in order to develop NER for Romanian and English, and section 4 shows statistical results for the name entities detection. In the end, section 5 describes a few conclusions and discussion for our future work.

2. State of the art

Starting with the first edition of *The Message Understanding Conference* (MUC) in 1990s, a new concept, NER, was introduced. The attention was focused on extracting structured information about the activities of companies in unstructured text (such as newspaper articles). Five years later, at the 6th edition of MUC, the term *named entity* was introduced (Grishman and Sundheim, 1996). In this context *Named Entity Recognition and Classification* (NERC) was defined and classified as a task of *Information Extraction* (IE). With other words this computational area proposes to identify and classify the words or word groups that appear in a text and signify proper

names (Nadeau and Sekine, 2007). In many studies we can observe the impact of NERC using textual genres and domain.

NER has become a very important topic for many other subfields of natural language processing (Sekine et al., 1998; Borthwick et al., 1998; Masayuki and Matsumoto, 2003). Actually, NER supports a few development strategies that consist in the algorithms based on a collection of grammar rules, like those of Nadeau and Sekine (2007) and Iftene et al. (2011). The interest was to identify all possible situations, without creating ambiguity, to obtain higher performances. Tasks as a personal name disambiguation (Mann and Yarowski, 2003), named entity translation (Fung, 1995; Huang, 2005), acronym identification (Nadeau and Turney, 2005) have become a part in our developed NER system.

3. The methodology

We describe the external resources which make up the basis of NER development (the corpus structured from journalistic, literature, religious and administrative texts), the pre-processing tools (POS-tagger[1] and PALinkA annotator[2]) and statistical results, in terms of precisions of automatic entities recognizer.

3.1. External resources

As already mentioned above, the entity recognition system that we have developed is a hybrid one, because it has a component that is based on knowledge, but also an automatic one used for disambiguation.

The knowledge component consists of a handcrafted gazetteer, one for the Romanian language (list with entities names from four different texts genres[3]) and another one for the English language (list with entities names from journalistic texts). It contains a collection of lists with entities for each of the predefined categorises: PERSON, GEOGRAPHICAL, ORGANISATION, URL.

We present the lists for each category included:

- class PERSON – person name (first name and surname);
- class GEOGRAPHICAL – name of continents, countries, cities, towns, landforms;
- class ORGANISATION – name of associations, institutions, political parties, companies, enterprises;
- class URL – web addresses.

Each of these classes, in addition to the lists specified, contains one manually created list of *knowledge* which contains representative entities falling within the class, but with names that are not part of that representative vocabulary: therefore, because there are three predefined classes (PERSON, GEOGRAPHICAL, ORGANISATION), we will

[1] http://nlptools.infoiasi.ro/WebPosRo/

[2] *Perspicuous and Adjustable Links Annotator* (PALinkA), developed by Constantin Orăsan, is a tool successfully used in several projects in NLP-Group@UAIC-FII.

[3] Journalistic Corpus – 100 text files (approx. 40 sentences for each file); Literary Corpus – *Quo Vadis* Novel (7281 sentences); Administrative Corpus (approx. 1035 sentences); Religious Corpus – (33830 sentences).

get three lists of knowledge: *person_knowledge*, *geographical_knowledge*, *organisation_knowledge*.

For example:

- In Romanian, we can consider the following entities in knowledge:
  - *person_knowledge* – Al Pacino;
  - *geographical_knowledge* – Kotte Sri Jayewardanapura;
  - *organisation_knowledge* – The Telegraph.
- In English, we can consider the following entities in knowledge:
  - *person_knowledge* – Wojtek Lazarek;
  - *geographical_knowledge* – Spitsbergen;
  - *organisation_knowledge* – Belangenvereniging Tankstations (BETA).

Depending on the language we refer to, the number of knowledge lists entries oscillates; in case of Romanian, the largest number of entries is found in *geographical_language* and the lowest is in *organisation_knowledge*. In English, the situation is changed: *geographical_knowledge* represents the list with the lowest number of entries and the largest number is found in *person_knowledge*.

Moreover, all the lists that make up the gazetteer for the Romanian language were built in two ways: with Romanian diacritics (*comma*) and Turkish diacritics (*cedilla*); this is because, during the construction of the corpus, we found both forms of diacritics in texts.

3.2. Resources used in learning process

In this stage, from the corpus mentioned above for the Romanian language, we used only two types of texts: journalistic and literary one. After annotation with PALinkA (Orăsan, 2003), we got the journalistic corpus and the literary corpus. The first one was manually created and is composed of 100 text files with information extracted from publications with a general circulation from different areas (politics, sports, culture, etc.), and the literary corpus consists of the novel "Quo Vadis" written by Henryk Sienkiewicz. The other part of the corpus (administrative and religious) will be used in the evaluation process.

For the English language, the corpus was also manually created and is composed of text files with information found on the internet, extracted from publications with a general circulation from different areas (politics, sports, culture, etc.).

For both languages, in order to create the Gold corpus, we took text files from the initial corpora, (for Romanian we took sixty one text files from the journalistic corpus and for English, we took forty five text files from the news corpus) and we used two pre-processing tools: POS-tagger (Simionescu, 2011) and PALinkA annotator. Because of the fact that PALinkA annotator receives as input only *Extensible Markup Language* (XML) files, the collection of text files chosen from corpus passed first through the process of parsing, made by the POS-tagger. Thus, we obtained XML files with its specified parts of speech and with representative attributes.

Moreover, we moved on to the annotation step. Each XML file from the corpus was annotated to the entity level. An entity is any proper name identified in the text which

can be classified into one of the categories from the predefined set. In this way we obtained the gold corpus.

3.3. The Multilingual Named Entity Recognizer

The entities recognition system that we developed is language independent; therefore, works on both languages: Romanian and English. The program receives as input an XML file resulting from the processing of text with the Part-Of-Speech tagger and the output also returns an XML file, but contains specifications referring to identified entities and the category type in which they were classified.

Entities recognition process has two phases: initialisation and processing. As already noted, because it is an independent language system, in the initialisation procedure it loads the resources: lists of entities in both languages (Romanian gazetteer, compound of lists with entities names from four different texts genres and English gazetteer, compound of lists with entities names from journalistic texts).

After the initialisation step it goes forward to the parsing process. Thus, each paragraph is divided into sentences and for each sentence it creates a list made up of all its component words. From the resulting list, sub-lists are formed by combining words, bowing to the maximum length, as follows:

- we have the word list - Word - containing words $<w_1, w_2, w_3, \ldots, w_n>$
- forming sub-lists:

$<w_1, w_2, w_3, \ldots, w_n>$
$<w_1, w_2, w_3, \ldots, w_{n-1}>$
$<w_1, w_2, w_3, \ldots, w_{n-2}>$
$<w_1, w_2, w_3>$
$<w_1, w_2>$
$<w_1>$
$<w_2, w_3, \ldots, w_n>$
$<w_2, w_3, \ldots, w_{n-1}>$
$<w_2, w_3, \ldots, w_{n-2}>$
$<w_2, w_3>$
$<w_2>$
$<w_3, \ldots, w_n>$
$<w_{n-2}, w_{n-1}, w_n>$
$<w_{n-2}, w_{n-1}>$
$<w_{n-2}>$
$<w_{n-1}, w_n>$
$<w_n>$

Now, each resulted sub-list is checked if it represents or not an entity; more specifically, if it is found or not in the resources lists. As entities are identified, they are retained in a new list – Entities – which is initially empty. Note that before each entity is checked, it is first checked if that sub-list is already included in the list of identified entities; if so, it is ignored and it goes forward to check the next sub-lists.

The validation of a sub-lists if it is entity or not and setting type step are limited to operations of decision. Therefore, each score is calculated for each category of entities (before this calculation, all scores are initialized to zero and increase with predefined values set according to relevance in the text of each category): if sub-list is found in the ORGANISATION resources category, the score for this category increases, if it is

found in GEOGRAPHICAL resources category, then the score for that category will increase, and if a sub-list is found in the lists with name of person representative for the class PERSON, then the score will increase for the PERSON category.

So, following the chain of checks, the minimum condition that a sub-list being an entity is at least one if the score is not zero, the starting value from which we started; otherwise, we conclude that the sub-list is not an entity. If the case is positive, at least one of the scores is different from zero, we consider the sub-list as an entity and the last thing is to set the type. For the disambiguation phase, based on the corpus, we automatically created a list of trigger words and new scores are calculated.

The last step is to compare the scores obtained during the searching of entities, and the entity type is given by the category with the highest score. Like we said before, all these specifications will be added in the XML file from the output.

To be user friendly, the program is available as a web application, with an interface as shown in the following picture:

Figure 1: NER Web Application

As it can be seen easily, the user has available more features such as:

- can set the language of the text on which the recognition of entities is to be made;
- may choose to manually type raw text or directly upload an xml file;
- can begin the process of recognizing entities by pressing the Recognize Entities button.

An overview of the result can be seen below[4]:

- as we can observe (Figure 2), each category of entities is represented by a colour; as a result, each entity identified in the text will be put into a box with coloured background corresponding to the class in which was classified;

[4] Each sentence from the text is framed individually; this emphasizes the process of splitting each paragraph in sentences.

Figure 2: Work Session NER

- because the text is pre-processed by a POS-tagger, when clicking on any word in the text, at the top of the frame, you will see its morphological analysis (Figure 3).

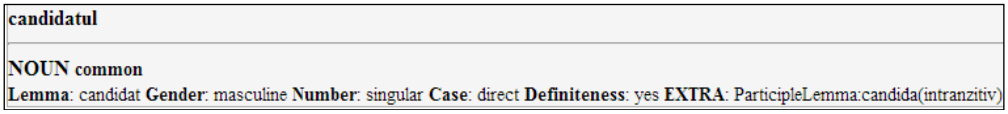


Figure 3: POS-tagger part of NER

4. Statistical evaluation

In evaluating the performance of a name entity recognition system, it is recommended to calculate the *precision* and *recall* parameters (Bikel et al., 1997). Because at the moment we work only with two types of corpora[5] (the journalistic and the literary one) from a total of four (as mentioned above), for this stage of the research, we will consider only the first parameter.

Here we present the results for the precision parameter for both languages (see formula 1):

$$P = \frac{\#correctly\_identified\_Es}{\#automatically\_annotated\_Es} \qquad (1)$$

where Es = entities.

