UNITED INSTITUTE OF INFORMATICS PROBLEMS
OF THE NATIONAL ACADEMY OF SCIENCES OF BELARUS

International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015"

NOOJ 2015

Abstracts

June 11–13, 2015, Minsk, Belarus

Minsk
UIIP NASB
2015

SCIENTIFIC COMMITTEE

Xavier Blanco	Autonomous University of Barcelona, Spain
Krzysztof Bogacki	University of Warsaw, Poland
Héla Fehri	University of Gabès, Tunisia
Yuras Hetsevich	United Institute of Informatics Problems of the NAS of Belarus, Minsk
Svetla Koeva	University of Sofia, Bulgaria
Peter Machonis	Florida International University, USA
Slim Mesfar	University of Manouba, Tunisia
Johanna Monti	University of Sassari, Italy
Max Silberztein	Université de Franche-Comté, France
Marko Tadic	University of Zagreb, Croatia
François Trouilleux	Université Blaise-Pascal, France
Simonetta Vietri	University of Salerno, Italy
Igor Sovpel	Belarusian State University
Alexander Zubov	Minsk State Linguistic University, Belarus
Vladimir Golenkov	Belarusian State University of Informatics and Radioelectronics

ORGANIZING COMMITTEE

United Institute of Informatics Problems of the NAS of Belarus, Minsk

Yuras Hetsevich
Barys Lobanov
Tatsiana Okrut
Julia Baradzina
Dzmitry Dzenisiuk
Alena Hiuntar
Stanislau Lysy
Lesia Kaigorodova

University de Franche-Comté, France

Max Silberztein

УДК 004.91

International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015" : Abstracts (11–13 June, 2015, Minsk, Belarus). – Minsk : UIIP NASB, 2015. – 80 p.
ISBN 978-985-6744-89-4.

This volume contains the abstracts of the International conference "NooJ 2015". The research presented covers different aspects of natural language processing using NooJ, including formalizing such levels of linguistic phenomena as syllabification, phonemic and prosodic transcription, multiword units and discontinuous expressions, local and structural syntax; transformational syntax and paraphrase generation, semantic analysis and machine translation, etc.

Abstracts are published in the form presented by authors.

У дадзеным зборніку прадстаўлены тэзісы дакладаў Міжнароднай канферэнцыі "NooJ 2015". Разглядаюцца розныя аспекты апрацоўкі натуральнай мовы з выкарыстаннем лінгвістычнага асяроддзя распрацоўкі NooJ, улічваючы фармалізаванне такіх напрамкаў лінгвістычнага аналізу як склададзяленне, фанетычная і прасадычная транскрыпцыі, устойлівыя выразы і дыскрэтныя слоўныя канструкцыі, лакальны і структурны сінтаксісы, трансфармацыйны сінтаксіс і перафразаванне, семантычны аналіз і машынны пераклад і г. д.

Тэзісы друкуюцца ў выглядзе, пададзеным аўтарамі.

Scientific Editors:

DSc in Engineering B.M. Lobanov,
PhD in Engineering Yu.S. Hetsevich

ISBN 978-985-6744-89-4

of Sciences of Belarus, 2015

Organized by

United Institute of Informatics Problems of the National Academy of Sciences of Belarus

In cooperation with

- Université de Franche-Comté
- NooJ International Association

PREFACE

We are delighted to welcome you to NooJ 2015 International Conference, which both serves to present the latest research in natural language processing with NooJ, as well as to provide the opportunity to renew old friendships and make new acquaintances.

NooJ is both a corpus processing tool and a linguistic development environment that allows linguists to formalize several levels of linguistic phenomena:

– typography and spelling;
– syllabification, phonemic and prosodic transcription;
– lexicons of simple words, multiword units and discontinuous expressions;
– inflectional, derivational and agglutinative morphology;
– local and structural syntax;
– transformational syntax and paraphrase generation;
– semantic analysis and machine translation.

The traditional topics to be covered by the NooJ 2015 conference include the following: linguistic resources: typography, morphology, lexical analysis, local syntax, structural syntax, transformational analysis, paraphrase generation, semantic annotations, semantic analysis; corpus processing: corpus linguistics, information extraction, discourse analysis, business intelligence, NLP applications.

In addition, this year, the organizers use the opportunity to broaden the scope of the conference by adding the following new topics: spelling, syllabification, phonemic and prosodic transcription.

NooJ provides linguists with regular grammars, context-free grammars, context-sensitive grammars, unrestricted grammars as well as their graphical equivalent (finite-state, recursive and contextual graphs) to facilitate the description of each phenomenon. NooJ's multi-layer approach allows linguists to accumulate elementary descriptions and describe phenomena that cross linguistic levels. As a corpus processing tool, NooJ allows users to apply sophisticated linguistic queries to large corpora in real time, in order to construct indices and concordances, annotate texts automatically, perform semantic and statistical analyses, etc.

NooJ is open-source, freely available at www.nooj4nlp.net and over 20 linguistic modules can already be freely downloaded, as well as a manual, video tutorials, references, etc.

The success of the NooJ series of conferences lies not only in the opportunity for NooJ users and researchers in Linguistics and in Computational Linguistics to meet and share their experience as developers, researchers and teachers, but also to attend special tutorials helping to build

NLP applications using NooJ as well as to present and discover the recent developments of NooJ itself.

Overall, the given book includes 42 abstracts and covers different aspects of NooJ usage, providing us with interesting and multidimensional views, ideas and relevant information that, hopefully, will be helpful for your future work. We would like to thank the presenters for their willingness to share their latest research and ideas, as well as all the experts for their valuable advice. With their efforts, this conference would not be possible. We hope you will enjoy Minsk.

TRANSLATING ARABIC ACTIVE SENTENCES INTO ENGLISH PASSIVE SENTENCES USING NOOJ PLATFORM

H. Ben Ali[1], A. Rhazi[2] , M. Aouini[3]
[1] University of Monastir, Tunisia;
[2] University Cadi Ayyad, Marrakech, Morocco;
[3] University of Franche-Comté, Besançon, France

The present paper will focus on the problems resulting from the translation of Arabic active sentences into English passive sentences. This translation may lead to many difficulties resulting from the disparities between the source language (Arabic) and the target language (English) at the syntactic level. For example, in Arabic we make the difference between masculine and feminine. In Arabic, the active verb follows the following pattern: **فَعَلَ، يَفْعَ لَ** for singular as well as dual and plural masculine and فَعَلَتْ، تَفْـعَ ل for singular feminine, dual feminine as well as plural feminine and some irregular masculine plurals. Whereas, in English we don't make this difference and the verb pattern in the passive form would be as follows: **to be + past participle**. This way, the verb would be dependent on the subject, which originally the Direct Object of the active sentence. So, in English passive sentences where subjects can be replaced by she/he/it, the verb which is in the simple present, for example, would be as follows: **am/is/are + past participle**. With subjects that can be replaced with you/we/they, the verb which is in the simple present would be as follows: **are + past participle**.

These problems and others must be taken in consideration when translating Arabic active sentences into English passive sentences [1–10]. Word order, for instance, should be taken into account. For that reason, structural adjustments should be introduced in translation if natural equivalence is to be achieved. It is noticed that Arabic tends to use less passive than English and, furthermore, does not have a natural method of expressing the agent in a passive sentence. Let's take this example: the active sentence structure أكل الولد التفاحة Verb + Subject + Direct Object is transformed into passive as follows: أ كِ لَت التفاحة . As you notice, the agent of the action which is الولد is not kept into the passive sentence as it is considered rather unnatural. Whereas in English, the agent of the action is not avoided only in case the agent is a personal pronoun. Otherwise, the agent of the action should be kept in an English passive sentence. The Arabic active sentence أكل الولد التفاحة is translated into English passive as follows: **the apple was eaten by the boy**. In this translated sentence, you can notice that the pattern X was done by Y is kept; unlike the Arabic structure X was done.

Applying grammatical rules (order, structure, tense …), the translation of this Arabic active sentence "أكل الولد التفاحة " into English passive sentence gives the following **"the apple was eaten by the boy"**.

References

1. Silberztein, M. La formalisation des langues: approche de NooJ / M. Silberztein // ISTE Eds. – London, 2015. – 426 p.

2. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2003. – Mode of access : http://www.nooj4nlp.net. – Date of access : 17.02.2015.

3. Trouilleux, F. Un analyse de surface non déterministe pour le français / F. Trouilleux // Actes de TALN 2009 (Traitement automatique des langues). – Senlis : ATALA, 2009.

4. Ben Hamadou, A. Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform / A. Ben Hamadou, O. Piton, H. Fehri // Proceedings of the NooJ 2010 Intern. Conf. – University of Thrace Ed., Greece, 2011. – P. 192–202.

5. Ben Hamadou A. Recognition and Arabic-French translation of named entities: case of the sport places / A. Ben Hamadou, O. Piton, H. Fehri //Finite-State Language Engineering with NooJ: Selected Papers from the NooJ 2009 Intern. Conf. – Sfax : Centre de publication Universitaire, 2010. – P. 271–284.

6. Lafferty, J. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data / J. Lafferty, A. McCallum, F. Pereira // Proceedings of the Eighteenth Intern. Conf. on Machine Learning (ICML-2001). – Morgan Kaufmann Publishers Inc. San Francisco, USA, 2001. – P. 282–289.

7. Finkel, J.R. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling / J.R. Finkel, T. Grenager, C. Manning // Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005). – Association for Computational Linguistics Stroudsburg, USA, 2005. – P. 363–370.

8. Gross, M. The Construction of Local Grammars / M. Gross // Finite-State Language Processing. – Cambridge : MIT Press, 1997. – P. 329–354.

9. Mesfar, S. Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard / S. Mesfar. – Université de Franche-Comté, France, 2008. – 464 p.

10. Fehri, H. Reconnaissance automatique des entités nommées arabes et leurs traduction vers le français / H. Fehri. – Tunis, 2012.

SEMANTIC TAGS FOR NOOJ RUSSIAN DICTIONARY

V. Benet
INALCO, Paris, France
e-mail: vincent.benet@inalco.fr

The linguistics resources for Russian language include morphological information but only few semantical basic annotation related to morphology such as verbs of motion, animate and inanimate features.

The aim of this communication is to present how semantic tags were included to the main Russian dictionary and the way to work with them.

A HIERARCHY OF SEMANTIC LABELS FOR SPANISH DICTIONARIES

X. Blanco
Universitat Autònoma de Barcelona, Bellaterra, Spain
e-mail: Xavier.Blanco@uab.cat

In the frame of the Labelsem project (FFI-2013-44185-P funded by the Spanish Ministerio de Economía y Competitividad), we are elaborating a comprehensive hierarchy of semantic labels for dictionaries of Spanish. A "semantic label" of a dictionary's entry is a lexical unit (simple or complex) or (more rarely) a syntagm that corresponds to the genus proximum of the entry in question. For example, the semantic label for *arañazo*, *corte* or *herida* would be LESIÓN ("injury" "lesion").

The semantic label constitutes, therefore, the syntactic head of the entry's formal definition. It must be noted that the definiendum is nearly always not the entry form (the lemma) but a propositional form that includes the semantic actants of the lemma. For instance, we do not define *herida* (*wound*) but *herida de X por parte de Y en Z con W* (*wound of W from Y in Z with W*) (X and Y being animates, Z a part of X and W a physical object). Other propositional forms accepted by *herida* would be the object of diferent definitions.

The fact that these labels are actual linguistic signs of Spanish and not a metalinguistic device implies that their regular semantic, syntactic and restricted lexical coocurrence with the definiendum can be and must be controlled before they can be attributed to a lemma. This control plays a key role in the elaboration of the hierarchy as it constitutes the central criterion for the attribution of labels. It is a distinctive trait of our hierarchy since most sets of semantic labels are actually made up of metalinguistic entities. Moreover, by labeling in this manner, we obtain both a minimal paraphrase of the lemma's signified and a syntactical substitute in any context [1].

If this moment our hierarchy comprises approximately 700 labels. The total number of labels in the hierarchy can not be fixed in advance since we can not arbitrarily restrict ourselves to some sets of hyperonyms (the hierarchy is inductively build, we need the genus or next kind for each lemma in the dictionary). Of course the usual inheritance mechanism can be used to form quantitatively manipulable sets of lemmas. The label with the greatest semantic extension is ALGO ("something") , followed by HECHO ("fact") and ENTIDAD ("entity"). A HECHO [2] is always a semantic predicate, while an ENTIDAD can be a semantic object or a predicate (quasi-predicate). The subordinate labels of HECHO are around fifteen: ACCIÓN ("action"), ACTIVIDAD ("activity"), ACTITUD ("attitude")... The subordinate labels of ENTIDAD are around twenty: ACUMULACIÓN ("accumulation"),

CREACIÓN (“creation”), CONJUNTO (“set”)... We will discuss the semantic labels of a sample of the dictionary containing 1,000 lemmas corresponding to HECHOS and 1,000 lemmas corresponding to ENTIDADES.

It is worth to emphasize that our hierarchy of semantic labels is language dependent. As a result, it can not be directly used for translation or for multilingual search operations. However, different mechanism of connections or equivalences between hierarchies can be proposed in order to consider translinguistic applications. We will examine the relationships between our hierarchy and the French hierarchy upon which our work is based [3–6].

References

1. Blanco, X. Les étiquettes sémantiques comme genre prochain: le cas des verbes / X. Blanco // Verbum. – Paris, 2007. – Vol. XXIX (1–2). – P. 113–125.

