C.R.I.T. Centre de Recherches Interdisciplinaires et Transculturelles EA 3224

16th INTERNATIONAL NOOJ 2022 CONFERENCE

Book of Abstracts

Hybrid Conference

Rosario, Santa Fe, Argentina

June 14-16, 2022

Conference Venue

ECU (Espacio Cultural Universitario, UNR)
Rosario, Santa Fe, Argentina

Silvia Susana Reyes
Andrea Rodrigo
Max Silberztein
Carolina Tramallino (Eds.)

Organized by/Organizado por

Facultad de Humanidades y Artes, UNR, Centro de Estudios de Tecnología Educativa y Herramientas informáticas de Procesamiento del Lenguaje (CETEHIPL), Centro de Estudios de Adquisición del Lenguaje, Cátedras: Lingüística General I, Lengua Española III Prof. y Lic. en Letras, Programa de investigación “Enseñar y Aprender con tecnologías. Lenguaje y Educación”. Diplomatura de Estudios avanzados en aplicación de herramientas informáticas en la enseñanza-aprendizaje de la lengua. Revista *Aprendo con NooJ*. Revista *Infosur. Investigación y Desarrollo*. IES 28 "Olga Cossettini". Asociación Civil "Los Orejanos". NooJ Association.

Declarada de Interés Municipal de la ciudad de Rosario
DECRETO N° 61.850- Expte. N° 263.057-P-2022 C.M.
Resolución Ministerial 1860/2022
Declarada de Interés Educativo por el
Ministerio de Educación de la Prov. de Santa Fe
Aprobación y acreditación dentro del Decreto 3029/12 incluyendo el Taller de Lingüística Computacional

Agradecimiento especial: Al Honorable Concejo Municipal de la ciudad de Rosario, y concejalas María Eugenia Schmuck y Nadia Amalevi. A Miguel Rabbia, Senador provincial por el Departamento Rosario y Marcelo Lewandowski, Senador Nacional. Al Director del Centro de Estudios Interdisciplinarios, UNR, Prof Darío Maiorana.

Table of Contents/Sumario

RA: A New Linguistic Engine for NooJ

Max Silberztein
Université de Franche-Comté
max.silberztein@gmail.com

I will present the new linguistic engine for NooJ. It is written in the Swift programming language, which is compatible with various operating systems (Windows, MacOS, LINUX) and is much more efficient than programming languages such as Java or C# which depend on intermediary virtual machines to run on each operating system. The linguistic engine contains functionalities to manage NooJ's dictionary and regular, context-free and context-sensitive grammars, as well as the corresponding types of automata and transducers, to parse and manage texts' annotation structure. RA does contain minor differences and incompatibilities with NooJ that I will discuss.

Morphological & Lexical Resources

The architecture of Santi-Morf's Guesser Module

Prihantoro
Universitas Diponegoro
prihantoro@live.undip.ac.id

A morpheme-level annotation system for Indonesian may struggle to analyze certain types of Indonesian words, such as proper names, misspelt words, and newly coined words. In this paper, I report the creation of the Guesser module, one of four core modules used in SANTI-morf, a new morpheme-level annotation system for Indonesian, implemented using NooJ (Silberztein 2003). To carry out morpheme level annotations, SANTI-morf uses two types of NooJ resources: morphological grammars and dictionaries. SANTI-morf always produces 100% coverage, even when applied to unrestricted texts. This stands in contrast to other existing morphological analyzers such as MorphInd (Larasati et al. 2011), in which unknown words are labeled X, and Two-Level Morphological Analyser for Indonesian (Pisceldo et al. 2008).

SANTI-morf first applies the Annotator, the core module of SANTI-morf that performs initial annotation. Words unknown to the Annotator are then passed on to the Disambiguator. Unlike the Annotator in which the root of a word is typically specified in one of the dictionaries. The disambiguator fully relies on morphological grammars to carry out the annotation of roots and other morphemes. The grammars in the Disambiguator are designed using two cues: orthographic and linguistic cues. In total, the Guesser is equipped with five morphological grammars encoded in five unique priorities. In each grammar, the rules are divided into two groups. The first group, whose rules include +UNAMB operators, is always prioritized. Thus, in total, there are ten layers of rules. Upon experimentations, this complex architecture allows SANTI-morf to achieve the optimum performance (99% precision-recall, with 100% coverage) with the smallest degree of ambiguity.

For instance, the word *diahokkan* 'to be treated like Ahok' is a relatively newly coined polymorphemic Indonesian word, composed of three morphemes: *di-* (active verb prefix), *ahok* (proper name, but in this context, the first letter is written with lower instead of uppercase), and *-kan* (causative suffix). The Guesser properly tokenises this word into three morpheme units and assigns a correct morphological tag for each morpheme.

Reflecting on the complexity of the architecture of the Disambiguator, I argue that an alternative scheme to set priorities for NooJ users is required. The scheme is simulated in this paper and when implemented, can reduce the number of grammars from five into just one. While the technical details to implement this scheme is not discussed, NooJ users, in general, will find this scheme useful, particularly those developing a large scale and full coverage annotation system, just like SANTI-morf.

References

Larasati, S. (2011). *MorphInd*. Retrieved from Larasti. Available at: https://septinalarasati.com/morphind/

Pisceldo, F., Mahendra, R., Manurung, R., & Arka, I. W. (2008). A Two Level Morphological Analyser for the Indonesian Language. In N. Stokes, & D. Powers (Eds.), *Proceedings of Australasia Technology Association Workshop* (pp. 142-150). Hobart: ACL.