As shown in Table 1, the precision results for the automatic detection of named entities in Romanian corpus is high. The reason is clear: we have a long process in manually annotation by specialists in linguistics and the corpus passed through three supervision

[5] http://nlptools.infoiasi.ro/LanguageIndependentNamedEntityRecognizer/

and related corrections ("Quo Vadis") and the journalistic corpus was annotated by the second author and supervised by the first one.

Table 1: Statistical results for the detection named entities for Romanian corpus

Text genres	Number of sentences	Number of entities	Precision
Journalistic	3952	2354	98.70%
Literacy	7281	4337	99.20%

Table 2: Statistical results for the detection named entities for English corpus

Text genres	Number of sentences	Number of entities	Precision
Journalistic	7986	8290	47.00%

As shown in Table 2, the precision result for the automatic detection of named entities in the English corpus is very weak; the fact that the corpus is not a proper one (only a few articles on the social and political issues) and the gazetteer is insignificant as compared with the Romanian one.

After we finish annotating the administrative part of the corpus and the religious one, we will have complex corpora, as desired, compound of four text genres: journalistic, literary, religious and administrative. When this step is passed, we will be able to compare the result of our system with others and can state the performance of our natural language processing tool.

5. Conclusions and future work

This paper described a system for named entity recognition from Romanian and English text. The results obtained from Romanian texts (journalistic and literary) were much better than those from English texts, but we will go further to improve the system and transform it in a very important Natural Language Processing tool.

Therefore, future development will involve implementing rules for recognition of entities classified in NORMATIVE ACTS (organic laws, governmental laws, ordinance, city council decisions, etc.) and TIMEX (temporal expressions). Identification of time expression that involve also creating rules with GGS tool. Part of our corpus (administrative and religious), used only for testing the system, puts in evidence entities which cannot be assigned to one of the general three categories (person, location and organisation). Furthermore, we intend to implement these categories in our system.

In order to get a better score, we want to add new semantic rules for a higher level of disambiguation.

Our highest goal is to combine this named entity system with the anaphora resolution system; in this way no entity will be lost and the score will increase remarkably.

Acknowledgements

In order to perform this research the first author received financial support from the Erasmus Mundus Action 2 EMERGE Project (2011 – 2576 / 001 – 001 - EMA2). Both authors are also grateful to all colleagues from NLP-Group@UAIC-FII that have authored the natural language processing tools used in this research.

References

Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R. (1997). Nymble: a High-Performance Learning Name-finder. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 1997.

Borthwick, A., Sterling, J., Agichtein, E., Grishman, R. (1998). Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. In *Proceedings of the 6th Workshop on Very Large Corpora*.

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of the 33rd Association for Computational Linguistics*.

Grishman, R., Sundheim, B. (1996). Message understanding conference- 6: A brief history. In *Proceedings of COLING*.

Iftene, A., Trandabăţ, D., Toader, M., Corîci, M. (2011). Named Entity Recognition for Romanian. In *Proceedings of the 3th Conference on Knowledge Engineering: Principles and Techniques Conference* (KEPT2011), 2, 19-24.

Masayuki, A., Matsumoto, Y. (2003). Japanese Named Entity Extraction with Redundant Morphological Analysis. In *Proceedings of the Human Language Technology conference – North American chapter of the Association for Computational Linguistic*.

Mann, Gideon S., Yarowsky, D. (2003). Unsupervised Personal Name Disambiguation. *Proceedings of the 9th Conference on Computational Natural Language Learning*.

Nadeau, D., Turney, P. A. (2005). Supervised Learning Approach to Acronym Identification. *Proceedings of the 18th Canadian Conference on Artificial Intelligence*.

Nadeau, D., Sekine, S. A. (2007). Survey of Named Entity Recognition and Classification. *Linguiticea Investigationes*, vol. 30, n.1, 3-26.

Orăşanu, C. (2003). PALinkA: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, ACL'03*.

Sekine, S., Grishman, R., Shinnou, H. (1998). A Decision Tree Method for Finding and Classifying Names in Japanese Texts. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

Simionescu, R. (2011). *POS-tagger hibrid*, dissertation, "Alexandru Ioan Cuza" University of Iaşi.

MAPPINGBOOKS: LINGUISTIC SUPPORT FOR GEOGRAPHICAL NAVIGATION IN BOOKS

DAN CRISTEA[1,2], IONUȚ CRISTIAN PISTOL[1]

[1] *"Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science*
[2] *Institute for Computer Science, Romanian Academy – the Iași branch*
{dcristea, ipistol}@info.uaic.ro

Abstract

This paper describes a way in which Natural Language Processing (NLP) can be employed to provide support for real-life geographical navigation guided by referenced locations and other entities found in texts. It is shown how NLP techniques, such as name entity recognition, anaphora resolution and entity linking, supported by digital geographical resources, can provide localisation and cross-references of places, trajectories and other types of relations detected in texts. The architecture and functionality of a prototype application is briefly presented.

Key words — data mining, geographical references, mobile application.

1. Introduction

In this paper we describe the *MappingBooks* project[1], intended to build a technology that complements a book content with information acquired from external sources. The goal is to develop a new type of electronic product with a high impact in education and tourism. The main users envisioned are high school pupils learning the geography of Romania and tourists visiting Romania. The technology combines text analytics, web cartography and mapping, augmented reality techniques and ambient intelligence/ubiquitous computing. Name entities mentioned in school manuals and touristic guides are localised on hypermaps and put in correlation with the reader's location and related data. If the name entities refer to toponyms, they are supplemented with diagrams or graphical materials. For example, if the school book contains a mention of the *Ceahlău Mountain* and the young reader happens to be in its proximity, not only that a localisation of her/himself will be signalled on an electronic map that includes the mountain, but information fetched from external sources, such as background layers, as discrete raster and vector data, and additional multimedia information, mainly represented as pop-ups and hyperlinks, all depending on the textual context in which the toponym occurs, will be displayed on the reader's mobile screen. As for the tourist guide, up-to-date information regarding the planning of a trip, accommodation possibilities on the mountain, etc., in general data not directly available in the guide, will be displayed near the map.

[1] First developed as term projects by students in the Artificial Intelligence course at the Faculty of Computer Science of the „Alexandru Ioan Cuza" University of Iași (UAIC).

As such, a *MappedBook* is a book connected with locations in the real and virtual world, which, to the discretion of the user, could be sensitive to the instantaneous location of the reader, as seized by her/his mobile, and that signals, at appropriate moments, events in the real or virtual world related to the location of the toponyms and other entity names the book contains.

After a short presentation of the current related technologies, we make a general overview of the system and state some conclusions.

2. State of the art and beyond

MappingBooks will apply, enhance and invent new e-content and language processing technologies, capable to recognise and annotate different types of name entities, to link these mentions onto the real and virtual world, and to synchronise, by applying ambient intelligence and ubiquitous computing techniques, the mentions in the book with the actual position of the user and the images captured by the mobile user's device.

In order to do that, basic text analytics include, minimally: segmentation at sentence level, tokenisation, part-of-speech tagging, lemmatisation, noun-phrase chunking, named entity recognition and anaphora resolution. These are well known techniques that have reached technological maturity for the language under investigation, which is Romanian. In past projects like CLARIN[2], METANET4U[3] and ATLAS[4] modules and processing chains operating on the lines described have been realised. In ATLAS, for instance, individual tools have been combined (Anechitei et al., 2013) in processing chains by using CAS objects of U-Compare, a UIMA dialect (Ferruci and Lally, 2004). A previous research developed at UAIC (Cărăușu, 2011; Ciucanu, 2011) in correlation with a postdoctoral project (Dumbravă, 2010-2013) used patterns of pragmatic knowledge inferred from a manually annotated Romanian text[5] to transform a textual description of a journey onto trajectories on Google maps.

Linking book mentions onto the real and the virtual world involves a very good name entity recogniser for Romanian. Some progress has been made in this respect, as for instance the ANNIE-GATE module[6], the resources reported by Dumitrescu and Barbu Mititelu (2013) and the recent work of Potolincă (2014) who acknowledges the acquisition of a gazetteer containing 103,601 names and a collection of 400 hand-written patterns. For English, as part of the MUC-7 competition, the system LTG (Mikheev et al., 1998) reports an F-measure of 93.4%, when human performance does not overpass 97%.

Important in *MappingBooks* is the discovery on the web of data linked to books' mentions. For this, entity linking techniques are applied. A rich collection of methods is summarised in the exceptional survey of Năstase et al. (2013). The project intends to develop ad-hoc one-document and cross-document linking techniques, by exploiting the

[2] http://www.clarin.eu/external/
[3] http://metanet4u.eu/
[4] http://www.atlasproject.eu/
[5] "Iter in Chinam" by Milescu Spătarul – about 1675-1678.
[6] A Nearly-New Information Extraction system, at: http://gate.ac.uk/sale/tao/splitch6.html

state-of-the art anaphora resolution techniques (Bagga and Baldwin, 1999; Saggion, 2007; Singh et al., 2011), mainly spotting persons and locations. The UAIC's RARE system, using a mixed approach which combines symbolic rules with learning techniques, has been recently proved to work well for Romanian and other languages (Anechitei et al., 2013). Moreover, dynamic links will be created between mentions in the processed books and their reflexions in the virtual space. The project will develop an integrated technology which will put in the same melting pot known and new techniques to create multi-dimensional mash-ups that combine textual, geographical and, to some extent, temporal information in order to present them adequately to a reader, intermediated by a specially-designed human-computer interface (HCI). The links should be sensible to the context of the book, the moment of time the access is initiated by the reader, the momentary location of the reader and her/his orientation in space.

3. General architecture of the system

3.1. Text pre-processing

Figure 1 displays a diagram of the interconnected modules of the system. The Text Analytics module (TA) receives as input the original text of the book in multiple document formats, including PDF and DOC. A process, known as boiler-plating, removes images, captions, footnotes, headers, footers and page numbers, i.e. any elements breaking the cursive unfolding of the text. However, the original layout of pages is also preserved and the page coordinates of all linguistic elements (words) are recorded in order to allow highlighting of specific word sequences that mark entities, spatial relations, or paths on the user interface, and to know where to place balloons or flying windows evidencing linking information.

Once cleaned from formatting, the content text is passed through some of the UAIC's web services (Simionescu, 2011) that perform: sentence segmentation, tokenisation, POS-tagging, lemmatisation and noun phrase chunking, including also the identification of head words of noun phrases.