2. Blanco, X. Etiquetas semánticas de hecho como género próximo en la definición lexicográfica / X. Blanco // Quaderns de filología. Estudis lingüístics. – València, 2010. – Vol. 15. – P. 159–178.

3. Mel’čuk, I. Semantics. From Meaning to Text / I. Mel’čuk. – Amsterdam / Philadelphia : John Benjamins Publishing Company, 2012. – 436 p.

4. Mel’čuk, I. Introduction à la linguistique / I. Mel’čuk, J. Milićević. – Paris : Hermann, 2014. – Vol. 1. – 375 p.

5. Polguère, A. Étiquetage sémantique des lexies dans la base de données DiCo / A. Polguère // Traitement Automatique des Langues. – Paris, 2003. – Vol. 44(2). – P. 39–68.

6. Polguère, A. Classification sémantique des lexies fondée sur le paraphrasage / A. Polguère // Cahiers de lexicologie. – Paris, 2011. – Vol. 98. – P. 197–211.

A HYBRID APPROACH TO EXTRACTING AND ENCODING DISORDER MENTIONS FROM CLINICAL NOTES

M. Chernyshevich, V. Stankevitch
IHS Inc. / IHS Global Belarus, Minsk
e-mail: Marina.Chernyshevich@ihs.com

This paper describes the clinical disorder identification and encoding system developed by IHS R&D Belarus team to participate in the international shared task organized by the Conference on Semantic Evaluation Exercises (SemEval-2015) [1]. This task aims at the recognition of entities belonging to the disorders semantic group of the Unified Medical Language System (UMLS) [2] and normalization of these entities to a specific UMLS Concept Unique Identifier (CUI).

The proposed system consists of two components: a CRF-based approach with a rich set of lexical, syntactic and semantic features to recognize disorder entities and empirical ranking to encode disorders to UMLS CUIs. We formulated disorder mention identification as a sequence labeling problem at token level and used Conditional Random Fields (CRF) [3]. To facilitate feature generation for supervised CRF learning, sentences were pre-processed with IHS Goldfire Linguistic Processor that performs the following operations: word splitting, part-of-speech tagging, parsing, noun phrase extraction, semantic role labeling within expanded Subject-Action-Object (eSAO) [4]. We conducted several experiments with different tagging conventions and decided to use the ILO (Inside-Last-Outside) tagging scheme, where tag I represents the beginning and the inside token of an entity, L represents the last word of entity and O is not a member of a disorder structure. The traditional BIO (Begin-Inside-Outside) tagging scheme showed the classification accuracy lower by 5,5 %. The following list of features for CRF model was defined: token text, 5-token window, POS-tag, letter case, 3- and 4-letter n-grams starting and ending the token, out-of-domain word frequency, boolean value showing whether the word is a part of a longer noun phrase or not, semantic class (body part, process, units of measure, drug etc.), document section (the id of the section to which the token belongs) and a UMLS lookup-based two-level (word and phrase) feature. We propose a simple sieve-based algorithm that applies the following string matching rules to select candidate UMLS concepts for a disorder entity (each rule assigns the score of confidence). Exact match: disorder and UMLS concept contain exactly the same extent text, excluding modifiers and determiners, with the same word order. Relaxed match: all informative words (excluding preposition, conjunctions, stop words etc.) from disorder are included in the UMLS concept. Partial match: at least

one informative word from disorder is included in the UMLS concept. Variants match: possible variants are generated for the disorder entity using synonyms, corrections and suggestions from our inhouse autocorrection and autocompletion. All found candidate UMLS concepts were ranked by a set of empirical parameters: score of confidence, TF-IDF of the intersecting words, total number of disorder variants in the UMLS having the CUI, number of times the UMLS concepts was already mentioned in this document, number of occurrences of the UMLS concept in the unlabelled corpus. The top ranked UMLS concepts were selected as the system's output. If some concepts have the same ranking score, the first one by CUI number was selected.

Evaluation on the test data set showed that our system achieved an F-measure of 0,898 for entity recognition (rank 9 out of 40), an F-measure of 0,794 for UMLS CUI categorization (rank 19 out of 40) and a combined score of 0,690 (rank 17 out of 40). Additional research to evaluate the feasibility of solving this kind of problems with rule-based engines such as Nooj is to be done in the nearest future.

References

1. SemEval-2014 Task 4: Aspect Based Sentiment Analysis / S. Pradhan [et al.] // Proceedings of the 8th Intern. Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland. – Dublin, 2014. – P. 27–35.

2. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology / O. Bodenreider // Nucleic Acids Research. – 2004. – Vol. 32. – P. 267–270.

3. Lafferty, J. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data / J. Lafferty, A. McCallum, F. Pereira // Proceedings of the Eighteenth Intern. Conf. on Machine Learning (ICML-2001). – San Francisco, 2001. – P. 282–289.

4. Todhunter, J. System and method for automatic semantic labeling of natural language texts / J. Todhunter, I. Sovpel, D. Pastanohau // U.S. Patent 8 583 422. – November 12, 2013.

MIXED PROLOG AND NOOJ APPROACH IN JAPANESE BENEFACTIVE CONSTRUCTIONS

V. Collec Clerc
Laboratoire d'informatique fondamentale de Marseille,
Université Aix Marseille, France
e-mail: valerie.collecclerc@valtal.fr

This paper presents our research on analysis and generation of valid Japanese sentences in the context of polite donatory situations of communication. We have used NooJ [1] to extract the relevant constructions. We have resorted to the logic programming language Prolog to generate these constructions. This paper points out the links between NooJ and Prolog.

Our work mainly deals with the interpersonal language, which requires calibrating polite forms to social ranks [2–7]. We have focused on the morphosyntactic impact of the benefactive situations which are utterances of giving/receiving acts. The choices of verbs, modals and auxiliaries depend on the relative roles of addressees and the referred persons.

A NooJ dictionary links lemmas to types, subtypes and attribute values.

兄,N+Humble+KANA=あに+ROMAJI=ani+Hum+EN=
elder brother+DE=älterer Bruder

In Prolog, data are expressed by a set of facts: woman(由美子), and enhanced by rules: human(X) :- woman(X). The context of utterance must be known for correct parsing or generation. It is called universe of discourse. Interpersonal language requires knowing the relationships between communication partners. We must devise a NooJ dictionary for named entities which meets the context (university, company, etc.).

由美子,ENAM+human+SEX=f+student+UNIVERSITY=
東京大学+ROMAJI=yumiko

Our system uses morphological, syntactic and pragmatic rules. Morphological rules express verb and adjective inflections. These rules are equivalent in NooJ and Prolog. We have created NooJ grammars to recognise syntactic patterns which comprise benefactive auxiliaries. Let us examine this sentence: *romaji de jusho wo kaitekudasaimasen ka?* (Couldn't you write the address in Latin characters for me?) *Kudaisamasenka* is the negative interrogative form of the benefactive auxiliary *kudaru*. We have built grammatical rules in Prolog to generate equivalent sentences.

Pragmatic rules are written in Prolog. They express the choice of words, modals, constructions according to the relative role of the components of the utterance. For instance, from a basic sentence like *Sakubun/wo/miru* (essay/case particle/see) which involves a situation in which a student makes

a request to the teacher, some modifications lead gradually to the final sentence *Sakubun wo goran ni natte kudasaimasen ka?* (Could you please have a look at my essay?) This first modification takes the donatory situation into account by turning the verb *miru* into its *te*-form and adding the benefactive auxiliary *kureru* to it: *Sakubun wo mittekureru*. The second modification applies the hierarchical link. When a student addresses to a teacher, the former must use honorific verbs. Here *miru* and *kureru* are replaced with their honorific suppletive forms, respectively *goran ni naru* and *kudasaru*: *Sakubun wo goran ni natte kudasaru*. The third modification considers the relationship between the speaker and the listener by adding the polite suffix *masu* to the verb *kudasaru*: *Sakubun wo goran ni natte kudasaru*.

This kind of rules must have its counterpart in NooJ. Therefore, we have inserted the required data to recognise standard social roles like student and professor in our NooJ dictionary for named entities. We have also designed an additional dictionary to identify the donatory verbs, their valences and the associated auxiliaries. NooJ grammars use the constraints on those data for instance the subject and the object of the utterance belong to the same university.

References

1. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2003. – Mode of access : http://www.nooj4nlp.net. – Date of access : 11.12.2014.

2. Kozai, S. Exceptionally exceptional expressions in Japanese. Japanese Linguistics 101 / S. Kozai // 6th Intern. Conf. on Languages, E-Learning and Romanian Studies. – Isle of Marstrand, Sweden, 2011. – 11 p.

3. Nariyama, S. Annotating Honorifics Denoting Social Ranking of Referents / S. Nariyama, H. Nakaiwa, M. Siegel // Proceedings of the 6th Intern. Workshop on Linguistically Interpreted Corpora (LINC-2005). – Jeju Island, Korea, 2005. – P. 91–100.

4. Sugimura, R. Japanese honorifics and Situation Semantics / R. Sugimura // Proceedings of 11th Intern. Conf. on Computational Linguistics. – IKS Bonn, Germany, 1986. – P. 507–510.

5. Keigo wo totonoeru (Treatment of the honorific form) / S. Tanaka [et al.] // Asakura Nihongo Shin-Kôza. – Asakura Shoten, Tokyo, 1983. – Vol. 5.

6. Terrya, K. Interpersonal grammar of Japanese. A systemic functional grammar of Japanese / K. Terrya. – London : NY : Continuum, 2007. – Vol 2. – P. 135–205.

7. Japanese Linguistics: Critical Concepts / N. Tsujimura (ed.). – Vol. II : Syntax and Semantics; vol. III : Pragmatics, Sociolinguistics and language contact. – Routledge Library of Modern Japan, Routledge, 2005.

SEMI-AUTOMATIC INDEXING AND PARSING INFORMATION ON THE WEB WITH NOOJ

M.P. di Buono
University of Salerno, Fisciano (SA), Italy
e-mail: mdibuono@unisa.it

Due to the large amount of data available on the Web, indexing information represents a crucial step in order to guarantee fast and accurate Information Retrieval (IR) [1]. Indexing content allows to find relevant documents on the basis of a user's query. Numerous researches discuss the use of automated indexing, considered faster and cheaper than manual systems [2–3]. However, using algorithms, in order to produce the index, entails low precision, due to the indexing of common ALUs or sentences, low recall, caused by the presence of synonyms, and generic results, arised from the use of too broad or too narrow terms [4].

In this paper we propose a system based on NooJ for developing a search engine able both to process online documents starting from a natural language query and to return information to a user.

IR systems usually are based on inverted text index and they process each document separately in order to retrieve terms which appear in free-text query. This procedure may cause overlapping in results and decreasing the positive predictive value.

In order to develop our system we use a software to automate the routine allowing to use NooJ and its Linguistic Resources (LRs) for analyzing the user request.

The system workflow is also based on a representation model applied both to user query and to documents, and on a match between those two elements.

The representation model proposed is developed on a semantic annotation process to guarantee the interoperability between metadata. In fact, queries may include some restrictions on metadata, such as URL, domain, etc., which are typically different for each document. In order to support these queries, the representation model uses ontological schema to map Atomic Linguistic Units with concepts for avoiding overlapping and indexing shared content just once. Semantic association is also used to infer Boolean relationship between elements in a free-text queries and relative meta-data. We may define our system as an architecture based both on document-term matrix and n-gram index. This index data structure is developed on trees consisting of tagged Atomic Linguistic Units and also on grammars containing rules of sentence recursive structures and co-occurrences.

We test our system on a corpus dumped from the Italian Wikipedia Database and we apply a standard zig-zag join as query evaluation system, since

an indexed nested loop join can be way faster and exploits a temporary or permanent index on the inner input's join attribute.

References

1. Anderson, J.D. The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part 1&2 (Research and the nature of human indexing) / J.D. Anderson, J. Perez-Carballo // Information processing and management. – 2001. – Vol. 37(2). – P. 231–277.

2. Seth, M.A. Notes on automatic indexing / M.A. Seth [Electronic resource]. – 2004. – Mode of access : http://taxonomist.tripod.com/indexing/autoindex.html. – Date of access : 28.01.2015.

3. Tulic, M. Automatic indexing / M. Tulic [Electronic resource]. – 2005. – Mode of access : www.anindexer.com. – Date of access : 28.01.2015.

4. Hjorland, B. Semantics and knowledge organization / B. Hjorland // Annual Review of Information Science and Technology. – 2007. – Vol. 41. – P. 367–405.

THE ANNOTATION OF COMPOUND SUFFIXATION STRUCTURE OF QUECHUA VERBS

M. Duran
UFTC, Paris, France
e-mail: duran_maximiliano@yahoo.fr

In the Quechua language we only find around 1500 simple verbs. In NooJ (Silberztein 2003) we can find several strategies to increase this verb lexicon by applying grammars to generate new verbs by derivation [1–5]. In this way we obtain a large set of new verbs which are known as composed verbs.