Prihantoro. (2021). *SANTI-morf: A new morphological annotation system for Indonesian (PhD Thesis).* Lancaster: Lancaster University.

Silberztein, M. (2003). *NooJ Manual.* Available at: https://atishs.univ-fcomte.fr/nooj/downloads.html

NooJ Morphological Grammar of Adjectives and Adverbs for Medieval Latin

Linda Mijić
Department of Classical Philology - University of Zadar
lmijic@unizd.hr
Anita Bartulović
Department of Classical Philology - University of Zadar
abartulo@unizd.hr

This paper continues the work on natural language processing of Medieval Latin. Our corpus consists of 385 wills drawn up in the Zadar commune between 1209 and 1409 – 84 of them unpublished and 301 published ones. The latter were published over the period of more than a hundred years, with varying publishing rules applied. Therefore, we first had to correct typographical errors in the text and even up the use of different kinds of brackets within damaged words.

In our previous paper we focused on orthographic variations of words, characteristic for Medieval Latin, and we annotated common nouns from the corpus. In this paper we extend NooJ's resources with a dictionary of adjectives. Latin is a highly inflected language, and Latin adjectives have three declensions. They are inflected for number (2), case (6) and gender (3), and are divided into two declension classes (the first and second declensions, and the third declension) with further subclasses. There are three degrees of adjectives: the positive, the comparative and the superlative. Regardless of the declension of the positive adjective, the comparative belongs to the third declension, and the superlative to the first and second declensions. The comparison of adjectives is divided into four types: a) regular, b) irregular, c) periphrastic, and d) defective comparison with different subtypes.

We also include all adverbs (of place, time, and manner) in the dictionary. Adverbs are an uninflected word class. However, most adverbs of manner are formed from adjectives (i.e. *notus* > *note*, *acer* > *acriter*, *sapiens* > *sapienter*), and the comparative and the superlative of adverbs from adjective bases (i.e. adj. *felix*, *-icis*: comp. *felicior*, *felicius*, sup. *felicissimus, -a -um*; adv. *feliciter*: comp. *felicius*, sup. *felicissime*).

Finally, we propose to present morphological inflectional paradigms for the recognition of adjectives and derivation paradigms for the comparison of adjectives and the formation and comparison of adverbs of manner.

References

Mijić, L.; Bartulović, A. (2021). Formalizing Latin: An Example of Medieval Latin Wills. Bekavac, B., Kocijan, K., Silberztein, M. and Šojat, K. (Eds.). *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities*. Springer, Cham International Publishing, 24–36.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. Wiley & Sons.

Stotz, P. (1996-2004). *Handbuch zur lateinischen Sprache des Mittelalters*, 1–5. Verlag C. H. Beck.

Formalizing the Ancient Greek Participle Inflection with NooJ

Silvia Susana Reyes
CETEHIPL, Universidad Nacional de Rosario, Argentina
sisureyes@gmail.com

Ancient Greeks were particularly fond of participles. And from a linguistic viewpoint, the Ancient Greek participle constituted a complex part of speech, an all-in-one unit combining conjugation and declension, a verbal adjective. It was a non-finite verbal form expressing tense/aspect and voice, also inflected for case, gender and number, like an adjective.

The purpose of this paper is to present a preliminary formalization of the Ancient Greek participle using the Modern Greek NooJ Module. To ensure its morphological processing, automatic recognition and generation, we recovered the secondary operators of the properties.def file of the Modern Greek NooJ Module from the doctoral thesis of Lena Papadopoulou and added four definitions related to the verb: V_Mood = PAR (participle), and the three tenses, besides the PR (present), in which the participle was conjugated in Ancient Greek: FU (future), AO (aorist), and PF (perfect). The set of conjugational and declensional properties includes: V_Mood = PAR; PAR_Case = nom | acc | gen | voc | dat; PAR_Gender = n | f | m; PAR_Number = s | p; PAR_Tense = PR | FU | AO | PF; PAR_Voice = act | mid | pas | mp.

Our corpus is made up of 150 animal fables written in prose by or associated with Aesop, a Greek fabulist and storyteller, who was born around 620 BCE. His fables offer a great variety of participles in their full forms.

What follows is a brief account of the stages involved in the inflectional formalization of Ancient Greek participles. First, we extracted the participles manually from the corpus and classified them according to the specific mood suffixes (-ντ- and -μεν-) involved in the morphological formation of the present, aorist and future participles, which connect the verbal stems inflected for tense and voice with the adjectival endings indicating case, gender and number. Second, we processed some participles of the 1st or thematic conjugation derived from active verbs in -ω, and from deponent or middle-passive verbs in -μαι. Third, according to the last accented vowel/diphthong of the verbal root or stem, or to the accented stem vowel/diphthong plus consonant(s) sequences (VC, VCC) before lemmas verbal endings (present indicative first person singular verbs), we created dictionaries (.dic files) and designed inflectional grammars (.nof files) to compile these dictionaries (.nod files). Fourth, after performing Linguistic Analysis and getting the TAS, we entered NooJ regular expressions or strings of characters in the Locate window to check that participles were successfully recognized and inflected for tense (present, aorist and future), voice (active, middle, passive and middle-passive), gender (masculine, feminine and neuter), number (singular and plural), and case (nominative, vocative, accusative, genitive and dative).