3.2. Entities and relations

The output, an XML file, is passed to the Name Entity Recognition module (NER), which is responsible for marking names and their types. Presently, we recognise 8 categories of names: *people*, *institutions/buildings*, *streets/areas*, *cities*, *districts*, *countries/regions*, *mountains*, *waterways*. If an ambiguity occurs (the same name identifies two cities in different countries or a city and a river, for example), the module asks help from the geographical information module (GEO), which returns the most probable category, usually that one which is placed on the map in a closer geographic vicinity to other entities recently mentioned in the same document. The NER module incorporates also an Anaphora Resolution algorithm, in order to link different mentions of the same entity (through pronouns or other nominal constructions). As such, all references to the same entity in the book are chained onto a unique semantic representation.

Marked entities are then crawled on the web by the Entity Crawler module (EC), in order to establish external links. Sources of external information are Wikipedia, museum sites, whether reports, etc. The attached data varies according to the type of each entity and the context it is mentioned. For instance, for a *country* type entity, area, population, official language, national feats and other information could be fetched in. For an *institution* type as is a museum – visiting hours, entry fees, exhibitions currently hosted by the museum, etc.

The last step of the linguistic processing chain is realised by the Relations Detection module (RD), which is responsible for putting in evidence semantic relations mentioned in the text. To exemplify, the set of patterns (written as regular expressions, as used by the RD module) below detect *built-by* relations, that link *institution/building* type of entities with entities of type *person*.

```
@built-by
#Ent_[1]+[\s\d\W\p{L}]+a fost refacută[\s\d\W\p{L}]+#Ent_[0]+
#Ent_[1]+[\s\d\W\p{L}]+a fost terminată[\s\d\W\p{L}]+#Ent_[0]+
#Ent_[1]+[\s\d\W\p{L}]+este creaţia lui[\s\d\W\p{L}]+#Ent_[0]+
#Ent_[1]+[\s\d\W\p{L}]+realizat de[\s\d\W\p{L}]+#Ent_[0]+
#Ent_[1]+[\s\d\W\p{L}]+a fost construit[\s\d\W\p{L}]+#Ent_[0]+
```

Other relations are descriptions of trajectories. For instance, the ones below (written as XML patterns, as used by the geographical relations detection module) links two location entities as being near to each other.

```
@near
<word>in</word><word>apropiere</word><word>de</word><loc/><word>se</wo
rd><word>află</word><loc/>
<loc/><word>se</word><word>invecineaza</word><word>direct</word><word>
cu<loc/><loc/>
<word>din</word><loc/><word>*</word><word>spre</word><loc/>
<word>pe</word><loc/><word>spre</word><loc/>
```

All patterns described above are built from extracted examples, by observing common types of relevant text fragments. However, in the project we envisage to build a corpus of annotated examples out of which an automatic learning system will infer more diversified patterns. At the end of the described chain, a heavily annotated XML file is accumulated. Then, the entity names, as identified by the NER component, are enriched with information of geographical nature.

3.3. Adding geographical data

The Geography module (GEO) uses free databases (Google GeoCoding API[7], GeoNames[8]) to complement the information with geographical layers, as is the actual

[7] https://developers.google.com/maps/documentation/geocoding/

positioning of the entities found in the document on real maps (provided by GoogleMaps). Real-world geographical relations (such as a city being part of a country) are also determined by the GEO module. Let's notice that, very often, this kind of semantic relations are not directly specified in the original text. Sometimes GEO also realises a disambiguation step, as described before. Besides these functionalities, GEO also accumulates a JSON database of all entities found in the document (institutions, hotels, various landmarks, etc.). This database will be used to display important locations near the current position of the user, as can be seen in Figure 3.

Finally, the Maps and Trajectories module (M&T) uses Google Maps APIs to trace routes (geographical trajectories) out of the XML notations that complement the descriptions of trajectories in the text. When patterns containing sequences of location entities are fired, routes are generated. The coordinates determined by the GEO module are used to compute the distances between locations. Routes need to have at least three locations in order to be marked as such.

3.4. The client-server model

The system is developed as a Client-Server architecture in which various modules continuously exchange data. The modules described up to this point run on the Server. Also there resides the file marked in Figure 1 as the Portrait, which keeps all relevant information about the user.

Two other modules run on the Client – in our case, the user's mobile device: the HCI Interface (INT), displaying appropriate maps, diagrams, associated information from web sources and highlights on the text, and the Augmented Reality (AR) module.

The client interface receives from the Server and displays the current section of the active document, which is fragmented in chunks smaller than page-size. The reason for the segmentation is to minimize the volume of circulated data. As seen in Figure 2, the interface displays the current text chunk and the location of a selected entity (in this case "Bucium"), as well as the additional information fetched by the EC module (when it exists). Figure 3 shows the current position of the device (determined by the mobile's GPS) as well as minimum 3-4 locations nearby (sorted dynamically).

The Augmented Reality module (only sketched in the current prototype) is intended to augment the image captured by the mobile's camera with fleshes pointing the direction of relevant objectives revealed by the current chunk of text (important buildings and streets in a town, mountains in the open, etc.).

3.5. Current development state and evaluation status

This architecture offers flexibility and reusability. The described design, as proved during the development of the prototype realised by students, has also the advantage of modularity, the individual modules being built, added and tested as unrelated small projects by groups of students. The prototype was coded in Java and uses JSON databases and Android APIs.

[8] http://www.geonames.org/about.html

MappingBooks is designed as a mobile application, allowing users to access the content they have rights to (by entering an access code or by uploading a document themselves).

After the prototype system is closer to the final stage of development an evaluation stage is planned. Evaluation can be done both at module level and as a whole system. Since the individual modules are changeable with any equivalent tool providing the same functionalities, and since the evaluation of those modules is a task optimally performed by their developers and to be discussed when describing those tools individually, we will focus mostly on evaluating the general performance of the whole system. Quantitative evaluations will be done by comparing the output of the system against annotations in a corpus containing relevant examples. But the nature of the system makes it more suited to a qualitative evaluation, to be performed by people in interested social categories: students, authors and editors, general consumers of geographical textbooks and travel guides. This type of evaluation requires a system near completion, so it will be carried out and its results described in a forthcoming paper.

4. Conclusions

The project advances the state-of-the-art in several directions. First, it invents a technology able to automatically annotate toponyms in Romanian texts and link them onto digital maps. Then, the hypermap structures are automatically updated based on user-dependent contexts. Starting from the book text, the user receives support which is presented in a graphical interactive manner. A mixed reality technology indicates on the screen of a graphic tablet or of a smart phone, superposed on an image captured by the device camera, the position of certain geographical guiding marks (mountains, hills, localities, roads, railway lines, outstanding buildings, etc.), simultaneously with highlighting them on the text of the book. Moreover, exploiting collaboration with the Faculty of Geography at UAIC, the project creates an inventory of spatial and geocoded data, which can supplement the map with information related to the context of the identified toponyms. The project also intends to augment the linguistic resources for the Romanian language with a manually annotated corpus (a Geography manual or a touristic guide) at geo-content, out of which minimum 500 instances of geographical concepts will be generated and integrated as instances onto the Romanian WordNet (Tufiș and Cristea, 2002), as open data for research.

Acknowledgements

The authors thank the students in Computer Science at UAIC, who, during the first term of the university year 2013-2014 realised a prototype of the MappingBooks system as their class project in the *Artificial Intelligence* course, under the supervision of the authors and our colleague Mădălina Răschip. The research described in this paper constitutes preliminary work to the PN-II-PT-PCCA-2013-4-1878 Partnership PCCA 2013 grant, having as partners UAIC, SIVECO and „Ștefan Cel Mare" University of Suceava.

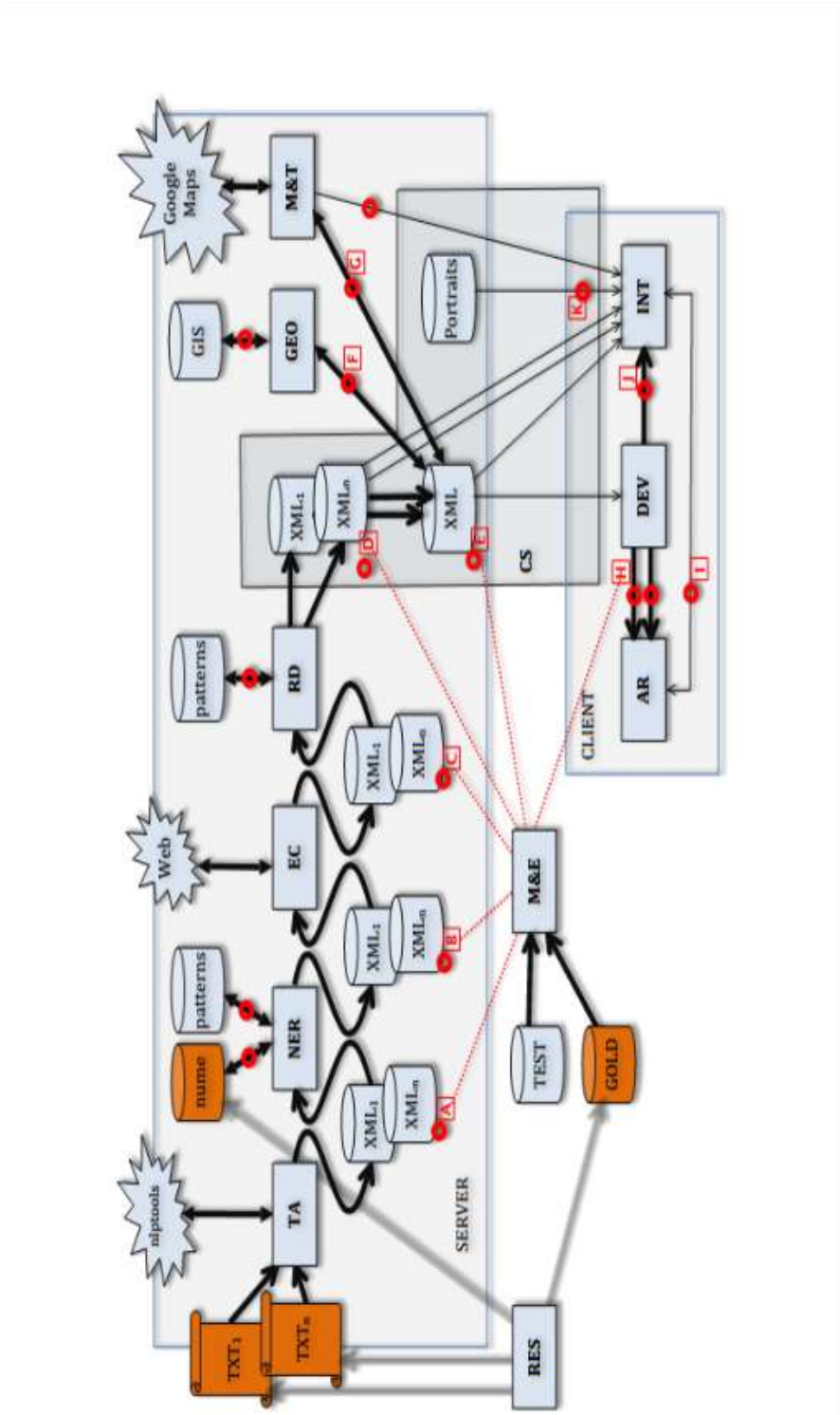

Google Maps
M&T
GIS
GEO
Portraits
INT
DEV
AR
CLIENT
CS
XML
RD
patterns
EC
Web
NER
numе
TA
nlptools
M&E
TEST
GOLD
RES
SERVER
A
B
C
D
E
F
G
H
I
K
L