Let us take the verb *llamkay* (to work), which is formed by the verbal lemma *llamka-* and the infinitive suffix *–y*. Interposing the suffix *–isi-* gives us the new derived verb *llamka-isi-y* (to help someone to work). Adding to the same lemma the suffix *–chka-* gives us a new verb *llamka-chka-y* (to keep working). The combination of these suffixes produces a new suffix *-chkaisi-* which could be agglutinated to the same lemma in order to obtain the new verb *llamka-chka-isi-y* (to keep helping someone to work). The permutated combination **–isichka-* is not grammatically correct. There are 26 interposable suffixes in Quechua which are capable of generating new verbs. They are interposed between the lemma and the ending of the conjugation. The combinatorial of two of them give us 676 possibilities of new verb generators, nevertheless only 292 are grammatically correct. Furthermore, the Quechua grammar allows the agglutination of up to four of them in a defined order. We have programmed several hundreds of morpho-syntactic NooJ grammars to generate all the correct ones. For instance for the case of one suffix we have the grammar:

VERSVIP1=<B>(chiy/FACT|chkay/PROG|ikachay/DISP|ikachiy/POLI| ikariy/DIS1|ikuy/COURT|isiy/COLL|kapuy/RAS|kuy/RFLX|llay/POL|muy/MU| nay/NOP|nayay/ENV|pay/REP|payay/FREQ|puy/APT|ray/PASS|rayay/DUR| riy/RI|rpariy/ASUR|rquy/PAPT|ruy/PASS2|sqay/PPA|tamuy/TAMU); For one-dimension derivations. Here FACT stands for factitive, PROG for progressive, etc.

A partial view of the grammar for the bi-dimensional combinations appears like this:

VERSVIP2=<B> (chichka/FACT+PROG | chiikachi/ FACT+ POLI |chiiku/ FACT+ COURT |chiisi/ FACT+ COLL |chiikapu/ FACT+ SOIN |chiku/ FACT+AUBE |chilla/ FACT+ POLI |chimu/ FACT + MU |chikamu/ FACT+ AAR |chkaisi/ PROG+ COLL |chkara/ PROG+ PASS …);

As a result and parallel to this work, we have been elaborating, a dictionary of composed, multi-lingual verbs, including their Spanish and French translations which will be added to our verb lexicon. We present about 500 verb

entries which look like the following sample:

saqey,V+FLX=VSIPSPP1+DRV=VERSVIP1+SP=dejar+FR=laisser.

We have applied both the dictionary and these grammars for automatically annotate a quechua text, built of a collection of eight quechua tales. We have obtained near 90 % of successful matches, 6 % of partial matches and 4 % are incorrect matches.

References

1. Bogacki, K. Derivational structure of Polish Verbs and the Expansion of the Dictionary / K. Bogacki, E. Gwiazdecka // Automatic Processing of Various Levels of Linguistic Phenomena. Selected Papers from the NooJ 2011 Intern. Conf. – Cambridge Scholars Publishing, New Castle upon Tyne, 2012. – P. 50–62.

2. Silberztein, M. Syntactic parsing with NooJ / M. Silberztein // Proceedings of the NooJ 2009 Intern. Conf. and Workshop. – Sfax : Centre de Publication Universitaire, Tunisie, 2010. – P. 177–190.

3. Silberztein, M. Automatic Transformational Analysis and Generation / M. Silberztein // Proceedings of the 2010 Intern. NooJ Conf. – Democritus University of Thrace, Komotini, Greece, 2010. – P. 221–231.

4. Silberztein, M. VariableUnification in NooJ V3 / M. Silberztein // Automatic Processing of Various Levels of Linguistic Phenomena. Selected Papers from the NooJ 2011 Intern. Conf. – Cambridge Scholars Publishing, New Castle upon Tyne, 2012. – P. 50–62.

5. Vietri, S. The Annotation of the Predicate-Argument Structure of Transfer Noun / S. Vietri // Formalising Natural Languges wih NooJ. – Cambridge Scholars Publishing, New Castle upon Tyne, 2013. – P. 88–99.

PROCESSING OF PUBLICATION REFERENCES IN BELARUSIAN AND RUSSIAN ELECTRONIC TEXTS

D. Dzenisiuk, Yu. Hetsevich
United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: d.denissyuk@gmail.com

Entity recognition is a subtask of information extraction that serves to locate and classification of elements in text into pre-defined categories such as the names of persons, organizations, expressions of time, quantities, monetary values, percentages, references etc. [1]. Information which one can get from the references is useful for libraries, publishing houses or research institutions, since it has information about authors or editors, publishers or universities, name and location of the publishing house etc.

The aim of the research is to develop NooJ grammars [2] for processing of references in Belarusian and Russian texts. First step is to develop a set of queries which allow retrieving information about author and co-authors of the publication, name of the publication, location and year, overall number of pages or certain page of the book or the journal. Second step is to combine those queries into syntactic grammars, which can be used on large corpora.

By the beginning of the conference is planned to improve grammars for the Belarusian and Russian languages, and also to make grammar for processing of references designed according to the standard of the Chicago Manual of Style.

References

1. Get a publication reference! [Electronic resource]. – 2014. – Mode of access : http://corpus.by/publicationReference. – Date of access : 10.12.2014.

2. Nooj Manual // NooJ: A Linguistic Development Environment [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 10.12.2014.

STUDY AND RESOLUTION OF ARABIC LEXICAL AMBIGUITY THROUGH THE TRANSDUCTION ON TEXT AUTOMATON

N. Ghezaiel[1], K. Haddar[2]
[1] Higher Institute of Computer and Communication Technologies of Hammam Sousse, Miracl Laboratory,Tunisia;
[2] University of Sfax, Faculty of Sciences of Sfax, Miracl Laboratory, Tunisia
e-mail: ghezaielnadia.ing@gmail.com

Arabic sentence is characterized by great variability in words order. In general, in the Arabic language, we put at the beginning of the sentence the most attractive word (noun or verb) and at the end we put the richest term to keep the sentence's meaning. This variability in the order of words causes syntactic ambiguities. So the grammar should contain all possible combinations of inversion rules of words order in the sentence. Note that the Arabic sentence can be either verbal or nominal. Arabic words can be ambiguous at the lexical level. For example, the word "ذهب" "Dhahab" can refer to the name "gold" in English, or the verb "go". The word "كتب" "katab" can belong to several grammatical categories: verb or name. The meaning of this word will be very different depending on its class name = Koutoub "books" verb = "write" is in passive voice: "Koutiba" or active voice "Kataba". To remove the ambiguity, some particles, some types of verbs and some syntactic constraints can be exploited. Indeed, there are particles that sub-categorize verbs such as particles of negation and others sub-categorize names like prepositions.

Finite automaton and particularly the transducers are increasingly used in the field of automatic language processing. Indeed, thanks to the transducers several local linguistic phenomena (e. g., recognition of named entities, morphological analysis) are treated appropriately. In addition, the use of transducers cascades allowed for robust parsing and with high accuracy on corpus. Note that a cascade of transducers is a series of transducers applied to text in a specific order to convert or extract patterns.

It is in this context that seen against this paper. Our approach is based on the application of cascade on a text automaton to simplify the removal of ambiguities. This technique requires a certain order in the transducer passage on the text automaton respecting the principle of elimination. Indeed, it first applies the most obvious and intuitive rules until arriving at the less one. The transducers can specify the lexical and contextual rules of the Arabic language allowing the lifting of ambiguities. The specified rules can either remove paths representing morpho-syntactic ambiguities or enrich the text automaton syntactically or semantically by new roads.

Regarding the text automaton, it is a way to express all possible morpho-syntactic tags and existing words in a sentence. Thus, each sentence of the text

can be represented by an automaton whose paths express all possible lexical interpretations. These different interpretations have different inputs presented in the designed dictionaries. The text automaton can be an efficient and visual way to show the morpho-syntactic's labeling ambiguities. Note that the application of a transducer on a text automaton provides a new text automaton as outputs, enriched by new paths witch are generated by the grammar. Some treatments (such as cutting into chunks or recognition of named entities), which identify and label occurrences of the forms described in the grammar, it is necessary to distribute the labeling of sequences recognized on whole path traveled during the recognition of the segment by the grammar. This type of process can be very tedious to write with the formalism of transducers because; it can specify the lexical and contextual rules of the Arabic language for the resolution of ambiguity.

For the experiment, we establish a set of lexical rules and constraints that are applied to a fairly representative corpus of text automata. The results obtained are satisfactory.

USING NOOJ FOR THE PROCESSING OF SATELLITE DATA

Yu. Hetsevich, J. Borodina
United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: yury.hetsevich@gmail.com

The paper describes the processing of satellite telemetry data and its conversion from reduced form into full orthographically correct word sequences for the further use in text-to-speech synthesis or other applications.

The purpose of a telemetry system is to reliably and transparently convey measurement information from a remotely located data generating source to users located in space or on Earth. Typically, data generators are scientific sensors, science housekeeping sensors, engineering sensors and other subsystems on-board a spacecraft [1]. By satellite telemetry data in our paper we also mean results of measurements made by nanosatellites from space and received on Earth by special software.

For the research we use text corpus which was created with the data from satellite telemetry converter software. The data is provided in the form of abbreviations, numbers and measurement units along with orthographical words (e.g. *Voltage of 5V system is 4,904 [V]; Temperature of the 145 MHz TX: 16 °C*).

Main difficulty which arises during the work with quantitative expressions with measurement units (QEMU) in relation to telemetry is language. Telemetry data are mostly collected in English as an international language for science, yet Belarusian national space tradition requires this information to be available in one of the national languages of Belarus: Belarusian or Russian. Therefore our task is not only to convert QEMU from reduced form into the full one, but also to automatically translate data from English to Belarusian. The grammar made in NooJ [2] is designed to be language independent and self-sufficient, i.e. there are no dictionaries applied.

The grammar takes as an input text sequences like *0,9708 A* and transforms first numerical part, then measurement units into Belarusian phrase: *нуль цэлых дзевяць тысяч семсот восем дзесяцітысячных ампера* ‘zero point nine thousand seven hundred and eight ten-thousandths of an ampere’. Due to the fact that Belarusian and Russian languages are both synthetic, declension paradigm of this numeral will be very complex, but our grammar can handle it.

This work continues the work which was presented in the previous NooJ conference in 2014 held in Sassari, Italy.

References

1. Synchronization, T.C. Report Concerning Space Data System Standards. Channel Coding – Summary of Concept and Rationale / T.C. Synchronization // Green Book. – November, 2012.

2. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2003. – Mode of access : http://www.nooj4nlp.net. – Date of access : 21.12.2014.

GRAMMARS FOR THE SENTENCE INTO PHRASE SEGMENTATION: PUNCTUATION LEVEL

Yu. Hetsevich, T. Okrut, B. Lobanov
United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: yury.hetsevich@gmail.com

This paper deals with, so called, punctuational phrases in the sentences and their intonation type marking in Belarusian electronic texts. Such marking allows to implement the algorithm of intonationally coloured (expressive) synthetic speech and to avoid its monotony.

There are 4 main categories of intonation types: finality (P), non-finality (C), interrogation (Q) and exclamation (E). The intonation types, constituting these categories, are designated by the category symbol (P, C, Q or E) and their subtypes marked the proper index (1-n).

The indicating model works as follows:

<PHRASE TYPE="C1">Амаль забылася</PHRASE> і <PHRASE TYPE="C7">здарэнне з лазняй</PHRASE>, а <PHRASE TYPE="C1"> потым </PHRASE> і <PHRASE TYPE="P4">лазню злізаў Дняпро</PHRASE>. <PHRASE TYPE="C7">Быццам і не было ні людзей</PHRASE>, ні <PHRASE TYPE="P4">закуранай нізкай будыніны на беразе</PHRASE>.

Each phrase in these sentences is tagged with an intonation index, namely: "C1" – non-final "i"-intonation with "i" (a Belarusian co-ordinating conjunction meaning "and") stating the end of the phrase, "C7" – non-final intonation of "comma staying before a co-ordinating conjunction" with a comma stating the end of phrase, "P4" – "full stop"– intonation with a full stop stating the end of phrase.

When marking intonation type of a phrase, not only punctuation marks are taken into account but also the nearest context in a text. For example, the intonation type "Q12" corresponds to each even phrase in a line of consecutive "Q1"– phrases – interrogative phrases containing an interrogative word:

<PHRASE TYPE="Q11">Як так атрымалася</PHRASE>? <PHRASE TYPE="Q12">Чаму ты мне нічога не сказаў</PHRASE>?

The algorithm developed represents the initial stage of prosodic processing in a speech synthesizer, the following step will be to provide segmentation of punctuational phrases into syntactic phrases.

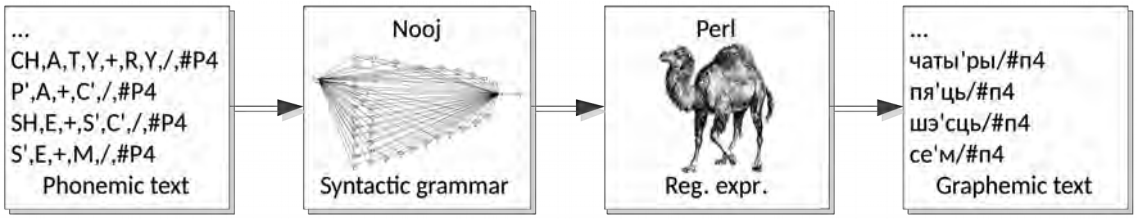
GRAMMARS FOR MAKING WRITTEN ORTHOGRAPHIC WORDS FROM TRANSCRIBED SPOKEN LANGUAGE

A. Hiuntar, V. Zahariev
United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: lena205593@gmail.com, zahariev@bsuir.by

There are two main problems for natural language processing on the transition step between morphological and phonetic levels of speech. The first one is a transformation from signed words to phonetically transcription for further processing with code signals corresponding to phonetic units [1]. This task is very common for example in text-to-speech synthesis systems [2]. The second one is an inverse problem: building of written orthographic words from transcribed spoken language. This task is very important item within the framework of automatic speech recognition (ASR) systems. In large vocabulary ASR systems, it is hard if not impossible to train separate statistical models for all words. In such systems, words are described as sequences of phonemes in a pronunciation lexicon, and statistical modeling is applied to phonemic units. In such case the problem of correct phoneme to grapheme transformation has a great significance.