References

Georganta, M. & Papadopoulou, E. (2011). Towards an Ancient Greek NooJ Module. Vučković, K. Bekavac, B. & Silberztein, M. *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik)*, Cambridge Scholars Publishing (2012).

Papadopoulou, E. (2010). *Diccionario monolingüe coordinado para enseñanza/aprendizaje del griego moderno por parte de hispanohablantes y para traducción automática griego-español*, Tesis Doctoral. Bellaterra: Universitat Autónoma de Barcelona.

Silberztein, M. (2015). *La formalisation des langues: l'approche de NooJ*. London: ISTE.

Silberztein, M. (2018). *NooJ Manual*. Available at: https://atishs.univ-fcomte.fr/nooj/downloads.html

Greek Module. Available at: https://atishs.univ-fcomte.fr/nooj/resources.html

Corpus Analysis with NooJ: Mining of Spanish Neologisms

Silvana Pierabella
Facultad de Humanidades y Artes - UNR
silvanapierabella@gmail.com

The paper presents the results obtained by using the open software NooJ (Silberztein,2016) in a research within lexicography and morphology framework of computational linguistics (Biderman,2001) with the purpose of extracting neologisms. According to Cabré, a neologism is considered a new lexical item that shows high level of adaptation to the language, either due to its frequency of occurrence or to its wide range of derivations (Cabré,1992); for example *hypertext* appeared formerly and short after derived in *hypertextual, hypertextualize, hypertextualized*. It is important to point out that neologisms are different from terms which renew exclusively scientific fields.

The methodology of corpus (Parodi,2008) is applied to 411 documents downloaded from CORLEC digital repository available on the internet. The files shared ASCII format or Byte-Marked Unicode UTF8 and belong to different domains (technology, journalism, sport, education) for pointing out the source of lexical innovations in Spanish.

First, the dictionaries developed by IES-UNR research group in charge of Andrea Rodrigo (Rodrigo-Bonino, 2019) are compared against the corpus in order to detect the lemmas that should be added with the tag [+NEO] and would be stored in "Lexical Analysis" of SP (Spanish) Module. Then, the new items are validated as neologisms from the quantitative point of view (Nazar-Vidal,2008) counting them up by means of NooJ Statistical Analysis. Finally, productive grammars are implemented in NooJ Morphology since they are based on lexical categories and allow the recognition of formal features shared by the tokens lately included in the Spanish dictionaries.

Thus, this study proves that NooJ is powerful enough to deal with corpus for updating electronic dictionaries and designing grammars that may be suitable for fulfilling complex tasks such as automatic classification of digital files, search engine improvement and organization of ontologies, taxonomies or data bases.

References

Biderman, M.T.C. (2001). *Teoría lingüística: teoría lexical y lingüística computacional.* São Paulo: Martins Fontes.

Rodrigo, A. & Bonino, R. (2019). *Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua.* Rosario: Ciudad Gótica.

Cabré, T. (1992). *La Terminología: Teoría, metodología, aplicaciones.* Barcelona: Editorial Antártida.

Nazar, R. & Vidal, V. (2008). Aproximación cuantitativa a la neología. *Actas del I Congreso Internacional de Neología en las lenguas románicas*, Barcelona.

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito RLA. *Revista de Lingüística teórica y aplicada*, 46 (1), 93-119. https://dx.doi.org/10.4067/S0718-48832008000100006

Silberztein, M. (2016). *Formalizing Natural Language: The NooJ Approach.* London: Wiley.

Automatic analysis of appreciative morphology. The case of nominal paronomasia in Colombian Spanish

Walter Koza
Universidad Nacional de General Sarmiento, Consejo Nacional de Investigaciones científicas y técnicas, San Miguel, Argentina
wkoza@campus.ungs.edu.ar
Viviana Román González
Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
viviana.roman@pucv.cl
Constanza Suy
Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
constanza.suy@pucv.cl

Within the studies of linguistic creation resources, paronomasia, understood as the substitution of one lexical item for another based on a partial homophony, has several studies focused on the stylistic phenomenon (López, 2005; Musté, 2005). However, approaches from the grammatical plane, specifically from formal theories, have been truly scarce (Bohrn, 2013), and the same can be said in relation to natural language processing. Regarding the latter, one of the problems observed refers to the difficulties of automatically assigning correct tags (both semantic and morphological), due to the ambiguity that these expressions generate. Thus, for example, 'alcamonías' (in Spanish, aromatic seasoning seeds) is a pluralia tantum word, but it used to be employed as a paronomasic form of 'alcahuete' (‘pimp’), which presents variations of gender and number (‘alcahuete’, ‘alcahueta’, ‘alcahuetes’, ‘alcahuetas’) and it can occurs in structures like those of:

a. Juan es un alcamonías [‘Juan is an alcamonías’]. (→ alcamonías: masculine, singular)
b. Juan y Pedro son unos alcamonías [‘Juan and Pedro are alcamonías’]. (→ alcamonías: masculine, plural)
c. María es una alcamonías [‘María is an alcamonías’]. (→ alcamonías: feminine, singular)
d. María y Susana son unas alcamonías. [‘María and Susana are alcamonías’]. (→ alcamonías: feminine, plural)

For this purpose, the present work analyzes a group of Colombian Spanish names with their corresponding paronomasic variants, from the Generative Lexicon Theory (Pustejovsky, 1995; 2011; 2013; Pustejovsky & Batiukova, 2019), with a view to a computational modeling for automatic recognition and generation. To do this, NooJ (Silberztein, 2016) is used. It is a tool for linguistic analysis that has various utilities such as electronic dictionaries and computer grammars.