Figure 1: The architecture of the MappingBooks system

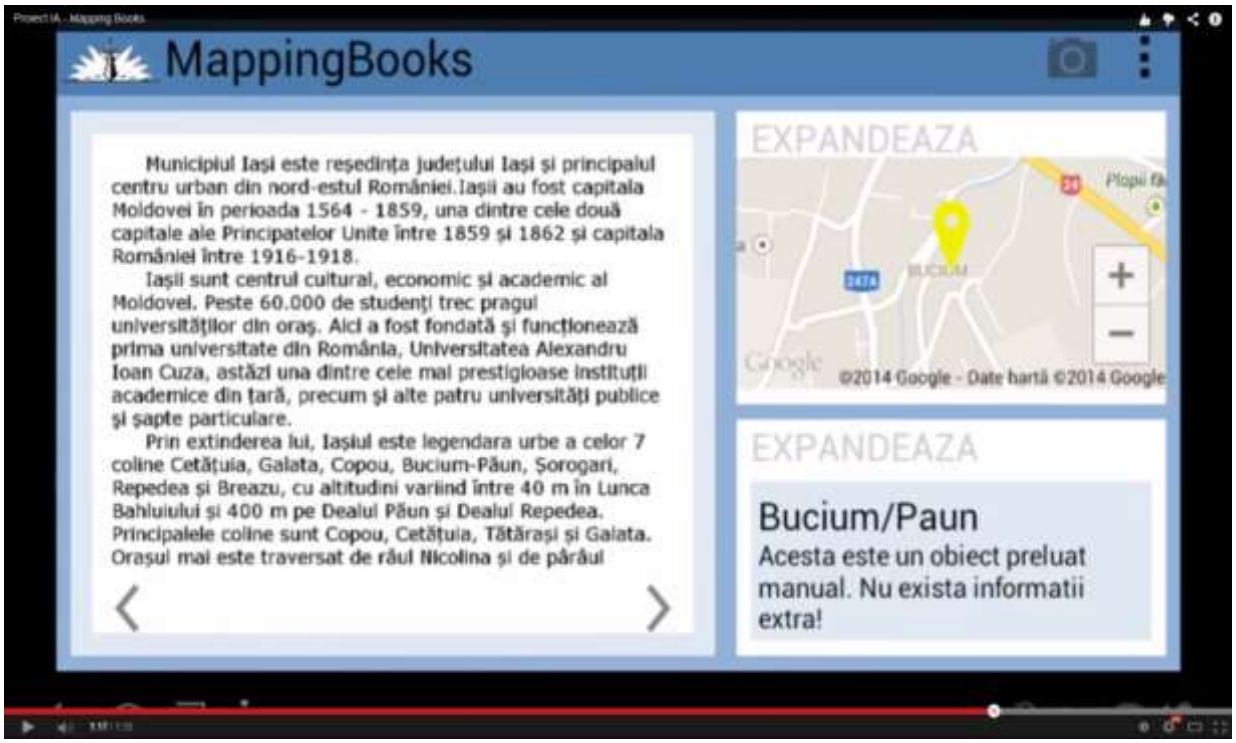


Figure 2: Application tablet interface: text and geolocations viewer

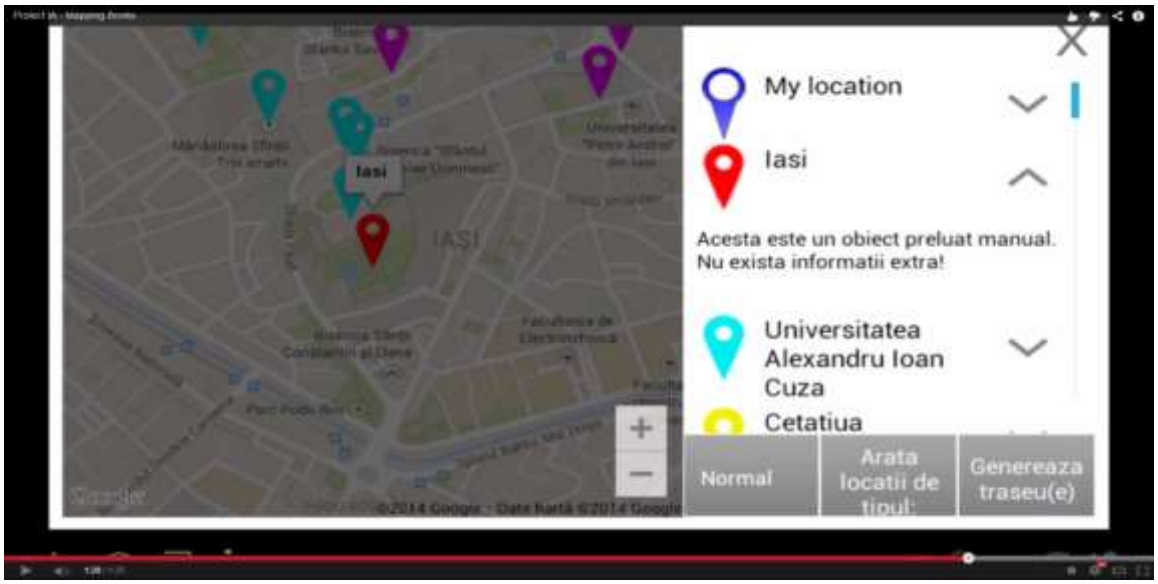


Figure 3: Application tablet interface: important locations near user device

References

Anechitei, D., Cristea, D., Dimosthenis, I., Ignat, E., Karagiozov, D., Koeva, S., Kopeć, M., Vertan, C. (2013). Summarizing Short Texts Through a Discourse-Centered Approach in a Multilingual Context. In Amy Neustein and Judith Markowitz (eds.) *Where Humans Meet Machines*. New York: Springer Science.

Bagga, A., Balwdin, B. (1999). Cross-document event coreference: annotations, experiments, and observations. In *CorefApp '99 Proceedings*.

Cristea, D., Ignat, E., Anechitei, D. (2012). ATLAS project – the Romanian Component, in L.Alboaie et al. (eds.) *BringITon! 2012 Catalog*, Editura Universității „Alexandru Ioan Cuza" Iasi, May, 26-27.

Cărăușu, G. (2011). Processing Spatial Relations In Old Texts And Their Transposition On Modern Maps (in Romanian: "Prelucrarea expresiilor spaţiale în textele vechi pentru realizarea echivalărilor topografice în hărţile actuale"), graduation thesis, Faculty of Computer Science, "Alexandru Ioan cuza" University of Iaşi.

Ciucanu, A. M. (2011). Iter in Chinam – Reconstructing the Journey of Milescu Spătarul from Russia to China (in Romanian "Iter in Chinam – Reconstituirea traseului lui Milescu Spătarul din Rusia până în China"), graduation thesis, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi.

Cristea, D., Dima, G. E. (2001). An Integrating Framework for Anaphora Resolution. In *Information Science and Technology*. Bucharest: Romanian Academy Publishing House, 4:3-4, 2001. In December 2003 this paper was distinguished with the 2001 "Grigore Moisil" Romanian Academy Award for Information Technology.

Cristea, D., Postolache, O. (2005). How to deal with wicked anaphora. In Antonio Branco, Tony McEnery, Ruslan Mitkov (eds.): *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, Benjamin Publishing Books.

Dumbravă, D. (2013). Nicolae Milescu's Iter in Chinam (1676). Visual, computational and encyclopaedic reconstructions, cercetare postdoctorală la Universitatea "Alexandru Ioan Cuza" din Iaşi, 01.10.2010 - 31 03.2013.

Dumitrescu, Ş. D., Mititelu, V. B. (2013). Instantiating Concepts of the Romanian Wordnet. In *Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, Miclăuşeni, 16-17 May 2013, "Alexandru Ioan Cuza" University Publishing House, Iaşi.

Ferruci, D., Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 10:3-4, 327-348.

Năstase, V., Nakov, P., Séaghdha, D. Ó., Szpakowicz, S. (2013). *Semantic relations between nominals*. California: Morgan & Claypool Publishers.

Potolincă, A. (2014). Geographic Entities Recogniser for Romanian. Graduation thesis, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași.

Saggion, H., (2007). SHEF: semantic tagging and summarisation techniques applied to cross-document coreference. *SEMEVLA '07 Proceedings*.

Simionescu, R. (2011). Hybrid POS Tagger. In *Proceedings of "Language Resources and Tools with Industrial Applications"*, workshop afiliated to Eurolan 2011 summer school, Cluj-Napoca.

Singh, S., Subramanya, A., Pereira, F., McCallum, A. (2011). Large-scale cross-document coreference using distributed inference and hierarchical models". *HLT '11 Proceedings*, 1.

Tufiş, D., Cristea, D. (2002). RO-BALKANET – lexicalised ontology, in a multilingual context, for Romanian language (in Romanian: RO-BALKANET – ontologie lexicalizată, în context multilinv, pentru limba română). In Dan Tufiş, Florin-Gheorghe Filip (eds.): *Limba Română în Societatea Informaţională – Societatea Cunoaşterii*. Bucureşti: Editura Expert, 137-164.