In our work we suggest solution of this problem based on Nooj framework [3]. A key feature of our finding is the use of syntactic level grammar for processing of phonetic units that are in fact atomic linguistic units of morphological, not syntactic level of speech. This fact allows us to build a more flexible model for phoneme-to-grapheme conversion, given the great opportunity of syntactic grammars, within the constraints imposed by the form of the incoming flow of phonetic units from the ASR speech analysis module. A higher level of abstraction, given by syntactic grammars, makes it possible to handle blank positions and pauses in the phonetic text, which is actually difficult to achieve, through the using of morphological grammars. Since its use implies a mandatory meaningful units, which can't be, for example, a space or any absence of the sign.

The general sequence of phonemic processing text as follows (fig.). Source phonemic (or allophonic) text derived from the resolver module of the speech signal is converted into a text object of Nooj system. This text is treated with a prepared set of syntactic level grammars. With these grammars performed a linguistic analysis of the text, and the results are exported from Nooj in XML format. Next, using a special Perl script is processing the given file, using a set of regular expression searching elements that Nooj recognized as grapheme units. On the output we get a graphemic text.



General processing scheme

Construction of grammars based on rules-based phoneme-to-grapheme conversion, both for basic variants of these types of transformations and complex options take into account the right and left contexts are considered in this report. Study and analysis of performance data grammars in terms of the number of errors, computational complexity and speed of the proposed algorithms convert the corresponding language resources are made. Practical results of this work we are going to use in natural language dialog module within mobile robot.

References

1. Dutoit, T. Applied Signal Processing: A Matlab-based Proof of Concept / T. Dutoit, F. Marques. – Springer Science, Buisness Media, LLC, 2009. – 456 p.

2. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 03.01.2015.

3. Transcription Generator [Electronic resource]. – 2013. – Mode of access : http://corpus.by/transcriptionGenerator. – Date of acces : 03.01.2015.

LANGUAGE MODELLING FOR ROBOTS-HUMAN INTERACTION

L. Kaigorodova[1], Yu. Hetsevich[1], K. Nikalaenka[2],
R. Prakapovich[1], S. Gerasuto[1], U. Sychou[1]
[1] United Institute of Informatics Problems of the NAS of Belarus, Minsk;
[2] Belarusian State University of Informatics and Radioelectronics, Minsk
e-mail: lesia.piatrouskaya@gmail.com; yury.hetsevich@gmail.com;
anak247@gmail.com; contacts@robotics.by;
rprakapovich@robotics.by; vsychyov@robotics.by

This work is the start for further design of language model for robots-human interaction. The goal is to interact with some number of robots in order to make them perform commands. With **NooJ** [1] this model can be designed in much easier way compared to other tools. The idea is to design the language that would be common and close to every-day language of the humans and that it would be able for machines to 'understand' it. Further design will be dedicated to replaying the model that has already been designed and the new data which is the new possible language constructions, phrases, linguistic units, etc. that can be expected from humans in order to interact with machines in their natural way.

At the start stage of the work we use deep syntactic analysis to get the model that is as simple as possible and yet far from underfitting the real model. We will use such concepts as *"Subject"*, *"Action"*, *"Object"* and *"Features"*. Using **NooJ Syntactic Grammar** we design graph model for combining all these concepts and linguistic units that will refer to them. Eventually we perform play-out routine to generate dictionary for robots using **NooJ Dictionary**. Some units from this dictionary will look like:

Робат_Віцебск прынясі лыжку, GUID=R1+Action=take+Object=spoon

Робат_Гродна едзь на_зарадку, GUID=R2+Action=Go+Charge

Робат_Брэст выконвай паварот направа,
GUID=R3+Action=turn+Features=right

Робат_Брэст выконвай паварот налева,
GUID=R3+Action=turn+Features=left

"Subject" (GUID in our example) refers to a robot's name. *"Action"* refers to an action to be performed by robots that is usually represented by a verb. *"Object"* represents a target of the action. And *"Features"* is an add-on to specify *"Object"* or *"Action"*.

Using such kind of concepts, which are natural for humans, and **NooJ** tools we can generate Language Model for robots-human interaction.

References

1. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 11.03.2015.

TRANSLATING SPACIAL AND TEMPORAL DEIXIS IN NEAR LANGUAGES: A COMPARATIVE CLASSIFICATION APPROACH WITH NOOJ

M. Kirova
Belarusian State University, Minsk
e-mail: maryiakirova@gmail.com

The following paper considers existing translation problems of temporal and spacial deixis [1] in 2 near language pairs (Belarusian-Russian and Dutch-German), and crosswise as well. The main problem to be solved concerns primarily the near languages and lies in their deictic semantics. For instance, Russian "здесь" and Belarusian "тут" do not correspond 100 % reciprocally, though they both possess the core meaning "here". In some cases temporal and local deixis can be expressed by the same means, so that it can be problematic to distinguish them. In some languages larger varieties of deictic shades of meaning are represented, and these can be significantly different even in the nearest languages. Many native speakers and even professionals affiliated either with translation or teaching a language do not interpret such slight differences in a right way, because they do not always perceive them as false friends or ambiguous categories. Therefore, a deep analysis in this field is essential, as its results could contribute to a better understanding of the language phenomena and more accurate translation.

During preparatory steps the most frequent deictic markers, their semes and sememes in the above mentioned languages were found out in respective language corpora and listed. The analyzed items are primarily adverbs. By virtue of NooJ typical models for all 4 languages are created. As there are various ways to express spacial and temporal deixis (both by means of syntax and morphology), many features of NooJ are helpful while exploring and classifying these categories. For example, syntactic grammars and dictionaries can be produced of the existing data, so that all possible matches for the deixis types of the forenamed languages are visually presented and ready to use for translation or teaching purposes.

References

1. Levinson, S.C. Pragmatics / S.C. Levinson. – Cambridge : Cambridge University Press, 1983. – 438 p.

RECOGNIZING VERB-BASED CROATIAN IDIOMATIC MWUS

K. Kocijan, S. Librenjak
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, Zagreb, Croatia
e-mail: krkocijan@ffzg.hr, sara.librenjak@gmail.com

Croatian language has very rich phraseme structure, as described in [1–3], as well as many others [2, 6, 12]. The authors have analyzed 2500 Croatian idiomatic expressions as defined by the Croatian Phraseme Dictionary [2], and sorted them according to their syntactic properties. On that basis, five major syntactic groups of Croatian idioms were found, and NooJ grammars [8] were constructed accordingly. We were able to recognize five syntactic types of idiomatic expressions: 1. <A> <N> (noun phrase with an attribute or apposition); 2. <V> <dir_object> (verbal phrase with a direct object); 3. <V> <indir_object> (verbal phrase with the optional indirect object that can disrupt the syntactic structure); 4. <A/V> kao <N> (comparative structure, verb or adjective as noun); 5. fixed structure. The first four types required building syntactic NooJ grammars in order to be recognized in all their varieties (split cases, inversions), and the fifth one was added directly to the NooJ dictionary. In this paper we will more closely describe those idiomatic expressions that are verb based.

The set of NooJ grammars for detecting idioms in Croatian is trained on digitized corpus made from Croatian literally text in which all of the processed idioms can be found. Subsequently, finished grammars were tested on web based Croatian corpus sample [4], and compared with manually marked results. Some statistical data about distribution of idioms in Croatian texts will be presented as well. Except for detecting idioms and providing statistical data, this work can be helpful in machine (aided) translation of Croatian, since MWUs are typically harder to process in MT and require special attention. A few can be understood in direct translation to e.g. English (*biti na vrhu jezika – sth is at the tip of the toungue, osvojiti srce – win smbs heart, naoružan do zuba – armed to the teeth*), but more are either only similar, but not direct translation (*komu se steže grlo –> lit. the throat is being clenched = have lump in one's throat; nositi srce na dlanu –> lit. carry the heart in the palm of one's hand = wear the heart on one's sleeve*), or completely different (*grlom u jagode –> lit. rush with your throat to the strawberries = jump the gun; ruku na srce –> lit. hand to the heart = truth be told*).

As the phrasemes are rooted in the tradition of the language and the society from which they hail from, they need a special treatment in computational linguistics [7, 9–11]. With this tool, Croatian idioms can be successfully detected and matched with their translation in corresponding

language, which would eliminate awkward and completely wrong automated translations. Thus, this work seeks to aid not only the successful development of additional language resources for Croatian language, but a possible assistance in future work relating to machine assisted translation.

References

1. Matešić, J. Frazeološki rječnik hrvatskoga ili srpskog jezika / J. Matešić. – Zagreb : Školska knjiga, 1982.

2. Menac, A. Hrvatski frazeološki rječnik / A. Menac, Ž. Fink-Arsovski, R. Venturin. – Zagreb : Naklada Ljevak, 2003. – 800 p.

3. Menac-Mihalić, M. Hrvatski Dijalektni Frazemi S Antroponimom Kao Sastavnicom / M. Menac-Mihalić // Folia Onomastica Croatica. – 2007. – № 12/13. – P. 85–361.

4. Agić, Ž. The SETimes.HR Linguistically Annotated Corpus of Croatian / Ž. Agić, N. Ljubešić // Proceedings of the Ninth Intern. Conf. on Language Resources and Evaluation. – Reykjavik, 2014. – P. 27–1724.

5. Bekavac, B. A Generic Method for Multi Word Extraction from Wikipedia / B. Bekavac, M. Tadić // 30th Intern. Conf. on Information Technology Interfaces. – Dubrovnik, Croatia, 2008. – P. 68– 663.

6. HrMWELex – A MWE lexicon of Croatian extracted from a parsed gigacorpus. Language technologies / Ljubešić [et al.] // Proceedings of the 17th Intern. Multiconference Information Society. – Ljubljana, Slovenia, 2014. – P. 25–31.

7. Machonis, P.A. English Phrasal Verbs: from Lexicon-Grammar to Natural Language Processing / P.A. Machonis // Southern Journal of Linguistics. – 2010. – Vol. 34(1). – P. 21–48.

8. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2003. – Mode of access : http://www.nooj4nlp.net. – Date of access : 18.12.2014.

9. Gavriilidou, Z. Processing Greek Frozen Expressions with NooJ / Z. Gavriilidou, E. Papadopoulou, E. Chadjipapa // Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2011 Intern. Conf. – Cambridge Scholars Publishing, 2012. – P. 63–74.

10. Vietri, S. Transformations and Frozen Sentences / S. Vietri // Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2011 Intern. Conf. – Cambridge Scholars Publishing, 2012. – P. 166–181.

11. Machonis, P.A. Sorting NooJ out to take Multiword Expressions into account / P.A. Machonis // Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2011 Intern. Conf. – Cambridge Scholars Publishing, 2012. – P. 152–165.

12. Tadić, M. Finding Multiword Term Candidates in Croatian / M. Tadić, K. Šojat // Proceedings of IESL2003 Workshop. – Borovets, Bulgaria, 2003. – P. 102–107.

COMPARISON OF LEXICAL AND GRAMMATICAL BASE OF BELARUSIAN N-KORPUS WITH DICTIONARY PROPERTIES' DEFINITION FILE OF BELARUSIAN NOOJ MODULE

U. Koshchanka[1], Yu. Hetsevich[2], V. Varanovich[3], A. Tretyak[3]
[1] The Center for the Belarusian Culture, Language
and Literature researches of the NAS of Belarus, Minsk
e-mail: koshul@gmail.com;
[2] United Institute of Informatics Problems of the NAS of Belarus, Minsk;
[3] Belarusian State University, Minsk

In this work, we plan to compare the matching parts of the Belarusian N-korpus and dictionary properties' definition file of the Belarusian module in NooJ.

Belarusian N-korpus is the first publicly available general Belarusian language corpus [1]. The Belarusian N-korpus currently contains ~93 885 texts (~57 729 850 tokens) taken from fiction, newspapers, journals and on-line editions. The texts of the corpus are grammatically annotated and contain metatextual information. The grammar database is available under Open Database License (ODbL) v1.0. The corpus engine is available under GNU General Public License, Version 3.

The Belarusian module for NooJ has been created during the period of 2011–2012 [2]. Basic results have been presented at the 14th international NooJ conference in 2011 (Dubrovnik, Croatia). The Belarusian NooJ module consists of the following parts: Texts, Dictionary, Grammars, Samples, Projects. They were composed for a basic acquaintance with the Belarusian language. This project is highly popular, as it allows the flexibility to customize the basic functionality of NooJ by further replenishment of parts of the Belarusian module for automatic annotation of Belarusian texts.

References

1. Belarusian N-korpus [Electronic resource]. – 2012. – Mode of access : http://bnkorpus.info. – Date of access : 10.02.2015.

2. Belarusian module for NooJ / Y. Hetsevich [et al.] // NooJ web-site [Electronic resource]. – 2012. – Mode of access : http://www.nooj4nlp.net/ pages/belarusian.html. – Date of access : 10.02.2015.