The methodology is comprised of the following steps: (i) selection of a list of Colombian Spanish names from the work of Varela (2016); (ii) creation of an electronic dictionary with associated semantic and morphosyntactic information; (iii) elaboration of morphological grammars, and (iv) elaboration of syntactic grammars for recognition and generation. The resources are tested on a corpus of texts extracted from the internet and through a group of informants made up of native speakers of the Colombian variety. The theoretical analysis demonstrates that, in some cases, paronomasia can imply a restriction of the polysemy (billete>Villegas) [‘bill>Villegas’] or generate a new semantic structure (mano>Manuela) [‘hand>Manuela’]. At the same time, the automatic analysis yields promising results to further investigate the phenomenon from this perspective.

References

Bohrn, A. (2013). ¿Qué me contursi? Mi mujica se fue con un vizcacha. Paronomasia en el español rioplatense. Kornfeld, L. & Kuguel, I. (Eds.). *El español rioplantense desde una perspectiva generativa*.
López, C. (2005). *La paronomasia como recurso conceptual, expresivo y humorístico en la lengua española actual*. Ph.D. Thesis: Universidad de Granada.
Musté, P. (2005). *Análisis lingüístico de un corpus de eslóganes y frases publicitarias de marcas comerciales*. PH.D. Thesis: Universidad Politécnica de Valencia.
Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, Massachusetts: MIT Press.
Pustejovsky, J. (2011). Coercion in a general theory of argument selection. *Linguistics*, 49(6), 1401-1431.
Pustejovsky, J. (2013). Type theory and lexical decomposition. Pustejovsky, J., Bouillon, J., Isahara, P., Hanzaki, H. & Lee, C. (Eds.). *Advances in Generative Lexicon Theory*. Nueva York: Springer (pp. 9-38).
Pustejovsky, J. & Batiukova, O. (2019). *The lexicon*. Cambridge: Cambridge University Press.
Silberztein, M. (2016). *Formalizing Natural Languages. The NooJ approach*. Londres: ISTE.
Varela, D. (2016). *Un sistema peculiar de creación de palabras en español: el caso de la homonimia parasitaria*. Ph.D. Thesis: Universidad Autónoma de Madrid.

A computational approach to recognition of deverbal nouns ending in <sión>

María José González
Administración Nacional de Educación Pública - Consejo de Formación en Educación, Uruguay
mgonzalez.uy@gmail.com

The synchronic morphology usually presents restrictions for the analysis of deverbal nouns ending in *-ión* and establishes a proliferation of suffixes such as *-ión, -sión, -ción.* The examination of this type of nominals is usually based on existing verbs in Spanish. And thus, the analysis is quite limited, since many of these nouns result from a Latin suffixation process and they are inherited from this language into Spanish. For example, a noun such as *cohesion* has lost the semantic relationship with **coherir* and therefore since there is no *coherir* in Spanish, the verb *cohesionar* has been created. However, it is possible to search this form in the history of Spanish and recover the Latin verb *cohaerere* and its participle *cohaesum* from which *cohesion* comes. The prioritization of form over content, that is, the prior consideration of the form versus semantic relationship and the diachronic approach allow the establishment of regular models of deverbal nouns. This means proposing bases such as **he´s* even if it does not exist in current Spanish.

The objective of this work is to demonstrate that it is possible to establish these models for the morphological analysis of deverbal nouns ending in <sión> from a computational approach that allows the automatic recognition of these formations by using free access software. From the point of view of diachronic morphology, the particularity of these nominals is that they are the result of Latin words formation rules. That is, its analysis reveals in all cases Latin bases, coming from the participial theme ending in <s>, selected by a single derivative morpheme *-ión.*

To process the morphological analysis, the NooJ tool will be used, a software designed by Max Silberztein to formalize and automate natural language. Based on the possibilities offered by the system to create derivational morphological grammars, the different models will be described considering the bases – often non-existent in modern Spanish – that account for the formation of these nouns. We will work with the Spanish module available in the tool. First, the respective morphological grammars will be created, then a dictionary of deverbal nouns that links these grammars will be designed and later the automatic analysis will be carried out from a *corpora* made up of twenty encyclopedia entries.

The advantage of the identification of these nominals with the design of morphological grammars is that, based on models created from the computational approach, it is possible to recognize all the nouns that satisfy the property of ending in <sión>. It also allows items to be grouped around the Latin base, “etymological paradigms” (Pena, 1999).

References

Bosque, I. & Demonte, V. (Eds.). (1999). *Gramática descriptiva de la lengua española*. Madrid: Espasa-Calpe.

Pena Seijas, J. (1980). La derivación en español. Verbos derivados y sustantivos deverbales. *Anejos de Verba*. Santiago de Compostela: Universidad de Santiago.

Pena Seijas, J. (1999). Partes de la morfología. Las unidades del análisis morfológico. Bosque, I. & Demonte, V. (Eds.). (1999). *Gramática descriptiva de la lengua española*. Madrid: Espasa-Calpe. (4305-4366).