EXTRACTING BACKGROUND KNOWLEDGE ABOUT THE WORLD FROM THE TEXT

LAVINIA-MARIA GHERASIM, ADRIAN IFTENE

"Alexandru Ioan Cuza" University of Iaşi, Faculty of Computer Science

{lavinia.gherasim, adiftene}@info.uaic.ro

Abstract

Currently, information extraction is an area of great interest aiming to obtain structured information from unstructured or semi-structured text. It is part of the Artificial Intelligence domain and it has many practical applications in various fields: medicine, finance, politic, Internet surveillance, etc. This paper aims to address the issue of information extraction from simple text, in a first phase, and then to show how it is augmented with information from Wikipedia pages by recognizing entities and identify properties and relations between them, in the second phase. We propose an approach in this direction due to the complexity of the Romanian language and the fact that resources in this language are more limited than in English. We have chosen Wikipedia in the second phase, because we think it brings many advantages by the way it is structured and because it gives a large volume of data that can be used for what we want. The project is focused on the analysis of the first paragraphs from Wikipedia articles, but it can be extended to full articles.

Key words — Information extraction, Wikipedia, Protégé.

1. Introduction

Nowadays, we have access to a lot of information from every domain and it is important to find a way to identify what we need, having tools which can extract only the relevant information. Thus, the text mining techniques that allow data classification, data filtering and information extraction from simple text consist in the usage of specific methods of natural language processing which can be very useful in our daily activities.

For instance, let us assume that we have an email with the *Subject*: "Meeting", *Date*: "10/19/2013", *Content*: "Dear Mr. Andrews, I'm writing to inform you that the congress will take place tomorrow, at the President Hotel, between 2 and 3 PM. Yours faithfully, John Smith". If there are hundreds of emails, it is very easy to lose important information. Therefore, an application able to identify the *date*, the *event*, the *time* and the *location* of the meeting and to create a calendar event with all these elements, will help the user by filtering only the information that he needs. To make this possible, we have to identify some elements in the text like dates, temporal or spatial clues ("tomorrow", "Hotel President"). Information extraction is a very complex process that supposes lexical analysis, followed by a syntactical one, combined with named entity recognition and classification. The previous approaches are specific to English and they were focused on the usage of large resources like Wikipedia (Milne et al., 2006) or Freebase (Bollacker et al., 2008).

Regarding the analysis of large data collections, another important part is related to sentiment analysis. For instance, let us consider the example of a user who wants to buy a product online. She/he has some expectations from it and he is interested in the opinion of other buyers. In this context, an application that can automatically extract information from the Internet, about the product, based on the opinions of those who have already bought it, would be very useful. In this way, the user saves time and she/he gets the information that can help her/him to decide whether to buy the product or not. In the same way, we can think about scenarios where sentiment analysis can be useful to politicians, big companies, investors and so on. In this area, there are approaches in a multilingual context (Gînscă et al., 2011), and even for Romanian (Gînscă et al., 2012). The limitations of these systems are clear and they can be overcome by completing the information obtained with information about the world that has the purpose of facilitating and clarifying the understanding of results.

In this context, the application we developed can help users who analyse a text, for a better understanding with more information about the world from large resources of data like Wikipedia, Freebase, DBpedia, etc.

2. Resources

2.1. Wikipedia

Wikipedia[1] is an online encyclopaedia, available in multiple languages, which can be edited by anyone. Its characteristics and the way in which it is organized help the process of information extraction, providing more accurate results. Until 2004, Wikipedia did not have a dedicated system for organizing articles. After this date, category pages have been introduced – they are in fact collections of links to various articles or other category pages. Thus, it became possible to assign an article to a certain category or to create links between categories, which created a classification system and hierarchy by grouping articles by their content. Now, Wikipedia resource consists of the following components (Stoutenburg and Kalita, 2009): articles, article redirects, article links, categories and disambiguation pages.

In the system that we developed, we used Wikipedia to validate named entities (if there is a Wikipedia page for that entity, then the entity exists, too) and to add information to the knowledge graph. For adding data to the graph, we used the Wikipedia API to extract the first paragraph from a given page (with the observation that the application can be extended to consider the content of the whole page) and to extract the information from the *infobox* section.

2.2. Freebase

Freebase[2] is a large online database, which can be edited in the same way as Wikipedia articles. This contains about 22 million entities grouped into *people*, *places* and *things*. Each entity is uniquely identified by its *ID*. Entities are connected in Freebase as in a graph. Its structure is represented by a set of nodes and a set of links that establish

[1] http://www.wikipedia.org/
[2] https://www.freebase.com/

relations between nodes (Bollacker et al., 2008). The information from Freebase is accessed using a query language MQL (Metaweb Query Language).

A topic from Freebase can be seen in many ways and thus a topic can have many types and for every type there exists a set of properties. For example, "*Bob Dylan was a song writer, singer, performer, book author and film actor*". The Freebase types are grouped into domains, each domain has an identifier, for example */music* is the Music domain. These IDs are conceptually arranged as in a file directory hierarchy. Every topic has an ID in the "*/en*" namespace, for example "*/en/Paris*" is the ID for Paris topic. This ID is used to uniquely identify a node in the graph and, in the same time, it is used to extract the corresponding values from the graph, when this is possible.

2.3. DBpedia

The DBpedia project is based on information extraction from Wikipedia articles. These consist mostly in free text, but they include structured information too, such as info boxes, images, geo-coordinates and links to other pages. All these data are extracted and organized in such a way, so that they can be queried. The information from DBpedia is represented like in a large ontology. It uses the RDF format to represent the information and the SPARQL language to query it. Every entity in DBpedia is identified by an URL reference like *http://dbpedia.org/resourse/Name*, where the *Name* is part of the URL Wikipedia article, represented by *http://ro.wikipedia.org/wiki/Name*.

3. System architecture

Firstly, the application identifies named entities in the text (which will be transformed into concepts and then into nodes in the graph that will be built), their features (properties of the nodes in the graph), followed by the identification of links between them (the edges of the graph). Then, in the graph, additional information about the world is added (both nodes features and links between them), using the mentioned resources Wikipedia, Freebase and DBpedia.

The project consists of three major modules: (1) the first one handles the processing of the original text (any text given by the user); (2) the second one extracts the text from the external resources and the (3) third module represents the obtained results using OWL (Web Ontology Language). Figure 1 gives an overview of the system architecture.

3.1. The analysis of the original text

POSTagging: Initially, we did the separation into phrases and the morphological analysis using a web service for Romanian language named PosTagger[3] (Simionescu, 2011), which takes a simple text as input and provides an XML as output. The file contains several tags, such as *lemma*, *POS* (part of speech), the *specification if a word is appropriate or not*, etc.

Named entity recognition: Based on this XML file, using two rules, we transformed proper names into named entities. The rules were these: (1) many successive capitalized

[3] http://nlptools.infoiasi.ro/WebPosRo/

words and marked by POSTagger as proper names are grouped into a single entity, (2) many capitalized words separated by linking words such as “din”, “de”, etc. (in En: *from*, *of*) were also grouped into one entity (for example “Venus din Milo” (in En: *Venus from Milo*), “Camera Internațională de Comerț” (in En: *International Chamber of Commerce*), etc.). We validated these entities using Wikipedia or our external resources (an entity is considered valid if it has a corresponding Wikipedia page or it exists in our external resources).

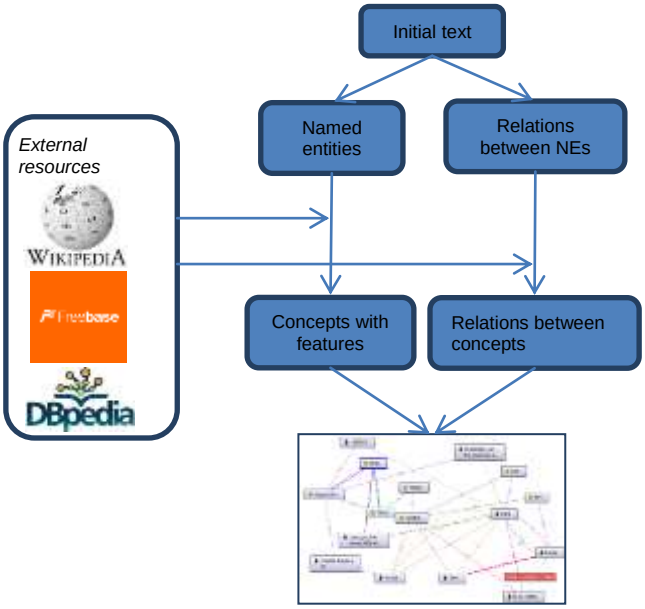


Figure 1: System architecture

Named entities classification: Next, we tried to identify the type of the entities found in the previous step. For this, we analysed the text and we looked at the words in the neighbourhood of the entity in order to make the classification. Thus, in the situations where we found in the text expressions such as “Palatul Ghica” (in En: *Ghica Palace*), “Muzeul Luvru” (in En: *the Louvre Museum*), “fluvial Sena” (in En: *the Seine River*), we considered the corresponding type “palace”, “museum” or “river” for them. In the other situations, we used our external resources, where we have entities of the type *location*, *organisation*, *people* and *other*. In addition to this, we created a list of subtypes for each of the four types using the class hierarchy from DBpedia and Schema.org. Thus, we can identify that an entity of the type “museum” or “palace” is actually an entity of the type *location*.

Anaphora resolution: we did the anaphora resolution both in the original text and in the Wikipedia articles. (1) In the original text we started from the classification of the named entity. Thus, after the classification of the entity in a certain class, we used it to replace the references to the class in the text by its corresponding entity. For instance, after we classified “Palatul Ghica” as “palace”, all the occurrences of the word “palat” (En: *palace*) (that were not followed by the word “Ghica”) have been replaced by the expression “Palatul Ghica”. (2) At the level of Wikipedia articles, we took advantage of their structure and of the fact that the first paragraph refers to the presented concept from the current Wikipedia page. Thus, we considered that all expressions such as: “A

fost inaugurat" (in En: *was inaugurated*), "Este conceput" (in En: *is designed/conceived*), "Este considerat" (in En: *is considered*), etc. refer to the main concept, described in that Wikipedia page. The same approaches are used for the pronouns in the text.