SEMANTIC TAGGING OF THE SENTIMENT WORDS WITH NOOJ

D. Le Pesant
MoDyCo (CNRS & Université Paris Ouest Nanterre), France
e-mail: denis.lepesant@orange.fr

After having built a thesaurus of the sentiment words, I am implementing it into NooJ, in order to have available a tool allowing such semantic tags as the ones given below, in a novel of Emile Zola:

Mais, deux ans plus tard, Antoine tomba au sort. <SN TYPE="CAUS">
Sa mauvaise chance</SN> <PRO TYPE="EXP">le</PRO> <VERB CATEG="Affect" CLASS="Emotion" SSCLASS="toucher" REG="neutre" SENS="">toucha peu</VERB> ; <PRO TYPE="EXP">il</PRO>
<V CATEG="Affect" CLASS="Espoir" SSCLASS=""REG="neutre" SENS="">comptait </V> <PROP TYPE="CONJ" TYPE="OBJ">que sa mère
lui achèterait un homme</PROP>. Adélaïde, en effet, voulut le sauver

The tags "CAUS", "EXP" et "OBJ" mean respectively "Cause of a feeling", "Experiencer", "Object of a feeling". The sentiment verbs are tagged according to the main Category ("Affect"), the Class ("Emotion", "Espoir"), the Subclass ("toucher"), the Register ("neutre"). The category SENS (meaning) is so far empty ; it should potentially include either paraphrases and definitions, or translations into a foreign language.

The sentiment words are probably, at least in French, the largest semantic category (more than 3500 simple as well as polylexical verbs, nouns and adjectives, divided into about 80 classes). The methodological principles of the linguistic classification was outlined in [1]. In several conferences, I presented the tasks of implementation in progress. The first interesting results are showed in [1, 2].

In my talk at the NooJ International Conference 2015, I intend to display for the first time:

– a "NooJ.dic" table corresponding to the complete and final form of the thesaurus of the sentiment words;

– the main types of local grammars "NooJ.nog" corresponding to the lexicon of the sentiment words;

– a numerical evaluation of the quality of the semantical tags (silence and noise rates, rate of partly wrong tags) inside a big corpus (79 novels of Balzac).

References

1. Le Pesant, D. Vers un thésaurus syntactico-sémantique des mots d'affect / D. Le Pesant // Cahiers de Lexicologie. – 2011. – Vol. 2, № 99. – P. 117–132.

2. Présentation d'un thésaurus des mots d'affects: théorie, méthode et applications / D. Le Pesant [et al.] // Les émotions dans le discours. Emotions in Discourse. – Frankfurt am Main : Peter Lang, 2014. – P. 395–408.

CREATION OF GEOGRAPHICAL NAMES DICTIONARY OF ALASKA TOPONYMS

A. Loskutova
Belarusian State University, Minsk
e-mail: loskutovatosha@gmail.com

The first “Geographic Dictionary of Alaska” was published in 1902 as Geological Survey Bulletin 187. The Digital Age motivate us to new discoveries, let it be NooJ geographical dictionary of Alaska that can be much more useful in processing texts with toponyms in our age. The idea is that some amount of geographical names [1] is collected in a NooJ dictionary as a small part of the whole NooJ computer program for natural language processing. Then the dictionary will be used for toponyms search in text corpora.

Moreover there will be semantic marks, such as the part of speech of the word and its flection list.

Nowadays, one of the most important problems in translation field will be translation of Proper nouns. This task is considered to be rather sophisticated because there are several approaches how to convey the main essence of the words amongst cultures. Longman Dictionary of Language Teaching and Applied Linguistics defines the proper name as “a name which is the name of a particular person, place, or thing”.

All these issues will be reflected in the work of creation of geographical names dictionary of Alaska toponyms using NooJ.

References

1. Orth, D.J. Dictionary of Alaska Place Names / D.J. Orth. – Washington : United States Printing Office, 1971. – 1084 p.

ADDITION OF PHONETIC TRANSCRIPTIONS TO BELARUSIAN MODULE OF NOOJ

S. Lysy, A. Hiuntar, Yu. Hetsevich
United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: stanislau.lysy@gmail.com

To learn, study or perform a text processing for one or another language full and thorough description of the language is required. The authors of this article have noted that while much has been done in the development of different areas of language processing with NooJ, yet little attention has been paid to issues related to phonetic language features [1].

This article will describe a way to represent phonetic level of language for Belarusian module of NooJ. This will be done in two ways: via creating a dictionary including phonetic transcriptions and via developing morphological NooJ grammars for creating a phonetic transcription for orthographic words.

For the first part of this aim, a software tool which allows to quickly transform both single words and whole texts into phonetic transcription will be used [2, 3]. This software tool can generate three kinds of transcription: Cyrillic, simple Latin and International Phonetic Alphabet [4]. Apart from that there will be developed and implemented an algorithm, which adds phonetic transcription in three forms listed above for every word in the Belarusian dictionary.

For the second part of this aim, a morphological NooJ grammars will be developed. In Belarusian, one letter may be represented by different allophones depending on their surrounding letters or position in the word. The most common sound changes in Belarusian are assimilation, elision and positional fortition. For example, in the word *дуб* "dub – eng. oak", the last letter *B* changes into the sound [p] as a result of end-word fortition. These sound changes will present in the grammar as following: all the graphemes, which are surrounded by other particular graphemes will be given as an output the allophone match, for instance, grapheme B from the example above will be marked by P as a corresponding allophone.

The main purpose of this paper is description structure of the Belarusian language using NooJ, which will help in introducing and learning the norms of the literary pronunciation of this language. Moreover the results can be useful in dealing with other educational and linguistic problems.

References

1. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 22.12.2014.

2. Transcription Generator [Electronic resource]. – 2014. – Mode of access : http://corpus.by/transcriptionGenerator. – Date of access : 22.12.2014.

3. The system of generation of phonetic transcriptions for input electronic texts in belarusian / Yu. Hetsevich [et al.] // Pattern Recognition and Information Processing : Proceedings of The 12th Intern. Conf., 28–30 May, Minsk, Belarus. – Minsk : UIIP NASB, 2014. – C. 81–85.

4. International Phonetic Association [Electronic resource]. – 2005. – Mode of access : http://www.internationalphoneticassociation.org. – Date of access : 22.12.2014.

MORPHEME-BASED RECOGNITION AND TRANSLATION OF MEDICAL TERMS

A. Maisto, R. Guarasci
University of Salerno, Fisciano, Italy
e-mail: {amaisto, rguarasci}@unisa.it

The technical-scientific language of the medicine, provided with a number of technical lemmas that is larger than any other sub-code, is a part of the set of sub-codes that are organized in taxonomies and strong notional fields. Each term of this huge sub-dictionary, besides, occurs in texts with a very low frequency. For this reason, the majority of medical sub-domain terms could be defined as "rare events" [1]. This phenomenon could has a negative impact on the performances of the statistical and the machine learning methods. In general, free lexical resources for the medical domain, are often few and incomplete for every kind of language. Multilingual resources, in addition, are very rare and have a crucial role in every NLP systems.

The idea of the paper is to approach this large number of medical terms starting from a restricted dictionary (about 1000) of morphemes that, combined one another, allow the recognition of a huge number of terms, at least in two languages: Italian, English. This kind of approach, called Morpho-semantics [2–4], can be used to describe, in an analytical way, the meaning of the words that belong to the same subdomain or to the same "morphological family" (e. g. words: iper-acusia, ipo-acusia; subdomain: -*acusia* "otolaryngology"; description: *ipo*- "lack", *iper*- "excess", etc.).

We grounded the automatic creation of medical lexical databases on specific formative elements that are able to define a meaning in a univocal way, thanks to the regular combination of modules defined independently. Such elements do not represent mere terminations, but possess their own semantic self-sufficiency [5].

In order to build a multilingual medical thesaurus in which every lemma is automatically associated with its own terminological and semantic properties and with the respective English translations we created two small NooJ dictionaries of morphemes (an Italian Morphemes dictionary and an English one). Morpheme may belong to three morphological categories, Prefixes, Confixes, and Suffixes, which are provided with semantic annotations (that contains the meaning of the morpheme), terminological annotations (that refers to the medical class when it is possible to include the morpheme to a specific semantic class) and with the translation of the morpheme in the other language (e. g. iper, "hyper"). A Morphological Grammar finds every possible combination of Prefixes, Confixes and Suffixes and annotates the recognized medical term separating it in different units, according with the morphemes that

compose the words. A corpus of about 1000 Italian Medical Records has been analyzed with this resources configuration and, later, a syntactic translation grammar has been applied: for every combination of morphemes, the grammar transcribes as output the English transduction of the morpheme.

References

1. Möbius, B. Rare events and closed domains: Two delicate concepts in speech synthesis / B. Möbius // International Journal of Speech Technology. – 2003. – Vol. 6, № 1. – P. 57–71.

2. Norton, L. Morphosemantic analysis of compound word forms denoting surgical procedures / L. Norton, M.G. Pacak // Methods of Information in Medicine. – 1983. – Vol. 22, № 1. – P. 29–36.

3. Word segmentation processing: a way to exponentially extend medical dictionaries / C. Lovis [et al.] // Medinfo. – 1995. – Vol. 8, № 1. – P. 28–32.

4. Automatic Population of Italian Medical Thesauri: a Morphosemantic Approach / F. Amato [et al.] // 9th Intern. Conf. on P2p, Parallel, Grid, Cloud and Internet Computing. – Guangzhou : Springer, 2014. – Lecture Notes in Computer Science. – P. 432–436.

5. Iacobini, C. Composizione con elementi neoclassici / C. Iacobini // La formazione delle parole in italiano. – A cura di M. Grossmann & F. Rainer, 2004. – P. 69–95.

HOW TO AUTOMATICALLY ENRICH LINGUISTIC RESOURCES USING NOOJ: APPLICATION ON ARABIC MODULE

S. Mesfar, D. Najar
RIADI, ENSI, University of Manouba, Tunisia
e-mail: mesfarslim@yahoo.fr

This work will deal with the possibility of building new lexical resources or enhancing the existent ones using the NooJ linguistic engine. The preconized approach will be tested on the NooJ's Arabic module. Additionally, it will take into account the high agglutinative specificities of the language.

The study is composed of three main parts. First, we will try to extract lexical entries using existing general open resources (Wikipedia, Wiktionary,...) as well as available multilingual ontologies (BabelNet, Translatica, ProlexBase, etc.). Then, the extracted lists are cross-referenced with the existing NooJ resources in order to identify the remaining entries serving to enrich them. Furthermore, we give an overview of the enrichment algorithm using NooJ apply command-line program.

The second part of our work will focus on building local grammars formalizing some syntactic and syntagmatic rules combined with disambiguation patterns. These local grammars will focus on unknown words and try to propose the likely part-of-speech (Noun, Verb, Adjective or particle).

The last part of this work will deal with semantic and distributional enrichment of lexical resources using our automatic document classification program based on a large coverage terminological dictionary (including simple as well as compound lexical entries) [1–3].

Finally, in order to validate the whole preconized approach, we show results on general and thematic corpus automatically collected from newspaper websites.

References

1. Elleuch, I. Towards automatic enrichment of standardized electronic dictionaries by semantic classes / I. Elleuch, B. Gargouri, A. Ben Hamadou // The 2014 Conf. on Computational Linguistics and Speech Processing ROCLING 2014. – Jhongli, Taiwan, 2014. – P. 96–109.

2. Miller, G.A. WordNet: An on-line lexical database / G.A. Miller // International Journal of Lexicography. – 1990. – Vol. 3(4). – P. 235–312.

3. Savary, A. ProlexFeeder – Populating a Multilingual Ontology of Proper Names from Open Sources / A. Savary, L. Manicki, M. Baron // Laboratoire d'Informatique, Université François Rabelais Tours. – France, 2013. – 38 p.

LOCAL GRAMMARS AND FORMAL SEMANTICS: PAST PARTICIPLES VS. ADJECTIVES IN ITALIAN

M. Monteleone
Dipartimento di Scienze Politiche, Sociali e della Comunicazione –
Università degli Studi di Salerno, Italy
e-mail: mmonteleone@unisa.it

Especially from the point of view of lexicographic descriptions, in Italian we may find a high level of categorial ambiguity between two specific parts of speech, i. e. past participles and adjectives. Very often, the words belonging to these two parts of speech are homographs and semantically contiguous. Above all, in most of their uses, it is possible to identify their correct linguistic function only by means of precise syntactic analyses, which must be focused on both the left and right contexts co-occurring with the propositions to be examined. Such analyses must also infer about all the possible verb antecedents to past participles and adjectives, be they verb operators, support verbs or simple auxiliary verbs.

Therefore, in such cases, to define automatically if a given Italian word is a past participle or an adjective, it is crucial to disambiguate correctly the role of its immediate verb antecedent(s). For instance, and from a theoretical point of view, it is possible to observe that in Italian a past participle, which syntactically belongs to the category of verb operators, in most cases is preceded by the auxiliary verb avere (to have), in its both simple and compound forms. On the other hand, an Italian is very often preceded by the support verb essere (to be), or by the set of its possible substitutes, or finally by other forms of verb support extensions. At the same time, the right contexts of both past participles and adjectives may be formed by variable sets of word groups.

Therefore, in our article, we intend to use then NooJ [1, 2] with the following purposes:

– regarding Italian simple words, and at the level of lexicographical description, we will study and define the levels of categorical ambiguity existing between past participles and adjectives;

– subsequently, we will define the syntactic pattern in which this ambiguity can be solved;

– we will describe the construction of a set of local grammars to apply for the disambiguation and correct tagging of these parts of speech;

– finally, we will determine the levels of recall and precision of these grammars.

At the end of this four-step process, we aim at creating a set of formal semantic analysis tools [3–6], to exploit in NooJ to distinguish sentences with conventional operators (verb predicates) from sentences with support verbs and predicative adjectives.

References

1. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 12.12.2014.