Real Academia Española & Asale. (2009-2011). *Nueva gramática de la lengua española*. Madrid: Espasa. [NGLE]

Silberztein, M. (2003). *NooJ Manual.* https://atishs.univ-fcomte.fr/nooj/downloads.html

Tramallino, C. P. (2013). Análisis morfológico con herramientas informáticas. Reconocimiento de nombres en textos en español con el sistema NooJ. *Revista Lingüística y Literatura.* 63, 33-48.

Varela, S. (2005). *Morfología léxica: la formación de palabras.* Madrid: Gredos.

Syntactic & Semantic Resources

Ukrainian productive morphological grammars for recognizing unknown units

Olena Saint-Joanis
ELLIADD, Université Bourgogne Franche-Comté - CREE, INALCO, Paris, France
alena.saintjoanis@gmail.com

This paper aims to present productive morphological grammars for the Ukrainian module for NooJ that recognizes unknown units.

The Ukrainian language is very rich in vocabulary and derivations are numerous and varied. Moreover, oral variants, neologisms, or even parallel forms are also numerous and can have their own paradigms different from the paradigms of the normative language.

The lexical units of some classes are constructed mainly by derivation of the lexical units of other classes, as is the case particularly for the adjectives. The dictionary of the Ukrainian language for NooJ contains only 14,719 adjectives, which is far from being enough. To compare the online dictionary - https://goroh.pp.ua - lists 89,898 adjectives or adjectival forms (including diminutives and comparatives). So we must find a way to complete our dictionary.

Therefore, we have two solutions:

1) find the missing adjectives, then manually add them to our dictionary and adjust them to their paradigms (FLX) and derivations (DRV) for comparatives, superlatives and diminutives;

2) build productive morphological grammars capable of recognizing adjectives, not included in the dictionaries.

The first solution seems to us very time consuming, so we opt for the second.

Thus, we form grammars capable of producing adjectives derived from nouns, verbs, and other adjectives and displaying the lemma of these adjectives as well as all associated annotations. We know that the adjective is declined in gender, in number and in case, and, therefore, our grammars also contain all the elements to recognize the flexed forms.

We build grammars that are capable of producing nouns, participles, adverbs and derived verbs. We also treat the lexical units formed by the contraction of two lemmas in the same way. This work seems particularly useful to us because it allows not only to recognize the missing lexical units but also to add specific annotations in particular: "neologism", "oral variant", "parallel form" etc.

References

Gorpynyč, V. O. (2004). *Morphologiya ukraïnskoï movy*. Kyïv: Akademiya.

Plušč, M. Y. (2010). Gramatyka ukraïnskoï movy. Čatyna 1. Morfemika. Slovotvir. *Morfologiya. Pidručnyk dlya studentiv filologi čnyh spetsialnostei vyščyh nav čalnyh zakladiv*. Kyïv: Vyšča škola.

Silberztein, M. (2015). Joe loves Lea: Transformational Analysis of Transitive Sentences. *Formalising Natural Languages with NooJ* (9th International NooJ conference, Minsk, Belarus 2015). CCIS Series. Springer Verlag: Heidelberg (2016).

Vyhovanets, I. R. & Gorodenska, K. G. (2004). *Teorretyčna morfologiya ukraïnskoï movy*. Kyïv: Pulsray.

https://goroh.pp.ua/ - a public electronic library, which contains the main dictionaries of ornaments.

Annotation of procedural questions in standard Arabic using syntactic grammars

Essia Bessaies
RIADI, ENSI, University of Manouba Tunisia
essiabessaies@gmail.com
Slim Mesfar
RIADI, ENSI, University of Manouba Tunisia
mesfarslim@yahoo.fr
Henda ben Ghazela
RIADI, ENSI, University of Manouba Tunisia
henda.benghezala@ensi.rnu.tn

Most question-answering systems have been designed to answer short questions (precise answers such as dates, locations), and only a few researches concern complex questions.

In this paper, we present a method for analyzing medical procedural questions. The analysis of the question asked by the user by means of a pattern based analysis covering the syntactic as well as the morphological levels. These linguistic patterns allow us to annotate the question and its semantic features for extracting the focus and topic.

We start with the implementation of the rules which identify and annotate the various medical named entities. Our named entity recognizer tool (NER) is able to find references to people, places and organizations, diseases, viruses, as targets to extract the correct answer from the user. The NER is embedded in our question answering system.

The task of our system is divided in four phases: question analysis, segmentation and passage retrieval, answers validation and finally answers extraction. Each phase plays a crucial role in overall performance. We use the NooJ platform which represents a valuable linguistic development environment. The first evaluations show that the actual results are encouraging and could be deployed for further question types.

References

Low, B.T., Chan, K., Choi, L.L., Chin, M.Y & Lay, S.L. (2001). Semantic expectation-based causation knowledge extraction: A study on Hong Kong stock movement analysis. Cheung, D., Graham, J. W., & Qing Li (Eds.). *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science.* 2035. Berlin: Springer.

Khoo, C., Chan, S. & Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. *Proceedings of 38th Annual Meeting of the ACL.* 336–343.

Mesfar, S. (2007). *Named Entity Recognition for Arabic Using Syntactic Grammars.* NLDB.

Mesfar, S. (2008). *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard.* Phd Thesis, Franche-Comte University, France.

Silberztein, M. (2006). *NooJ Manual.* https://atishs.univ-fcomte.fr/nooj/downloads.html

Silberztein, M. (2015). *La formalisation des langues: l'approche de NooJ*. Londres: ISTE.