Identifying concepts: The same entity can appear many times in an article, in different sentences or in different ways. In the graph that we built, we considered concepts and every concept is unique. Therefore, different entities can refer the same concept: for instance "Luvru", "Muzeul Luvru", "Luvrul" (in En: *Louvre*, *the Louvre museum*, *Louvre*) indicate in fact only one concept (and they will be represented by a single node in the graph). Before adding an entity to the list of concepts, we check if it is not already in the list. This verification is done using the Freebase API, based on the property ID from Freebase: two concepts are equal if they have the same ID. In the previous example, all three entities are represented by the ID "*/en/Louvre*".

Finding relations between entities (*concepts*): Firstly, we identified these relations at text level (among entities), and then we transformed them at the graph level (among concepts). In order to establish relations between entities, we considered only entities from the same phrase. For this, we used several rules that cover various situations: (1) If there is an enumeration of entities in the same phrase, (2) If there are exactly two entities within a phrase, (3) If there is only one entity within a phrase, (4) Others. (1) An enumeration of entities suggests that there is a connection between them and they have to be seen as only one entity. After the identification of the mentioned enumerations, their components are joined into one entity and then the next step is to find relations. Thus, a relation will be established between the unified entity and the entities existing in the phrase before and after the enumeration. (2) Having exactly two concepts in a phrase is the easiest case, because we are sure that there will definitely be a relation between them. (3) The case with only one concept in a sentence identifies a feature of the concept, not a relation, and for (4) other situations the relations are established using patterns with regular expressions (similar to (Iftene et al., 2007)).

3.2. Using external resources

We have already seen how we used external resources to validate the named entities (using Wikipedia) and to identify concepts (using Freebase). In addition, we used these external resources as it follows:

Wikipedia, for adding features to concepts: In addition to using the first paragraph of Wikipedia pages, we decided to use the *infobox* section, which provides a summary of the most important information about the entity presented in the text. For this, we used the DBpedia resource which contains structured information from Wikipedia, in RDF format. Thus, from the DBpedia server, we downloaded the dataset corresponding to the *infobox* section for Romanian language: the tag "raw_infobox_properties".

Freebase, for identifying relations: First, for the identified entities, we extracted from Freebase the relations that are already known. Please note that the graph in Freebase is oriented and we can perform queries that return the links that came in or came out from a node. The properties that can be queried are: "*/type/reflect/any_master*"(matches any outgoing link to another object), "*/type/reflect/any_reverse*" (matches any incoming link

from another object) and "*/type/reflect/any_value*" (matches any link to a primitive value such as "/type/text", "/type/datetime" or "/type/float"). We have chosen the property"*/type/reflect/any_reverse*", because the search of relations between entities is done in both directions.

Protégé, to visualize the results: After text analysis is done, the information obtained is represented using ontology. For building the ontology, we used Jena API[4] which created an OWL file (Web Ontology Language). To visualize the results we used Protégé[5], an ontology editor developed by Stanford Center for Biomedical Informatics Research from Stanford University School of Medicine. In the ontology created, we associated an *Individual* for each concept. Then, the four types: Location, People, Organisation and Other represent *Classes*, properties correspond to *Data Properties* and relations to *Object Properties*.

4. Case study

To illustrate the results, we have chosen a Wikipedia article, where we will explain, step by step, what worked well and what did not and we will also present some statistics after the analysis of the text. The chosen article refers to the Wikipedia page which corresponds to the city of Paris[6].

4.1. POSTagging

In this step, we called the POStagging service and we obtain the lemma for every word, the POS and the information if a noun is proper or not. After this step, we obtained for "Paris", "Parisul" the same lemma: "Paris" and this helped us to unify them in the same concept later.

4.2. Identifying entities

This step is one of the most important since all next steps are based on it and a good precision of the results is essential for establishing relations between concepts. Entities such as "Paris", "Franța", "Sena", "Europa", "UNESCO"(En: *Paris*, *France*, *Seine*, *Europe*, *UNESCO*), etc. are correctly identified by the POStagging service as proper nouns. Problems often occur at the form "Parisul" (the articulated form of Paris), which is sometimes recognized by PosTagger as a proper noun and sometimes not. There were problems even for the form "Paris", when it appears at the beginning of the sentence even if it is followed by the word "regiunea" (En: *region*), which should help in the process of classification for the entity. In the end, we obtained an accuracy of 81.48% for entity identification and classification, but this value can be improved by using rules that take into consideration the entities identified in a previous step.

4.3. Anaphora resolution

At this phase, we identified the type of the main concept "capitală" (in En: *capital*) and

[4] https://jena.apache.org/documentation/ontology/
[5] http://protege.stanford.edu/
[6] http://ro.wikipedia.org/wiki/Paris

its subtype "oraș" (En: *city*). According to the rule explained above, in the sentences "Orașul este traversat de fluviul Sena" (in En: *The city is crossed by Seine*) and "Orașul în limitele sale administrative" (in En: *The city in its administrative limits*), "Orașul" (in En: *city*) is replaced by "Paris". Also, there is a pronoun in the text which refers to the main entity and which will be replaced, in the sentence "Acesta găzduiește sediul mai multor organizații internaționale" (in En: *It hosts the headquarters of many international organisations*). Another case is found in the expression "Este una dintre cele mai populate zone metropolitane din Europa" (in En: *It is one of the most populated metropolitan areas in Europe*), where "Paris" will be inserted at the beginning of the sentence.

4.4. Identifying concepts

From the 27 entities identified in 4.2, we got 11 concepts, 2 of which remained unclassified: "Clubul Paris" (in En: *Paris club*) and "Camera Internațională de Comerț" (En: *International Chamber of Commerce*). In Table 1, we can see how their types or subtypes are correctly identified, even if not all of them have a subtype.

Table 1: Concepts, ids, types and subtypes identified in the article "Paris" from the Romanian Wikipedia

Concept	**Id**	**Type**	**Subtype**
Paris	/en/paris	Location	City
Franța	/en/france	Location	Country
Sena	/en/seine	Location	River
Île-de-France	/en/paris_region	Location	Region
Europa	/en/europe	Location	-
UNESCO	/en/unesco	Organisation	-
Organizația pentru Cooperare și Dezvoltare Economică	/en/ocde_staff	Organisation	-
Camera Internațională de Comerț	/en/paris	-	-
Clubul Paris	/en/paris	-	-
Uniunea Europeană	/en/europe	Organisation	-
Locuri din Patrimoniul Mondial UNESCO	/en/world/heritage_site	Other	-

4.5. Identifying relations

The two steps described above involve (1) the extraction of relations already known from the Freebase or from Wikipedia infobox section, followed by (2) the identification of new relations in the analysed text. Initially, our knowledge base is populated by:

- Paris – *partially contains* – Sena (En: Siena);
- Franța (En: France) – *capital* – Paris;
- Franța – *contains*– Sena;
- Uniunea Europeană (En: European union) – *geographic_scope* – Europa;
- Name: *Paris*;
- Official name: *Ville de Paris*;
- Image: *Paris_-_Eiffelturm_und_Marsfeld2.jpg*;
- Type: *Capitală, Franța, Capitala Franței* (in En: capital);

- Motto: *Fluctuat nec mergitur*;
- Flag image: *Flag of Paris.svg*;
- Emblem: *Grandes Armes de Paris.svg*;
- Alias: *La ville lumière*;
- Subdivision type: *Regiunile Franței*;
- Subdivision name: *Île-de-France*;
- Census: *2007*;
- Type: *Oraș* (En: *city*);
- Population: *2,193,031*;
- Total area km: *86.9*;
- Density: *24,948*;
- Urban population: *10,142,825*.

For the second step, initial checking must be done to see if there are enumerations in the text, therefore the application first identifies "UNESCO, Organizația pentru Cooperare și Dezvoltare Economică, Camera Internațională de Comerț sau informalul Club Paris?". Before this, the entity "Paris" is found in the sentence (the original sentence was "Aceasta găzduiește sediul..." (in En: *It hosts the headquarters*), but according to the previous step it becomes "Paris găzduiește sediul..." (We can see now the importance of the anaphora resolution). Therefore, it will add a relation for Paris and each of the elements that appear in the enumeration, as follows:

- Paris - *găzduiește sediul mai multor organizații cum ar fi*– UNESCO;
- Paris - *găzduiește sediul mai multor organizații cum ar fi* - Organizația pentru Cooperare și Dezvoltare Economică.

According to the rule, there must be 4 relations, but the last two entities relate to the same concept Paris. Because we have the restriction not to build relations between the same concepts, 2 relations are eliminated. The second enumeration in the text consists only of two words connected by the conjunction "și" (in En: *and*): "Paris și regiunea Paris" (in En: *Paris and Paris region*). Because in the same sentence we have the entity "Europa", there will be established a relation with this:

- Paris – *produc mai mult de un sfert din produsul intern brut al* – Franței.

As we can see, the relation: "regiunea Paris – *produc mai mult de un sfert din produsul intern brut al* – Franței", is not found. Because, "regiunea Paris" is not part of the list of concepts, it is associated with the concept Paris, and this is the reason why we have only one relation here.

For the next rule, we studied the sentences where there are exactly two entities and we extracted the following:

- Paris – *este capitala și cel mai mare oraș din* – Franța;
- Paris – *este una dintre cele mai populate zone metropolitane din* – Europa (again, anaphora resolution helps);
- Paris – *este considerat unul dintre cele mai verzi și mai locuibile orașe din* – Europa;
- Paris – *conține 3.800 de monumente istorice și patru* – locuri din Patrimoniul Mondial UNESCO;
- Paris – *este cea mai mare din* – Europa.

There are situations when a certain relation must be associated with a feature of the entity, as it can be seen in the last example, where it would have been correct to say: aglomerația din Paris – *este cea mai mare din* – Europa (in En: *crowds in Paris - is the largest in – Europe*). Therefore, not only are the words between entities important, but also the words before them. Our proposed method uses the dependency path between words, which can be obtained with FDG Parser[7] and it is based on analysing the parts of sentence. Thus, for an entity that is of type Attribute, we will search for the words that depend on it, to the left. Another observation is related to the concord between subject and predicate (Paris – *produc mai mult de un sfert din produsul intern brut al* – Franței). When we have an enumeration, it would be more accurate to transform the verb which expresses the relation, by using the singular form of the verb for each component. Thus, the verb "produc" (En: *produce* at a plural form) should be replaced by "produce" (in En: *produces* is at a singular form).