2. Silberztein, M. NooJ: a Linguistic Annotation System for Corpus Processing / M. Silberztein // Proceedings of HLT/EMNLP. – Canada, 2005. – P. 10–11.

3. Silberztein, M. Analyse et génération transformationnelle avec NooJ / M. Silberztein // Proceedings of the 47th annual meeting of the Italian linguistic Society “Livelli di Analisi e Fenomeni di Interfaccia”. – Bulzoni, Rome, 2015.

4. Silberztein, M. La formalisation des langues: l´approche de NooJ / M. Silberztein. – Iste Ediciones, London, 2015. – P. 425.

5. Vietri, S. The Formalization of Italian Lexicon-Grammar Tables in a Nooj Pair Dictionary/Grammar / S. Vietri // Applications of Finite-State Language Processing: Selected Papers from the NooJ 2008 Intern. Conf. – Cambridge Scholars Publishing, New Castle upon Tyne, 2010. – P. 138–147.

6. Vietri, S. The Annotation of the Predicate-Argument Structure of Transfer Noun / S. Vietri // Formalising Natural Languges wih NooJ. – Cambridge Scholars Publishing, New Castle upon Tyne, 2013. – P. 88–101.

PARAPHRASING HUMAN INTRANSITIVE ADJECTIVE CONSTRUCTIONS IN PORT4NOOJ

C. Mota[1], P. Carvalho[1,2], F. Raposo[1], A. Barreiro[1]
[1] INESC-ID, Lisboa, Portugal;
[2] Uni. Europeia | Laureate International Universities, Lisboa, Portugal
e-mail: cmota@ist.utl.pt

Port4NooJ [1, 2] is a set of resources that allow the generation of paraphrases for Portuguese. These paraphrases feed the linguistic engine of the eSPERTo paraphrasing system (http://esperto.l2f.inesc-id.pt/), based on NooJ technology [3]. This paper presents an enhanced Port4NooJ that includes fifteen lexicon-grammar (LG) tables describing the distributional properties of 4,250 human intransitive adjectives [4], which add new paraphrasing capabilities to eSPERTo. Among other properties, these linguistic resources provide information on:

– syntactic and semantic nature of the subject modified by each adjective, which can correspond to a human noun, a complex noun phrase (NP) involving an appropriate noun, or to a finite or non-finite clause;

– copulative verbs (and aspectual variants) selected by each adjective;

– constraints related to the quantification of adjectives by an adverb or a degree morpheme;

– position of adjectives in adnominal context (pre- or post-nominal position);

– possibility of certain adjectives being optionally followed by an infinitive clause, with causal interpretation, or by a human NP introduced by the preposition *para* (*for*).

In addition to general properties, these resources also describe other constructions in which human intransitive adjectives may occur:

– generic and cross-constructions, where the adjective is the head of the NP;

– characterizing indefinite constructions, where the adjective occurs after an indefinite article;

– exclamative sentences expressing insult.

Initially, Port4NooJ contemplated paraphrases involving support verb constructions (*fazer uma apresentação* (*make a presentation* (*of*)) = *apresentar* (*present*)), compound adverbs (*de uma forma interativa* (*in an interactive way*) = *interativamente* (*interactively*); *com entusiasmo* (*with enthusiasm*) = *entusiasticamente* (*enthusiastically*)), relatives (*que foram escritos* (*that were written*) = *escritos* (*written*); *o papel que a Europa tem/desempenha* (*the role that Europe has/plays*) = *o papel da Europa* (*the role of Europe*)), active/passive

constructions (*A solta B* (*A releases B*) = *B é solto por A* (*B is released by A*)), among others.

The use of the linguistic knowledge described in the tables allows themapping of several other types of paraphrasing constructions resulting in a semantic relationship between predicate adjectives, nouns and verbs. The LG tables enable eSPERTo to paraphrase (i) morphologically related adjective, noun and verb constructions (*está zangado* (*he is angry*) = *zangou-se* (he *got* (*himself*) *angry*) = *esteve envolvido numa zanga* (*he was involved in a fight*); (ii) adjective constructions supported by different copulative verbs (*estar perdido* (*to be lost)* = *andar perdido* (*to walk around lost*)); (iii) constructions involving nationality and other membership relations (*de origem portuguesa* (*of Portuguese origin/roots*) = *português* (*Portuguese*) = *de Portugal* (*from Portugal*); *benfiquista* (*Benfica fan*) = *do Sport Lisboa e Benfica* (*a fan of Sport Lisboa e Benfica*)); (iv) cross-constructions (*o idiota do rapaz* (*the idiot of the boy*) = *o rapaz é um idiota* (*the boy is an idiot*)); appropriate noun constructions (*foi arrogante nos seus comentários* (*he was arrogant in his comments*) = *os seus comentários foram arrogantes* (*his comments were arrogant*) = *foi arrogante* (*he was arrogant*), (v) generic noun phrases (*é um indivíduo estúpido* (*he is a foolish individual*) = *é um estúpido* (*he is a fool*) = *é estúpido* (*he is fool*)), among others.

Our work follows previous attempts to integrate LG tables in NooJ, such as experiments with LG tables of transitive and neutral phrasal verbs in English [5], and the integration of a LG of Italian idioms [6].

References

1. Barreiro, A. Linguistic Resources and Applications for Portuguese Processing and Machine Translation / A. Barreiro // Applications of Finite-State Language Processing: Selected Papers from the NooJ 2008 Intern. Conf. – UK : Cambridge Scholars Publ., 2010. – P. 41–51.

2. Barreiro, A. Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation / A. Barreiro // Proc. of the 2007 Intern. NooJ Conference. – UK : Cambridge Scholars Publ., 2008. – P. 19–47.

3. Silberztein, M. La formalisation des langues: l'approche de NooJ / M. Silberztein. – UK : ISTE Ed., 2015. – 426 p.

4. Carvalho, P. Análise e Representação de Construções Adjectivais para Processamento Automático de Texto / P. Carvalho // Adjectivos Intransitivos Humanos, PhD Dissertation. – Universidade de Lisboa, 2007. – 452 p.

5. Machonis, P. English Phrasal Verbs: from Lexicon-Grammar to Natural Language Processing / P. Machonis // Southern Journal of Linguistics. – USA : SECOL, 2010. – Vol. 34(1). – P. 21–48.

6. Vietri, S. The Formalization of Italian Lexicon-Grammar Tables in a Nooj Pair Dictionary/Grammar / S. Vietri // Applications of Finite-State Language Processing: Selected Papers from the NooJ 2008 Intern. Conf. – UK : Cambridge Scholars Publ., 2010. – P. 138–147.

A LARGE TERMINOLOGICAL DICTIONARY OF ARABIC COMPOUND WORDS

D. Najar, S. Mesfar
RIADI, ENSI, University of Manouba
e-mail: dhekra.najar@gmail.com

NooJ is a linguistic development environment that allows formalizing complex linguistic phenomena such as compound words generation, processing as well as analysis. We will take advantage of NooJ's linguistic engine strength in order to create a new large coverage terminological compound word's dictionary for Modern Standard Arabic language. Classifying and annotating Arabic compound words would have a major impact on the disambiguation of applications working with Arabic texts. The diverse analyzers, based on morphological aspect, are not able to recognize multiword expressions. These multiword expressions are combinations of single terms expressing various meaning compared to basic terms. Morphological analyzers usually separate compound expressions into single terms. Therefore recognizing the entire compound words is essential to preserve the semantic of texts and to provide a crucial resource to better analysis and understanding of Arabic language.

Our work is composed of three sections. First, we will deal with a literature review on Arabic compound expression's categories which aims to dress a detailed topology. The structural variability of multiword expressions in Arabic language will be studied in order to measure the degree of morphological, lexical and grammatical flexibility of multiword expressions. Then, we will discuss present the electronic thematic dictionary of compound Arabic expressions and give detailed description of our methodology and guidelines.

This morphological dictionary contains lexical entries divided into more than 20 domains including medical, political, legal, religious, financial, economical, computer science, etc. Most of these entries belong to scientific and technical terminology. Our lexicon covers other types of compound words such as expressions that are traditionally classified as idioms, prepositional verbs, collocations, etc. We provide the syntactic phrase structure composition of the Arabic compound expressions, giving each entry of our lexical resource its component elements (noun+noun, noun+adjective, verb+preposition+noun…). Moreover, each lexical compound entry of our lexicon is associated with a set of semantic (Semantic information where we cover semantic fields) and distributional information, an inflexional paradigm as well as some derivational descriptions generating duals and plural forms (regular and irregular forms).

Organizing our specialized lexical entries in semantic field format brings many practical benefits; one of those is to allow classifying textual documents by category. The inflexional and derivational paradigms, which concern

syntactically flexible compound words, are using some specific morphological operators that will be described as well. A syntactically flexible multiword expression is a frequent combination of two words or more, characterized by high degree of morphological and syntactic flexibility. Finally, we show new results showing the rates of morpho-lexical coverage enhancement.

CONTEXT-SENSITIVE HOMOGRAPH DISAMBIGUATION WITH NOOJ IN BELARUSIAN AND RUSSIAN ELECTRONIC TEXTS

T. Okrut [1], B. Lobanov [1], Y. Yakubovich [2]
[1] United Institute of Informatics Problems of the NAS of Belarus, Minsk;
[2] Universitat Autònoma de Barcelona, Bellaterra, Spain
e-mail: tatberrie@gmail.com

When we read, we may encounter words having different phonological representations associated with singular orthographic representation. In speech synthesis, disambiguation of such words, or homograph disambiguation, serves an obstacle to overcome at the stage of text preprocessing. There are several major types of homographs we deal with in Belarusian and Russian: different lexemes of the same part of speech, different forms of the same lexeme and different lexemes of different parts of speech. Moreover, these homographic groups may be divided into the subgroups on the base of grammatical similarity of homographic word pairs [1]. Elements in such pairs differ at least in one grammatical feature, which influences the stress position in a word. For example, in the following homographic pairs of the group "different forms of the same lexeme" elements differ in number:

ГО́ДА (singular, "year") – ГОДА́ (plural, "years"),
О́ЗЕРА (singular, "lake") – ОЗЁРА (plural, "lakes").

Such similarity allows developing of one context-sensitive disambiguation algorithm for a number of homographic pairs at once. The authors have already developed a Russian syntactic NooJ grammar for disambiguation of 58 homographs referring to the homographic subgroup "Singular nouns of masculine or neuter gender in genitive case – Plural nouns in accusative or nominative case". Therefore, the goal of this research is to improve the grammar mentioned above and to develop a similar Belarusian disambiguation grammar using a context-sensitive approach.

References

1. Вырашэнне амаграфіі з дапамогай NOOJ для больш чым 50 амографаў рускай мовы / Т.І. Окрут [і інш.] // Контрастивные исследования и прикладная лингвистика: материалы Междунар. науч. конф., Минск, 29–30 окт. 2014 г. : в 2 ч. / М-во образования Респ. Беларусь, Минский гос. лингв. ун-т; редкол. : А.В. Зубов [и др.]. – Минск, 2014. – Ч. II. – С. 83–87.

SEMANTIC ANALYSIS FOR LOCATING EXPRESSIVE MEANS AND STYLISTIC DEVICES IN AUTHENTIC ENGLISH TEXTS, RANGING AND CLASSIFICATION

A. Patsiomkin[1], Yu. Hetsevich[2]
[1] Belarusian State University, Minsk;
[2] United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: andrei.patsiomkin@gmail.com

As natural language processing has become one of the leading fields in modern science, it is reasonable enough to use its achievement in terms of other scientific or educational spheres [1]. Thus, the study of natural texts can be used in education and this assumption led to the idea which could serve both the educational and scientific purposes.

The following work is aimed at creation and implementation of specific grammar and semantic rules, used in NooJ [2–4], in order to carry out a detailed and comprehensive analysis of natural language for the purpose of locating expressive means (EM's) and stylistic devices (SD's) in authentic English texts and ranging the given texts according to their lexical and semantic richness [5–7].

To achieve the abovementioned, a set of rules will be represented as the result of the work, with the help of which NooJ parser will perform thorough evaluation of incoming material followed by the final outcome, fitting the frames specified. The classification of the texts based on the results of the experiment will be worked out.

The work can be used in education at lexicology and stylistics classes of the English language as well as it's highly expected that the methods applied in this work may serve as a basis for future development of syntactic-semantic analysis of written texts.

References

1. Jurafsky, D. Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech recognition / D. Jurafsky, J.H. Martin. – New Jersey, 2008. – 1024 p.

2. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2003. – Mode of access : http://www.nooj4nlp.net. – Date of access : 21.12.2014.

3. Chomsky, N. Syntactic Structures / N. Chomsky. – Berlin – NY : Mouton de Gruyter, 2002. – 117 p.

4. Byrd, S. Natural language processing with python / S. Byrd, E. Klein, E. Loper. – Newton : O'Reilly Media, 2009. – 504 p.

5. Galperin, I.R. Stylistics / I.R. Galperin. – M. : Higher School, 1977. – 334 p.

6. Arnold, I.V. The English Word / I.V. Arnold. –M. : Higher School, 1986. – 295 p.

7. Calchei, M. Practising stylistics through lyrics / M. Calchei, E. Albu, A. Sudniţîna. – Moldova State University, Chişinău, 2014. – 40 p.