Silberztein, M. (2016). *Formalizing Natural Languages: the NooJ approach*. Hoboken NJ: Wiley-ISTE.

Zellig S. Harris' Transfer Grammar and its application with NooJ

Mario Monteleone
Dipartimento di Scienze Politiche e della Comunicazione - Università di Salerno, Italy
mmonteleone@unisa.it

In 1954, with his article entitled "Transfer Grammar" (published in “The International Journal of American Linguistics”, Vol. 20, No. 4, pp. 259 -270, University of Chicago Press), Zellig S. Harris was the first linguist to approach the nascent Automatic Translation (AT) from the point of view of structuralist and formal linguistics. This article, written in the pivotal period for the first TA attempts in the US, outlines a translation method that wants to:

- Formally measure the difference between languages, in terms of grammatical structures;
- Define the point of minimum difference (or maximum similarity) between any type of language pair;
- Define the difference between the languages as the number and content of grammatical instructions needed to generate the utterances of one language from the utterances of the other.

At the time, the purposes of Harris's article were therefore extremely innovative, since they considered translation as a process in which meaning transfers could only be achieved based on morphosyntactic analyses and evaluations. Moreover, it is worth stressing that at the time the first TA experiments performed word-for- word translations, without taking into account (not even statistically) the contexts in which the words co- occurred. As is known, this method proved to be unsuccessful, as regards the quality, time and costs of the translations made automatically. In 1966, this led ALPAC to end TA research in the US, and cut off the flow of funding to it.

As for contemporary times, the most AT used method is the statistical one of the recent Neural Machine Learning (NML), which is based on Neural Transition Networks (NTM), but which fails to guarantee constant satisfactory results, as demonstrated by portals like Google Translate and Reverso. On the other hand, what became of the method identified by Harris? Up to today, it has remained unrealized. During the 90s of the last century, there have been several important experiments regarding TA based on the formalization of language morphosyntax. All these experiments, including that of the Logos system, however, have long since been abandoned. At the same time, as we intend to demonstrate here, currently the only software tool capable of implementing Harris's TA method is NooJ. In fact, we will see how NooJ FSA / FSTs, if used for AT, can apply and develop the aforementioned methodological steps indicated by Harris. Furthermore, we will again demonstrate how also in its TA grammars NooJ can use Lexicon-Grammar methodological tools, which are able to formalize morphosyntax using always the same parameters for all natural languages, hence allowing a more straightforward automatic translation between language pairs. Therefore, the final aim of this study will be to demonstrate, with practical examples, how NooJ TA grammars are the most suitable application tool of Harris' Transfer Grammar.

References

Harris, Z. S. (1946). From Morpheme to Utterance. *Language.* 22, 161-83.
Harris, Z. S. (1951). *Structural linguistics*. Chicago: University of Chicago Press.
Harris, Z. S. (1954). Transfer Grammar. *The International Journal of American Linguistics.* 20-4, 259-270.
Harris, Z. S. (1970). *Papers in Structural and Transformational Linguistics.* Dordrecht: D. Reidel Publishing Company.
Harris, Z. S. Website Homepage www.zelligharris.org.
Silberztein, M. (2003). *The NooJ Manual*. Available at: https://atishs.univ-fcomte.fr/nooj/downloads.html
Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. London: ISTE-Wiley-Sons Inc. EAN: 9781848219021.

Formation and Evolution of Intensive Adverbs in *-mente* derived from the Adjectival Class <Causatives of Feeling>

Rafael García Pérez
Universidad Carlos III de Madrid
rafael.garcia.perez@uc3m.es
Xavier Blanco Escoda
Universitat Autònoma de Barcelona
Xavier.Blanco@uab.cat

In this paper, we intend to study the constitution and the evolution of a very specific group of Spanish intensive adverbs: the adverbs in *-mente* derived from adjectives belonging to the semantic class <causatives of feeling> (Blanco, 2006).

From a historical point of view, the connections between the meaning of manner conveyed by the adverbs in *-mente* and their meaning as intensifiers are still quite close. Nevertheless, the development of some intensifiers has been the result of a semantic and discursive reinterpretation which, in a broader sense, is considered a part of a grammaticalization process (Hopper & Traugott, 2003; Rhee, 2016). It has already been shown in García Pérez (2022) that, as is often the case in the evolution of other lexical classes (for instance, in partially limiting focusing adverbs: García Pérez, 2013), not all the intensifiers have undergone a simultaneous grammaticalization process. The construction of the paradigm entails rather a progressive enrichment, with variable lexical-semantic acquisitions and losses throughout history. Moreover, it is interesting to note that, although the adjectival class <causatives of feeling> can be considered quite homogeneous from a semantic point of view, not all adverbs in *-mente* derived from it have given rise to intensifiers. Some of the adjectival subclasses have been more prone than others to this type of formation: for example, the subclass <caus_fear> (*terrible* 'terrible'– *terriblemente* 'terribly'; *tremendo* 'awful' – *tremendamente* 'awfully'…) and, to a lesser extent, the subclass <caus_surprise> (*asombroso* 'astonishing' – *asombrosamente* 'astonishingly'). This research adopts a contrastive perspective. The evolution of Spanish adverbs will first be described and then the results will be compared to those derived from the analysis of the French corpora in order to determine similarities and/or differences between the two languages. It should not be forgotten that, although French and Spanish share a Romance core, they have also been attached to diverse sociocultural and political contexts. This communication is part of the COLINDANTE research project (PID2019-104741GB-100), *Ministerio de Ciencia e Innovación* (Spain), which aims to study intensive collocations in medieval French and Spanish (Blanco & García Pérez, 2021). The implementation of the lexicons and part of the corpus is carried out through the NooJ linguistic engineering platform (Silberztein, 2016).