When a relation between two entities is established, the main problem is which of the words best expresses the relation between them, and which ones should be removed, because they give additional information that is not relevant for what we want to illustrate. For instance, let us consider the following sentence "Paris și regiunea Paris, cu 552.1 miliarde € în 2009, produc mai mult de un sfert din produsul intern brut al Franței.", where the most important relation is: Paris - *produce mai mult de un sfert din produsul intern brut al* – Franței, all the other details should be ignored (eventually an independent property can be added for the value of the PIB).

4.6. Representing the ontology with Protégé

After the processing explained above, the result is a knowledge base, which can be seen in the figure below, as it is illustrated in Protégé (the concepts are nodes with features and the relations between concepts are edges):

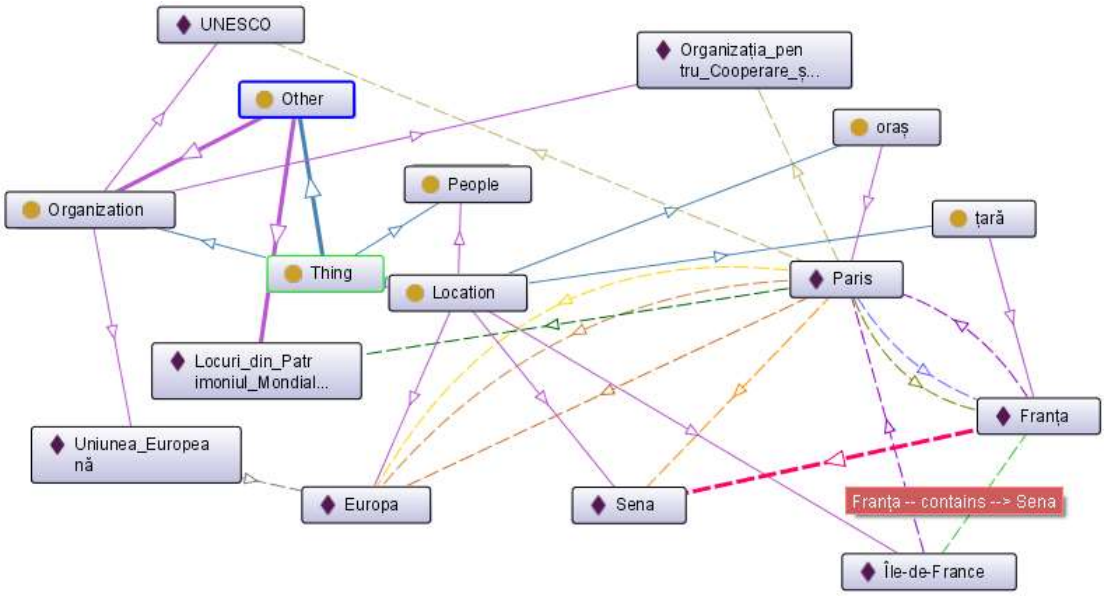


Figure 2: The final graph for Paris

[7] http://nlptools.infoiasi.ro/WebFdgRo/

5. Conclusions

This paper presents a way of augmenting the existing information in a text using the information extracted from large data resources such as Wikipedia and Freebase. The system uses a POSTagger for the Romanian language; afterwards it identifies and classifies the named entities. For these entities, the external resources mentioned above are used and concepts are identified (an entity that can appear in the text in various forms) and the relations between them.

The problems we faced started from the results returned by POSTagger that are not always accurate, but which can be improved with some post-processing. Other problems are connected to the relation identification among concepts in the text that is analysed. Sometimes, these relations are too detailed and they should be parameterised. In the future, using the analysed texts, we plan to build a resource with the features of the identified entities and with the relations among them. This resource would increase in time and it should be useful to those interested in analysing the named entities.

Acknowledgments

The research presented in this paper was funded by the project MUCKE (Multimedia and User Credibility Knowledge Extraction), number 2 CHIST-ERA/01.10.2012.

References

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of ACM SIGMOD international conference*.

Gînscă, A. L., Boroș, E., Iftene, A., Trandabăț, D., Toader, M., Corîci, M., Perez, C. A., Cristea, D. (2011). Sentimatrix - Multilingual Sentiment Analysis Service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (ACL-WASSA2011). Portland, Oregon, USA.

Gînscă, A. L., Iftene, A., Corîci, M. (2012). Building a Romanian Corpus for Sentiment Analysis. *In Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language" 8-9 December 2011, 26-27 April 2012*, Editura Universităţii "Al.I.Cuza", Iași, 63-71.

Iftene, A., Trandabăț, D., Pistol, I. (2007). Grammar-based Automatic Extraction of Definitions and Applications for Romanian. *In Proceedings of RANLP workshop "Natural Language Processing and Knowledge Representation for eLearning environments"*, September 26, 2007, Borovets, Bulgaria, 19-25.

Milne, D., Medelyan, O., Witten, I. H. (2006). Mining Domain-Specific Thesauri from Wikipedia: A Case Study.

Simionescu, R. (2011). POS-tagger hybrid, Master Thesis, Faculty of Computer Science, "Al. I. Cuza" University of Iasi.

Stoutenburg, S., Kalita, J. (2009). Extracting Semantic Relationships between Wikipedia Articles. *In Proc. 35th International Conference on Current Trends in Theory and Practice of Computer Science*, Czech Republic.

DEFINING HIDDEN SYNTACTICAL PATTERNS FOR AN ENCRYPTION/DECRYPTION SYSTEM

NICOLAE CONSTANTINESCU, MIHAELA COLHON

Department of Computer Science, University of Craiova, Craiova – România
nikyc@central.ucv.ro, mghindeanu@inf.ucv.ro

Abstract

The main purpose of the cryptography is to find a way to hide the information from the users which do not have the rights to know it, but to reveal to those that have certain things that authorize them to access the contents. In the present article we present a way to accomplish this request by introducing a new way of encryption: by meaning of the logical links between the propositional parts.

Key words — syntactic patterns encryption, morpho-syntactic tagging.

1. Introduction

Cryptographic algorithms as any other application of information theory have been developed to accomplish essentially the same thing: to have access to certain information by users which can prove this right.

In the present article we propose a method for extracting information using hidden syntactic patterns of written natural language texts. The proposed scheme is necessary to separate noise from underlying signal. The presented study addresses to Romanian texts but can be easily adapted to any other language because the main things of the encryption scheme is the set K of grammar rules which can be defined for each language to be specific.

While in stenography, the hidden information is in images, in this type of security information the hidden messages reside within the propositional data, but they, unlike stenography, will be also encrypted, as it will be shown in the article.

The structure of a sentence codifies the relationships that exist between the words of that sentence. It also indicates how the words are grouped into syntactic phrases, which words play a functional or a central role in the sentence meaning. Usually, all this information is stored in a tree representation.

Based on this syntactic information relative to a natural language statement, we define a new encryption method that will be called in what follows as **Sy**ntactic **P**atterns **En**cryption (SYPEN) method. The proposed method has the advantage that is hard to be decrypted but very easy to be formulated. The method can be applied on any form of written natural language text with the key defined by a set of certain grammar rules which achieve the described completeness rules from Subsection 3.5.

The article is organized as follows. It starts with the Introduction section and because we are not aware of any other similar study the Related Works section is missing. As a consequence, the second section is dedicated to the current syntactic parsing formalisms

developed for Romanian language. In the following section we present the proposed encryption/decryption system, a practical example and the completeness of the scheme.

2. Syntactic Parsing

The most common type of linguistic annotation is Part-Of-Speech (POS) tagging or, more accurately, morpho-syntactic tagging. The morpho-syntactic annotation describes the annotated words in terms of grammatical tagging (*Noun*, *Verb*, *Pronoun* ...) and morphological information (e.g. *gender* or *number*). Often, POS tagging can include lemmatisation, by indicating the lemmas of the words. Many studies consider that POS tags contain enough syntactic information to support word abstraction in any NLP system training. For example, the search space of a translation rules database can be greatly reduced by focusing only on POS tags instead of real words (Tufiş and Ion, 2007).

The traditional problem of morphological analysis for a given word form is to predict the set of all its possible morphological analyses. A morphological analysis consists of a Part-Of-Speech tag, possibly other morphological features, and a lemma (basic form of the word) corresponding to this tag and the combination of features.

POS Tagging, also called grammatical tagging, is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context – i.e. relationship with adjacent and related words in a given phrase or paragraph (Dang and Luo, 2008).

Base forms, also known as lemmas, base forms or ground form of the words do not contain any morphological derivation of the word (such as gender, number, tense, and so on) but are crucial in order to get the corresponding target word from the dictionary entries.

Remark 1. In a specific language, a word form is uniquely identified by its lemma and the corresponding morpho-syntactic information. The reciprocal is not true: more morpho-syntactic interpretations can correspond to a word form and they are disambiguated by the context.

Despite the various existing linguistic theories, which lead to different ways of viewing sentence structure and therefore syntactic analysis, studies related to this subject (Colhon and Simionescu, 2012; Diaconescu, 2003; Gîfu and Cristea, 2013; Hristea and Colhon, 2012) agree that at the heart of sentence structure are the relations among words. These relations refer either to grammatical functions (*subject*, *complement*, etc.) or to the links which bind words into larger units like *natural language constituents* (phrases or even sentences).

Dependency grammar (DG) is a class of syntactic theories developed by Lucien Tesnière (1959). Within this theory, any syntactic structure is determined by the grammatical relations existing between a word (a head) and its dependents. Any word should depend exactly on one other word (the head), with the exception of the main predicate in the sentence which depends on no other word. Several words may depend on the same head.

3. Syntactic Patterns Encryption – SYPEN System

3.1. Principle Description

The most important things in a communication which we will take in consideration in the present work are:

1. Confidentiality: Let A and B be two parts who want to communicate. They transmit their messages across a channel. If any E will intercept the meaning of the message it will be interpreted like a breaking of confidentiality.
2. Resistance to active attack: if A will transmit a message m to B, E (events-dropper) will not be able to modify m to m', for B. This means that B will receive m not $m'(\neq m)$.