NORMALIZATION OF TWEETS IN CROATIAN LANGUAGE USING NOOJ

T. Pejar[1], K. Kocijan[2], B. Bekavac[1]
[1] Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia;
[2] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
e-mail: tpejar@ffzg.hr

Existing natural language processing tools face various difficulties when applied to informal language, eg. the language used in the context of social networks. There are two main approaches to this problem: domain adaptation and normalization. This area of research is new and undeveloped with regard to Croatian. The dataset is a collection of tweets. Twitter has several advantages. First and foremost, tweets are public and therefore privacy issues are avoided. There are large quantities of text are available. Furthermore, restrictions imposed on tweet length encourage innovative use of abbreviations and contractions. Textual cues such as emoticons are similarly used to compensate for the lack of a visual channel. NooJ is a used for tweets processing in combination with existing Croatian language resources. When using those resources, the majority of unclassified words (the dictionary UNKNOWNS) may be said to belong to such informal language. These irregularities had to be classified and formally described in terms of NooJ grammars (inflectional, lexical and syntactic grammars) and dictionaries. This work is mostly focused on lexical normalization, while postponing syntactic irregularities for future work. The additions made to Croatian language resources for NooJ result in a 14 % reduction in unclassified words for those users whose tweets were analyzed during development, and a 12 % reduction for other users. Therefore, this work is far from complete, but provides a useful basis for future research.

MORPHOLOGICAL RELATIONS FOR THE AUTOMATIC EXPANSION OF ITALIAN SENTIMENT LEXICONS

S. Pelosi
Department of Political, Social and Communication Science
University of Salerno, Italy
e-mail: spelosi@unisa.it

In this abstract we propose a morphological strategy for the enlargement of electronic dictionaries of sentiment in the Italian language. Lexical databases of this kind can be truly useful into Sentiment Analysis tools, but their manual creation can become a very slow task; so, several solution for their automatic construction and testing have been developed in literature, e.g. conjunctions and morphological relations between adjectives [1]; context coherency [2]; word similarity [3]; pointwise mutual information [4].

In our research we will show the possibility to double the dimension, or even to create from scratch, new sentiment dictionaries thanks to derivational phenomena. Our inputs are a manually built Nooj lexicon of Italian adjectives of sentiment (5000 + entries) [5], a list of prefixes and suffixes, and a set of Nooj morphological grammars, able to put in relation sentiment words and affixes and to modify, in many different ways, the grammatical category, the semantic orientation (positive or negative), or the intensity (strong or weak) of the starting lemmas.

In detail, our work takes advantages from linguistic clues pertaining to the derivation, from semantically oriented adjectives, of quality nouns (e. g. with suffixes as *-ità*, *-ia*, *-ezza*, etc…) and adverbs in *-mente*, with the purpose of making them automatically derive the semantic information associated to the adjectives which they are morpho-phonologically related with. Furthermore, we use as morphological Contextual Valence Shifters a list of prefixes able to negate (e. g. *anti-*, *contra-*, *non-*, ect…) or to intensify/downtone (e. g. *arci-*, *semi-*, ect…) the orientation of the words in which they appear. Thus, if the just mentioned suffixes can interact with pre-existing dictionaries of the Italian module of Nooj, in order to automatically tag them with new semantic descriptions; the cited prefixes can directly work on opinionated documents, so Nooj can "understand" the actual orientation of the words occurring in real texts. The evaluation of the precision reached by the automatically built dictionaries (more than the 90 % for both nouns and adverbs) and the error analysis will be discussed in detail in the full paper.

We conclude this abstract by clarifying that the morphological method could be also applied to Italian verbs, but we prefer to avoid this solution because of the complexity of their argument structures. We decided, instead, to manually evaluate all the verbs described in the Italian Lexicon-grammar binary tables, so we could preserve the different lexical, syntactic and transformational rules connected to each one of them [6].

References

1. Hatzivassiloglou, V. Predicting the semantic orientation of adjectives / V. Hatzivassiloglou, K.R. McKeown // Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conf. of the European

Chapter of the Association for Computational Linguistics. – Stroudsburg, 1997. – P. 174–181.

2. Kanayama, H. Fully automatic lexicon expansion for domain-oriented sentiment analysis / H. Kanayama, T. Nasukawa // Proceedings of the 2006 Conf. on Empirical Methods in Natural Language Processing. – Sydney, 2006. – P. 355–363.

3. Mohammad, S. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus / S. Mohammad, C. Dunne, B. Dorr // Proceedings of the 2009 Conf. on Empirical Methods in Natural Language Processing. – Singapore, 2009. – Vol. 2. – P. 599–608.

4. Turney, P.D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews / P.D. Turney // Proceedings of the 40th annual meeting on association for computational linguistics. – Philadelphia, 2002. – P. 417–424.

5. Maisto, A. A Lexicon-Based Approach to Sentiment Analysis. The Italian Module for Nooj / A. Maisto, S. Pelosi // Book of Proceedings of the Intern. Nooj 2014 Conf., June 3–5. – University of Sassari, Italy, 2014.

6. Elia, A. Lessico e strutture sintattiche: introduzione alla sintassi del verbo italiano / A. Elia, M. Martinelli, E. D'Agostino. – Liguori, 2003. – 430 p.

FIRST ONE MILLION CORPORA FOR BELARUSIAN NOOJ MODULE

I. Reentovich[1], Yu. Hetsevich[1], V. Voronovich[2], E. Kachan[2], H. Kozlovskaya[2]
[1] United Institute of Informatics Problems of the NAS of Belarus, Minsk;
[2] Belarusian State University, Minsk
e-mail: mwshrewd@gmail.com

In this report first 1 million corpus for Belarusian NooJ module is represented. The given corpus has been built up of texts, patched up into sections of different subject lines. From the broad list of possible subject lines in the sections the corpus focuses on fiction, historic, medical, scientific, sociological literature and etc. And if being of the view that there is a great many of analogous subject lines, then this first 1 million corpus can be considered as the first subject collection of texts for Belarusian NooJ module.

The text corpus that is used in NooJ will be effective for the research activity development on the following respects:

1) the words polysemy processing in texts of different subjects;
2) the polysemic punctuation marks processing;
3) the new lexical items search.

Besides, the 1 million corpus will be for all intents and purposes applicable for solving many crucial tasks:

in general

– use this corpus in a linguistic development environment called NooJ [1] to optimize and expand the development of high-quality linguistic algorithms for the electronic texts pre-processing TTS block;

in particular

– conduct several experiments in order to specify at the minimum and, possibly, maximum level of various syntactic and morphological grammars using effectiveness for texts of each subject section;

– take thorough measures in order to create the *subject domain generator* (which will be then very useful for the formation of special subject-oriented NooJ dictionaries);

– in the most extent use the given corpus in the process of text-to-speech synthesis with the help of available programs [2], required for such process, and also when testing newly created applications;

– make comparative analysis of this corpus with the same corpora in other languages (taking into account all necessary rules, language features in texts of each current corpus, various possible emerging issues, while building syntactic and morphological grammars, etc.).

It is very essential that the first 1 million corpus for Belarusian NooJ module can be completely applicable in any line of linguistic research. And in

the near future the corpus is planned to be expanded up to approximately 5–10 million words.

References

1. NooJ: A Linguistic Development Environment [Electronic resource]. – 2015. – Mode of access : http://www.nooj4nlp.net/. – Date of access : 08.02.2015.

2. Corpus.by // Corpus.by [Electronic resource]. – 2015. – Mode of access : http://www.corpus.by/. – Date of access : 08.02.2015.

A PROPOSAL FOR THE TREATMENT OF CLITICS IN RIOPLATENSE SPANISH VERBS USING NOOJ

A. F. Rodrigo
Grupo InfoSur, Fines;
Universidad Nacional de Rosario, Argentina
email: andreafrodrigo@yahoo.com.ar

The presence of unstressed pronominal forms or clitics (*me*, *te*, *se*, *le*, *les*, *lo*, *la*, *los*, *las*) is a distinctive characteristic of the Spanish language. Clitics will never appear isolated, given that they phonologically depend on the verb [1] to which they are attached and have certain restrictions in terms of order. This way, the dative is always located in front of the accusative, for example. The intention of this paper is to show how NooJ [2] can be used for the automatic treatment of some of the sequences in Rioplatense Spanish, such as:

1. Te lo conté. (I told you).
2. Me la prestás.(You lend it to me).
3. Me lo das. (You give it to me).
4. Lo entendés. (You understand it).

In all the cases, we find sequences comprised of one or two clitics preceding the verb, i.e. as proclitic. Here we resume the proposal by Bonino (2013) [3], who dealt with strings of verbs and enclitics and how their automatic treatment is possible with NooJ. We are interested here in the possibility of incorporating to the verb a category [+PRON] that enables the possibility of establishing a sort of verbal typology in line with what is proposed by Solana (2008) [4]. Meaning that the verbs can be characterized by their admittance of clitic combinations or the lack thereof.

The grammar to be created should express that in (1), for example, the following sequences are grammatically correct:

1.1) conté;
1.2) lo conté;
1.3) te conté;
1.4) te lo conté;

being most common the expressions that contain a clitic ((1.2) and (1.4)), at least in Rioplatense Spanish. Also the above mentioned grammar should exclude:

1.5)* lo te conté.

Finally, we will not deal with duplications here, which are very frequent in this Spanish variant, such as in:

5. Lo vi a Juan. (I saw John).

This will be left for a subsequent work.

References

1. Fernández Soriano, O. El pronombre personal, formas y distribuciones, pronombres átonos y tónicos / O. Fernández Soriano, I. Bosque, V. Demonte // Gramática descriptiva de la lengua española. – Espasa, Madrid, 1999. – P. 1209–1273.

2. Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 14.12.2014.

3. Bonino, R. Una propuesta para el tratamiento de los enclíticos en NooJ / R. Bonino. – Rosario : Universidad Nacional de Rosario, 2015. – Vol. 7. – P. 31–40. I:\Abstracts_New template\infosurrevista.com.ar

4. Solana, Z. Clíticos como clasificadores de verbos / Z. Solana. – Rosario : Universidad Nacional de Rosario, 2008. – Vol. 2. – P. 61–72.

TOWARDS BUILDING OSTIS TECHNOLOGY-BASED SEMANTIC NLP APPLICATIONS USING NOOJ

K. Rusetski[1], D. Ilyushchenia[1], K. Nikalaenka[1], S. Lysy[2]
[1] Belarusian State University of Informatics and Radioelectronics, Minsk;
[2] United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: rusetski.k@gmail.com

Nowadays, applied intelligent (knowledge-based) systems are quite complex to operate and usually require fairly deep understanding of artificial intelligence concepts. Natural language interfaces (NLI) aim to simplify those interactions, bring machine closer to human. Furthermore, the NLI is essential for the next generation of educational and training systems [1]. NLI obviously makes substantial use of natural language processing (NLP) technologies. While lexical and syntactical aspects of NLP are fairly well understood today, further research is still needed in its semantical aspects. In this paper, we will have a closer look at semantical side of things, particularly at integrating various NLP software pieces to achieve a better understanding of natural-language user input and a better, more user-friendly, natural-language output from computer system.

The article will cover an approach to integrating NooJ linguistics processor with natural language interface technology of Open Semantic Technology for Intelligent Systems (OSTIS) project. OSTIS project provides common formal and technological base for various traditional and intelligent computer systems to interact and augment each other in useful ways. This approach is based on using unified semantic networks (USNs) [2] that are able to represent a wide range of knowledge, both declarative and programmatic, due to their versatility. USNs are particularly useful in formalizing natural language knowledge in a way that is succinct enough for humans and at the same time is interoperable with whichever OSTIS-based applied intelligent system that is in need for linguistic services. OSTIS-NooJ integration is beneficial for NooJ [3] in that it allows for both efficient representation of NooJ grammars and extending and enhancing NooJ's semantic analysis capabilities. OSTIS project benefits from this integration by using NooJ's powerful syntactic engine, thus further facilitating its natural language interface technology.

References

1. Yeliseyeva, O.E. Component design of intelligent tutoring system to prepare students for centralized testing in a foreign language / O.E. Yeliseyeva, K.V. Rusetski // Open Semantic Technologies for Intelligent Systems (OSTIS-2013); eds. V.V. Golenkov [et al.]. – Minsk : BSUIR, 2013. – P. 511–516.

2. Golenkov, V.V. Graphodynamical models of parallel knowledge processing / V.V. Golenkov, N.A. Guliakina // Open Semantic Technologies for

Intelligent Systems (OSTIS-2012); eds. V.V. Golenkov [et al.]. – Minsk : BSUIR, 2012. – P. 23–52.

3. Silberztein, M. Nooj Manual [Electronic resource] / M. Silberztein. – 2014. – Mode of access : http://www.nooj4nlp.net. – Date of access : 07.12.2014.

UKRAINIAN DATA AND KNOWLEDGE BASE AND ITS ADAPTATION TO NOOJ

M. Sazhok, V. Robeiko, D. Fedoryn, R. Selyukh, O. Yukhymenko
International Research/Training Center for Information Technologies and Systems, Kyiv, Ukraine
e-mail: sazhok@gmail.com

The considered data and knowledge base (D&KB) has been initially developed for Ukrainian speech technology and systems like grapheme-to-phoneme, speech-to-text and text-to-speech conversions [1]. The D&KB elements are extracted and/or validated over basic dictionary and text corpus.

The basic dictionary is extracted from the electronic lexicography system subset containing 151 962 lemmas, including over 10 thousand names, that totally makes 1,90 million word forms [2]. Due to shared spelling the actual word form vocabulary consists of 1,83 million words that have different either spelling or primary lexical stress position. The developed tools extracted 4686 paradigm generation structures and 1755 stress generation structures. Each dictionary entry contains a word machine stem and references to a paradigm and stress generation structures.