References

Blanco, X. (2006). Un inventario de clases semánticas para los adjetivos predicativos de estado. *Verba*, 33, 235-260.

Blanco, X. & García Pérez, R. (2021) Las estructuras comparativas intensivas aplicadas al adjetivo *negro* en español medieval en comparación con el francés. *Romanica Olomucensia,* 33(1), 21-39.

García Pérez, R. (2013). La evolución de los adverbios de foco particularizadores. *Iberoromania*, 77, 90-107.

García Pérez, R. (2022). Fuertemente atados: adverbios intensificadores en *-mente* y colocaciones en castellano medieval. *ELUA*, 35, 273-292.

Hopper, P. J., and Traugott, E. C. (2003 [1993]) *Grammaticalization*, Cambridge: Cambridge University Press.

Rhee, S. (2016). On the emergence of the stance-marking function of English adverbs: A case of intensifiers. *Linguistic Research*, 33(3), 395-436.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*, London: Wiley-ISTE.

Detection of adjunct prepositional phrases expressing time in the output of Estonian students of Spanish as third language: a computational approach

Virginia Rapún Mombiela
Universidad de Tartú, Estonia
virginia.rapun@ut.ee
Carolina Tramallino
Universidad Nacional de Rosario, Argentina
carolinatramalllino@gmail.com
Romina Arnal
Universidad Nacional de Rosario, Argentina
arnalromina@gmail.com

This article studies temporal adjunct prepositional phrases produced by Estonian students of Spanish as third language (L3+). L3+ is understood here as a non-native language that is being used or learnt by a subject with knowledge of one or more native languages and one non-native language at least (Hammarberg, 2010, p. 97). We will place the focus on structures that modify the verb, that are made up by a prepositional phrase, and that express a temporal reference (*at 5 in the morning*, *from 2 pm*…). These structures include prepositions and articles, two elements that have been shown to pose great difficulties to Estonian students (Kruse, 2018; Nikitina, 2019). It is not surprising, since articles do not exist as a part of speech in Estonian and, although the language has prepositions, cases or postpositions are preferred (Kruse, 2018, pp. 126-127). Samples used for this study are taken from first- and second-year undergraduate students of the Degree in Spanish Language and Literature of the University of Tartu (Estonia). Informants are mainly female individuals aged between 18 and 23. They are plurilingual speakers whose native language is Estonian. All of them have passed the entrance examination for university, meaning that they have an intermediate level of English and have learnt it for 10 years at least. Our work tool is the program NooJ. This software will enable us to create two syntactic grammars: one to recognize adjunct noun and prepositional phrases within verb phrases containing time nouns in Spanish; another to describe the same structures in the students' interlanguage. Interlanguage is here defined as a particular linguistic system that varies as the learner acquires the target language, "an intermediate system between the code of the native language and that of the target language" that has its own characteristics (Alexopoulou, 2011: 87-89). After carrying out this operation, we will be able to automatically identify and count students' utterances that coincide either with Spanish language or their interlanguage. By doing so, we aim to take into consideration the total number of productions of first- and second-year students, as well as to describe the stage of both groups in their attempt to approach the final goal (Alexopoulou, 2011: 88). After analyzing the results, in the conclusion, we will talk about possible interlinguistic and intralinguistic causes for the appearance of idiosyncratic structures and we will discuss the implications of formal instruction based on the needs of the students. Automatic analysis will enable us to validate the proposed hypothesis, according to which the interlanguage of students whose first language lacks articles and does not use prepositions presents a higher amount of idiosyncratic structures in which these parts of speech are involved.

References

Alexopoulou, A. (2011). La función de la interlengua en el aprendizaje de lenguas extranjeras. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 9, 86-101.

Hammarberg, B. (2010). *The languages of the multilingual: Some conceptual and terminological issues*. 48(2-3), 91-104. https://doi.org/10.1515/iral.2010.005

Kruse, M. (2018). *La transferencia en personas plurilingües. Los falsos amigos como un obstáculo y una oportunidad en la enseñanza y aprendizaje de lenguas extranjeras.* [Tesis]. Universidad de Tartu.

Nikitina, I. (2019). *Análisis de errores en el aprendizaje de los artículos en español por parte de aprendientes de habla estonia y rusa*. Unpublished MA thesis, University of Tartu.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. ISTE.

Tramallino, C. P. (2021). Avances en el tratamiento computacional en corpus de aprendientes de español como lengua segunda y extranjera. *Quintú Quimün. Revista de Lingüística*, *5*, 051.