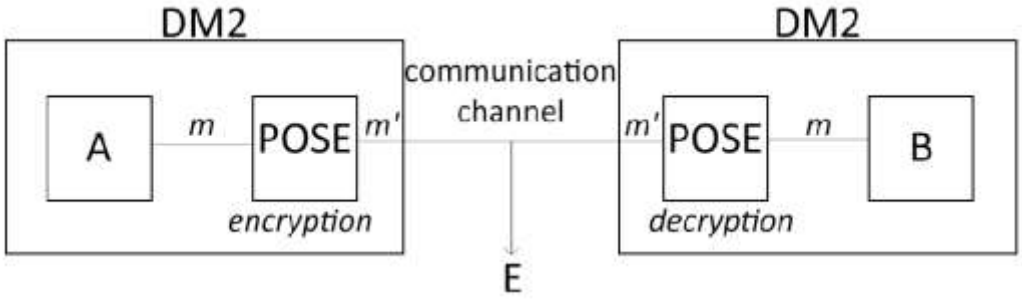


Figure 1: The SYPEN scheme (based on Part of Speech Encryption)

With these conditions in mind, we define our principle scheme as in Figure 1. In this case we will have a model of encryption which is concluded in the next tuple:

(P, C, K, e, d) where:

P = is a set of messages to be used by the communication parts;

C = is a set of encrypted messages to be transmitted in the unsecured channel;

K = is the encryption key;

e = is the encryption function;

d = is the decryption function.

3.2. SYPEN Scheme

In the present work we propose a new way to encrypt and decrypt messages, by hiding the meaning of usefully information into a normal text which is constructed over an internal grammatical set of rules. This is achieved by setting a set of syntactic restrictions for the considered natural language in which the texts are written. Then, the encryption key is the way of joining the resulted constructions.

The task of every syntactic parser is to annotate the syntactic units of a phrase (words or syntactic phrases) with part-of-speech information, POS tags and to represent the syntactic structure of the sentence in a tree form.

We choose to parse the sentences with a Dependency Parser in order to obtain the sentences words annotated for their head-words and the corresponding dependency relations.

We will define *P* the tuple of the encryption scheme as follows:

P = the set of constructions in Romanian which respect the defined internal grammatical rules;

P' = a set of phrases from Romanian language which can contain useful information for *A* and *B* without imposing any grammatical restrictions; obviously $P' \supseteq P$.

K = a set of grammatical rules used in both ways: interpretation and generation (production) for elements of *P';*

e = a way to construct from a message *m* (which *A* wants to transmit to *B*) a set of phrases $T(T \in P')$ – encryption function

d = a way to reconstruct the message *m* from *T* – decryption function.

From these we can conclude:

- Let $p \in P$ be a sequence of words arranged according to the defined set of grammatical rules (the component K)
- We embedded the text of *p* with some "white information" (also in Romanian) and in this manner *p* becomes *p'*. In our formalism *p'* = *T*.
- Following the previous steps one can observe that inside the resulted *p'* is the message *m*:

  From $K = \bigcup_{i=1}^{n} K_i$; $m = \bigcup_{i=1}^{n} m_i$; $p = \bigcup_{i=1}^{n} p_i$ and $m_i \overset{K_i}{\rightarrow} p_i$, whre *i* is the index of the sentence within the message that will be sent from *A* to *B*, we have:

$$\bigcup_{i=1}^{n}(m_i \overset{K_i}{\rightarrow} p_i) = \bigcup_{i=1}^{n} m_i \xrightarrow{\bigcup_{i=1}^{n} K_i} \bigcup_{i=1}^{n} p_i = m \overset{K}{\rightarrow} p$$

$$p' = p + \text{"}white\ information\text{"}$$

$$p' become\ T$$

3.3. SYPEN Case Study

Let be a sentence: "O fată frumoasă citește o carte" (in English "A beautiful girl reads a book"). By parsing this sentence[1] we get the following XML annotations for the sentence words:

```
<s id="1">
```

[1] The Dependency Parser we used for the parsing task was developed based on the Dependency Treebank for the Romanian language that was built by the Natural Language Processing of Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iaşi.

```
<w id="1" lemma="un" ana="Tifsr" deprel="det." head="2">O</w>
<w id="2" lemma="fată" ana="Ncfsrn" deprel="subj." head="4">fată</w>
<w id="3" lemma="frumos" ana="Afpfsrn" deprel="a.adj."
head="2">frumoasă</w>
<w id="4" lemma="citi" ana="Vmip3s" deprel="ROOT" head="0">citește</w>
<w id="5" lemma="un" ana="Tifsr" deprel="det." head="6">o</w>
<w id="6" lemma="carte" ana="Ncfsrn" deprel="n.pred"
head="4">carte</w>
<c>.</c>
</s>
```

The basic layers of resulted annotation include: borders of each sentence (marked as `<s></s>` elements and identified by unique identifiers – attribute id) and words (marked as `<w></w>`) and including unique `ids`, part of speech (attribute `ana`), lemma (attribute `lemma`) and dependency information (dependency relation – attributes `deprel` and `head`).

We will define the key *K* (all the following rules are compulsory):

1. A rule of the grammar associates, in the parsing structure, an attribute adjective to each subject noun. For the considered example this adjective is "frumoasă" (in English "beautiful").

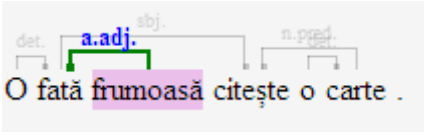


Figure 2: The adjectival relation of the sentence[2]

2. A rule of the grammar associates, in the parsing structure, an article to each subject noun. In our example, this article is "O" (in English "A").

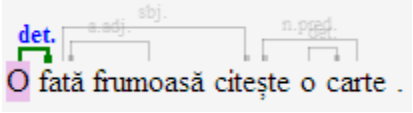


Figure 3: The determiner relation pointing to the subject

3. A rule of the grammar associates, in the parsing structure, an article to each complement noun. For our example, this article is "o" (in English "a").

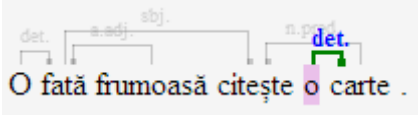


Figure 4: The determiner relation pointing to the complement

[2] The caption was made using the web-service from the address http://nlptools.infoiasi.ro/WebFdgRo/

4. Adding "white information", which in our method means to add sentences that do not obey any of the above grammar rules, the message will become:

   "Afară vremea este indefinită. O fată frumoasă citește o carte. Cartea poate fi dintre cele citite înainte." (in English "Outside the weather is indefinite. A beautiful girl reads a book. The book can be one of the read before").

Let *e* be defined as:

1) $K_1 \rightarrow denote\ an\ approve$

   $K_2 \rightarrow denote\ an\ action$

   $K_3 \rightarrow denote\ a\ time\ schedule\ as$:

   a) $\exists\ of\ it: tomorrow$

   b) $\nexists\ of\ it: each\ day$

With these in mind we can interpret the example as:

Da, Acționează, Mâine (in English: Yes, Action, Tomorrow)

The rest of the message ("white information") will be ignored because it does not respect any of rules from K ($K_i \in K$). The "white information" represents any sentence which has no interpretation in SYPEN scheme. When a decryption function will interpret an encrypted message it will ignore any sentence which has no parsing scheme in accordance with the decryption rules. The meaning of "white information" is to raise the cracking complexity of the scheme. Like in any encryption algorithm, the strength of it consists on key strength and management.

3.4. Completeness of SYPEN

In case of encryption scheme, it is necessary to have:

$$d(e(T))_k = m$$

This means *A* will encrypt a message and will transmit it trough an unsecured channel and *B* will receive it, and if he has d and *K* will be able to reconstruct the message *m* from *T* (Blake-Willson et al., 1997).

Let *m* be the message and the transformation rules from $e(m, K)$ are an implication$e \rightarrow d$, this means that a message *m* can be incorporated into *T* (but inverse is not a valid implication).

From these we will have:

$$e(m, K) = T\ and\ d(T, K) = m$$

There is a hiding message scheme, not an encryption, from standard definition, is a new way to transmit messages, like in stenography scheme robustness (Herzog, 2002). In the second one, the message is hidden into an image, in our scheme is encrypted into a semantic scheme from a set of phrases.

The scheme did not transmit a clear message with "noises", but using a key structure will transform some sentence in totally different assertions, as we explained in the case study section.

4. Conclusions

The present article shows a new way to ensure the information security principle and achieve the goal of less computation power to make the encrypted text and to extract the information from it. It is an application of logic analysis in context and it has enough "necessary time breaking" to be applied on practical applications. The development of the scheme will consist in the implementation of the algorithm and the second step is to analyse different information communication to find if there are logical constructions which can have SYPEN – compatible algorithms implemented.

Acknoledgments

Our research was funded by research projects PN-II-PT-PCCA-2013-4-0614 and UCV 43C/27.01.2014.

References

Blake-Willson, S., Johnson, D., Menezes, A. (1997). Key Agreement Protocols and their Security Analysis , *The Sixth IMA International Conference of Cryptography and Coding*, Cirencester, England, 1355, 30-45.

Colhon, M., Simionescu, R. (2012) Deriving a Statistical Syntactic Parsing from a Treebank. In: *Proceedings of International Conference on Web Intelligence, Mining and Semantics WIMS' 12*, June 13-15, Craiova, Romania, XIX, 439-452.

Diaconescu, Ş. (2003). Natural Language Understanding Using Generative Dependency Grammar. In *Research and Development in Intelligent Systems*.

Dang, C., Luo, X. (2008). WordNet-based Document Summarization. In *Proceedings of 7th WSEAS Int. Conf. on APPLIED COMPUTER & APPLIED COMPUTATIONAL SCIENCE* (*ACACOS '08*), Hangzhou, China, 383-387.

Gîfu, D., Cristea, D. (2013) Towards an Automated Semiotic Analysis of the Romanian Political Discourse, Computer Science Journal of Moldova, 21:1(61), 36-64.

Herzog, J. C. (2002) Computational soundness of formal adversaries, *Master's thesis*, Massachusetts Institute of Technology.

Hristea, F., Colhon, M. (2012). Feeding Syntactic Versus Semantic Knowledge to a Knowledge-lean Unsupervised Word Sense Disambiguation Algorithm with an Underlying Naive Bayes Model, *Fundamenta Informaticae Journal*, 119:1/2012, 61-86.

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

Tesnière, L. (1959). *Elements de syntaxe structurale*. Paris: Klincksieck

Tufiş, D., Ion, R. (2007). Parallel Corpora, Alignment Technologies and Further Prospects in Multilingual Resources and Technology Infrastructure", in C.

Burileanu and H-N Teodorescu (Eds.), *Proceedings of the 4th Conference on Speech Technology and Human-Computer Dialogue*, SpeD 2007, Iaşi, Romania.

INDEX OF AUTHORS