The basic text corpus is derived from a hypertext data downloaded from several websites containing samples of news and publicity (60 %), literature (8 %), encyclopedic articles (24 %), and legal and forensic domain (8 %). A text filter, used for text corpus processing, provides conversion of numbers and symbolic characters to relevant letter sequences, removing improper text segments and paragraph repetitions. Hereafter, we refer to the basic corpus as 275 M corpus. In accordance to the corpus summary shown in table, we observe 6,64 word forms per lemma in average, whereas this relation is twice greater, 12,3, within the basic dictionary. Adding 200 000 most frequent words to the vocabulary we reduce OOV to less than 0,5 %.

To solve the graphematic analysis a respective table is provided. The table describes graphemes detected in the text corpus. Attributes for Ukrainian letters includes alphabet position, case conversion, phonematic features like vowel/consonant. Valid inner word characters (apostroph, dephis, hyphenation and lexical stress mark) are marked with respective labels. Other characters are accomplished with pronunciation and attributes to distinguish numbers, punctuation and other important types of symbols.

The developed toolkit allows for operating with the D&KB end exporting dictionary data for use in any linguistic software, for example in NooJ.

275 M basic text corpus summary

Words	Sentences	Vocabulary			OOV	Homographs
		All words	Known words	Known lemmas		
275 288 408	1 752 371	1 996 897	801 040	120 554	2,51 %	16 729 476

References

1. Sazhok, M. Distinctive features for Ukrainian real-time speech recognition system / M. Sazhok, V. Robeiko, D. Fedoryn // Proc. of the Intern. All-Ukrainian Conf. on Signal/Image Processing and Pattern Recognition UkrObraz'2014. – Kyiv, Ukraine, 2014. – P. 66–70.

2. Словники України [Electronic resource]. – 2014. – Mode of access : http://lcorp.ulif.org.ua/dictua. – Date of access : 16.12.2014.

NAMED ENTITY RECOGNITION FROM ARABIC-FRENCH HERBALISM PARALLEL CORPORA

M.A.F. Seideh[1], H. Fehri[1], K. Haddar[2], A. Ben Hamadou[2]
[1] University of Gabès, Tunisia;
[2] University of Sfax, Tunisia
e-mail: almedyfall@gmail.com

Bilingual lexicons play an important role in Natural Language Processing (NLP) such as Information Retrieval Interlingua (IRI) (ex. Multilingual information extraction) and machine translation (MT) (eg. Learning languages). However, they are very expensive to enrich manually, especially with regard to specialized dictionaries (in a particular domain). That is why in recent years many studies have used the alignment techniques to automate the bilingual dictionaries construction process [1–4]. These works showed that alignment of simple forms and named entities from parallel corpus is a relatively well-controlled task for Latin script languages. However, the pairing of parallel texts that are not using the same writing is a complex task.

In the context of the construction of bilingual dictionaries, we use parallel corpus of multilingual texts to extract bilingual lexical correspondences, but inaccuracies and liberties taken by the translators in these corpora can affect the quality of the inferred lexical entries. Moreover, the kind of the corpus has a great influence on the vocabulary that can be extracted. It is for these reasons that we turned towards the free encyclopedia Wikipedia. This resource contains structured and translated information's into many languages, making it an ideal tool for the creation of bilingual dictionaries by extracting bilingual lexical correspondences. Compared to the corpus of raw text, Wikipedia contains many titles and hyperlinks. This particular structure facilitates the extraction and classification of data. We focus here on some type of lexical entries, called named entities, often composed of several words. In our work we deal with Herbal medicine, also known as herbalism or botanical medicine. Herbal medicine is a medical system based on the use of plants or plant extracts. In recent years, interests in herbal medicine become significant. Many international studies have shown that plants are capable of treating disease and improving health, often without any significant side effects.

The identification of herbalism named entities is not an easy task, as the list of named entities is extensible and their structures are not accurate. This task becomes more difficult when searching to identify the correspondence of French Arabic named entity into Arabic language. Indeed, the structure of a named entity in French and Arabic is not the same. So, the number of components of a named entity is not the same in the French and Arabic language. Therefore, the

position of a named entity is not a sufficient criterion to identify the same named entity in French and Arabic language from a parallel corpus.

The proposed method is based on NooJ platform bilingual dictionaries and identified recognition rules (siding criteria). Recognition rules are elaborated from a study herbalism French-Arabic parallel corpora. Furthermore, we built morphological grammars to solve problems related to the Arabic language and syntactic grammars to recognize Herbalism named entities in French and Arabic language. To evaluate our work, we used another parallel corpora different of the study parallel corpora. From the obtained results, we tend to build a bilingual dictionaries (French-Arabic and Arabic-French) of Herbalism named entities.

References

1. Fehri, H. Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model // H. Fehri, K. Haddar, A. Ben Hamadou // FSMNLP 2011. – Blois, France, 2011 – P. 134–142.

2. Goldman, P. Création automatique de dictionnaires bilingues d'entités nommées grâce à Wikipédia / P. Goldman, P.Y. Scherrer // Nouveaux cahiers de linguistique française 30. – 2012. – P. 213–227.

3. Silberztein, M. NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE / M. Silberztein, A. Tutin // Spécial Atala. – 2005. – Vol. 8(2). – P. 123–134.

4. Yu, K. Bilingual dictionary extraction from Wikipedia / K. Yu, J. Tsujii // Proceedings of Machine Translation Summit XII. – Ottawa, 2009. – P. 379–386.

TRANSFORMATIONAL ANALYSIS OF TRANSITIVE SENTENCES

M. Silberztein
Université de Franche-Comté, France
e-mail: max.silberztein@gmail.com

Transformational Grammars are grammars that can link one or more elementary sentences to complex sentences that contain the same vocabulary items, such as Joe loves Lea <=> Lea is not loved by Joe.

We discuss the implementation of a complete set of simple transformational operators via the use of unrestricted grammars which involve the use of Turing Machines. In NooJ, such grammars are implemented by recursive graphs that contain variables and constraints.

Implementing simple transformations is not enough, because even complex sentences can be further transformed. We discuss the architecture of a system capable of performing chains of transformations in cascade, and we show that this system contains theoretical and methodological flaws.

We then present the new NooJ transformational functionality, including the optimisation of syntactic grammars and the linguistic formalisation of noun phrases. We then show how simple transitive sentences can generate millions of paraphrases and transformed complex sentences.

FROM LINGUISTIC TO KNOWLEDGE PROCESSOR

I. Sovpel
IHS Inc. / IHS Global Belarus, Minsk
e-mail: Igor.Sovpel@ihs.com

The importance of automatic natural language processing in modern information technologies is, undoubtedly very high. In most cases it is related to the modeling of natural language for the purpose of automatic understanding of text. Such understanding can be described as representation of text content in terms of a certain knowledge system, whose structure is determined by the task at hand. In most cases, such modeling requires development of procedures of linguistic analysis of text, and their implementation within a linguistic processor (LP). The set of procedures used in the majority of NLP tasks can be defined as basic linguistic processor (BLP). The BLP is universal in relation to different natural languages and to the goal of their processing. It includes lexical, grammatical, syntactic and semantico-syntactic analysis of text, which insures the extraction of its formal semantico-syntactic structure, which, in turn, can be used to further extract the principal types of knowledge. Therefore, BLP can be described as the basis of any efficient framework in both, theoretical researches in NLP, and in practical solutions for tasks at hand.

The results of all above-mentioned stages of language analysis are concluded in the text's linguistic index (LI), which can be represented in the following way:

$$LI = \langle W, POS, SYN, REL \rangle,$$

where W – is the set of words from the text; POS, SYN, REL – are the maps of all the words from the set to their grammatical and syntactic tags, as well as the tags of semantico-syntactical relations, which pre-determine the principal types of knowledge.

Such model has one very important feature – it naturally allows the inclusion of new components, corresponding to the requirements of a task at hand. For example, for the task of automatic knowledge extraction (in case, when the text is considered as the main knowledge source), it is possible to add the components, which map the words to their knowledge types, both principal and attributive. In this case, LI is naturally transformed into the text's semantic index, which can be treated as an efficient text's knowledge representation model, which, firstly, unlike other known models, is not dependent on extraction procedures, but can be transformed into any of these models; secondly, it provides the natural language interface to the user in order to access the extracted knowledge.

The report contains the description of the most important points of the development and implementation process of BLP.

AUTOMATIC TRANSLATION FROM BELARUSIAN INTO SPANISH BASED ON USING NOOJ'S LINGUISTIC RESOURCES

A. Veka[1], Y. Yakubovich[2]
[1] Belarusian State University, Minsk;
[2] Universitat Autònoma de Barcelona, Spain
e-mail: helena1993huk@mail.ru

Interaction of nations with one another in the current world has become so developed and widespread that knowing foreign languages today is of high importance.

Of course, it is very advisable to have a good command of foreign languages you need in your everyday communication or for work, but it is not really possible to learn all of them and have an opportunity to make a translation from one language to another by yourself. Pretty often we need a dictionary to render words, phrases or complete sentences.

Our work is aimed at helping in translating words and phrases from Belarusian into Spanish. The whole process deals with using an indrawn Belarusian-Spanish dictionary, which contains pairs of Spanish equivalents for Belarusian words and expressions, to make a rendering a Belarusian text into Spanish.

The first step was to create a dictionary itself. Belarusian-Spanish dictionary represents a set of word and expressions supplied with different kinds of grammatical information, e.g. category, inflectional and derivational paradigms (conjugated forms of verbs), data about number and gender of nouns and adjectives. Also this dictionary was completed with a number of idiomatic expressions typical for both languages. That made the translation of idioms possible.

The fact that both Belarusian and Spanish don't have a strict fixed word order facilitates a lot the translation process. Thus, it is not necessary for the machine to make complex transformations of phases or sentences. More often it is enough just to translate them word by word in order to get an adequate translation of a text under consideration.

The process of translation consists in linguistic analyzing the text, which results in finding concrete text elements, i.e. isolated words, set expressions or special names of places or notions (proper names). Then Nooj finds equivalents for the text elements found in the dictionary. After the shifting of all the items for their analogues in the other language, we can operate with a ready-made translated sentence.

The purpose of this finding is to apply it in practice. Namely, for translating texts from Belarusian into Spanish. First of all, it is innovative because there are

no many electronic resources dealing with automatic Belarusian-Spanish translation of the whole texts. Then it is typical as this country has tight relationships with some Spanish-speaking countries, e.g. Venezuela.

A FRENCH-TAMAZIGHT MT SYSTEM FOR COMPUTER SCIENCE

F. Yamouni
Mouloud Mammeri University, Tizi Ouzou, Algeria
e-mail: fariyamo@yahoo.fr

Like Systran in other times, Google hastens to explain that this new service can help the user understand the general meaning of a text in a foreign language, but does not provide accurate translations. Great are user's expectations and they are not understand very well why the machine translation does not make faster progress [1]. Today, industrial and large-audience Machine Translation software are still producing poor quality results. For example when we use Babelfish to translate the compound term: Entrées sorties physiques, we obtain: Entries physical outputs, instead physical input output.

This really, more translations are not 'word to word' for example: Mémoire vive has 4 translations: RAM (Random access memory), Computing store, Random access storage, Read write memory.

Automatic translating software needs increasingly significant and varied terminological resources. For the technical languages or of speciality, work remains to be made to build electronic dictionaries. The construction of a terminology depends on the application in which one wants to use it. The selected terms and their degree of description are different according to whether one wants to build.

Our aim is to develop a MT system for computer science compound words, from French to Tamazight.

For example Mémoire à bulles magnétiques is translated "takatut s tlilac tidkiranin", and Mémoires à bulles magnétiques gives "tikatutin s tlilac tidkiranin".

Our electronic computer science dictionary for French compounds with 10 500 entries was developed with NooJ [2–4]. The computer science dictionary for Tamazight terms (1500 entries) is built, using Saad-Bouzefrane [5] dictionary. We build a NooJ bilingual dictionary French Tamazight with the French dictionary and add the translation in Tamazight for the entries. Each entry contains information about source language (Kabyle, Chleuh, Touareg, Mozabyte or Chaoui) .We study translations for the compounds (terms with length 2, 3, 4, 5 or more). We build syntactical translation grammars [6] with for input Fr and output TM.

References

1. La traduction automatique, de Babel Fish à Google Traduction [Electronic resource]. – 2014. – Mode of access : https: //www.actualitte.com/

societe/la-traduction-automatique-de-babel-fish-a-google-traduction-46891. – Date of access : 20.12.2014.

2. Aoughlis, F. Construction d'un dictionnaire électronique de terminologie informatique et analyse automatique de textes par grammaires locales / F. Aoughlis. – Tizi Ouzou : Université Mouloud Mammeri, 2010. – 188 p.

3. Aoughlis, F. A Computer Science Electronic Dictionary for NOOJ / F. Aoughlis // Lecture Notes in Computer Science 4592. – Springer, 2007. – P. 341–151.

4. Hildebert, J. Dictionnaire des technologies de l'informatique / J. Hildebert // Français/Anglais, La maison du dictionnaire (Paris). – NY : Hippocrene Books Inc., 1998. – Vol. 2. – 1786 p.

5. Saad-Bouzefrane, S. Lexique d'informatique (Français – Anglais – Berbère), Amawal n tsenselkimt (Tafṛansist – Taglizit – Tamaziɣt) / S. Saad-Bouzefrane. – Paris : L'Harmattan, 1996.

6. Ferhi, H. Reconnaissance automatique des entités nommées arabes et leur traduction vers le français / H. Ferhi. – Sfax : Université de Sfax, 2012. – 17 p.

CONTENTS