Williams, S. & Hammarberg, B. (1998). Language Switches in L3 Production: Implications for a Polyglot Speaking Model. *Applied Linguistics*, 19(3), 295-333 https://doi.org/10.1093/applin/19.3.295

Directed Motion Verbs in the Interlanguage of Spanish learners: Automatic Analysis using the NooJ system oriented to the Learning of Spanish as a Second and Foreign Language

Romina Arnal
Universidad Nacional de Rosario, Argentina
arnalromina@gmail.com
Carolina Tramallino
Universidad Nacional de Rosario, Argentina
carolinatramalllino@gmail.com
Virginia Rapún Mombiela
Universidad de Tartú, Estonia
virginia.rapun@ut.ee

This work is part of a research project about teaching Spanish as a Second and Foreign Language (ELSE by its acronym in Spanish) through computer tools. It aims at achieving the automatic recognition of verb phrases (VP) constructed with directed motion verbs (Demonte, 2011) in a corpus of written texts produced by Spanish learners whose mother tongue is different. These verbs describe a movement direction to a goal and require prepositional complements to indicate the end point, such as: *ira Argentina* (go to Argentina), *volver a París* (come back to Paris). These types of structures are idiosyncratically formed in written texts belonging to Spanish learners who have different mother tongues and are at the beginning level of learning (according to the Common European Framework of Reference for Languages - CEFR). In the analyzed corpus, the absence of preposition or its replacement by another preposition that cannot be combined with the verb are observed. Therefore, this study has the goal of detecting these structures commonly found in the first stages during the Spanish language acquisition process. It should be clarified that the research is part of the interlanguage (IL) current (Selinker, 1972; Alexopolou, 2010), which is defined from a psycholinguistic approach as a system that is in an intermediate position between the mother tongue and the studied language in which some elements are common to both systems. In this way, learners go through stages of competence levels that modify that system as they acquire lexis and new structures of the language they are learning. Furthermore, IL has the particularity of being systematicity established by its internal coherence at a certain point of its development. This is detected by the application of linguistic rules that respond to strategies and processes that are activated by the learners. For this reason, the entire production is taken into consideration and the aim is to automatically recognize both: the structures that coincide with the target language (Spanish) and those that deviate from it. In order to do this, the NooJ system (a software developed by Max Silberztein in 2002) was used to distinguish them by means of a label. The work methodology consisted of a linguistic description of the VPs in Spanish, the directed motion verbs and the combination with prepositions that indicate an endpoint: *a, hacia, hasta* (to, towards, until). The NooJ software allows the user to analyze natural languages from their lexico-semantic, syntactic and morphological aspects. In order to detect verbal phrases with prepositional motion verbs, a dictionary of verbs with the features [+movement] [+direction] was first created within the Spanish module (Argentina) and then, productive syntactic grammars of graphs were generated to detect the prepositional phrases selected by this type of verbs. In this way, the structures coinciding with the Spanish system were located and differentiated from other IL structures found in the corpus. Then, all the verbal phrases with directed motion verbs were counted in order to draw conclusions about the number of matching structures and interlanguage structures in the samples produced by beginning students. It is expected that these syntactic grammars may contribute to the learning of the relations between the Spanish lexicon and syntax.

References

Alexopoulou, A. (2010). La función de la interlengua en el aprendizaje de lenguas extranjeras. *Revista Nebrija de Lingüística aplicada*, 9.

Cano Aguilar, R. (Coord.) (1999) 29. Los complementos de régimen verbal en Bosque y Demonte. *Gramática Descriptiva de la Lengua Española, 2. Las construcciones sintácticas fundamentales. Relaciones temporales, aspectuales y modales.* Madrid: Espasa-Calpe, colección Nebrija y Bello.

Demonte, V. (2011). Los eventos de movimiento en Español: construcción léxico-sintáctica y micropará-metros preposicionales. Cuartero Otal, J., García Fernández, L. & Carsten Sinner (Hg.). *Estudios sobre perífrasis y aspecto*. München: Peniope (248S.) (Etudes linguistiques/Linguistische Studien, 5)

Selinker, L (1972). Interlanguage. *International Review of Applied Linguistics*, 10(3), 209-231.

Tramallino, C. & Arnal, R. P. (2021). Reconocimiento de sintagmas nominales construidos con indefinidos a través del sistema NooJ en corpus de español como segunda lengua. *Revista IRICE*, (38), 129-162. Available at: https://ojs.rosario-conicet.gov.ar/index.php/revistairice/article/view/1310

Tramallino, C., Beltrán, C.& Ricciardi, N. (2021). Localización y contabilización de sufijos nominales en corpus de aprendientes de español como segunda lengua. *Entrepalavras*, 11 (10esp), 412-436. doi: http://dx.doi.org/10.22168/2237-6321-10esp2116

Approach to locative constructions: using NooJ to automatically analyze Spanish *donde* (where)-structures

Silvina Lorena Palillo
Universidad Nacional de Rosario
silvinapalillo@yahoo.com.ar
Andrea Fernanda Rodrigo
CETEHIPL, Universidad Nacional de Rosario
andreafrodrigo@yahoo.com.ar

In the following work, we will automatically analyze two possible scenarios of locative constructions headed by *donde* (where*):* one following the precepts of traditional grammar and the other one following the precepts of generative grammar. According to the precepts of traditional and structural grammars, constructions such as *Voy <donde digan>* (I go wherever you say) are classified as adverbial clauses of place. These clauses are dependent, i.e. they add information about where the action described in the main clause takes place (cf. Gili Gaya, 1951). Adverbial clauses of place are grouped together with adverbial clauses of time and adverbial clauses of manner since they function as adverbs of place, time and manner, respectively. Following this, the adverbial clause of place *<donde digan>* (<wherever you say>) can be replaced by the Spanish adverb of place *<allí> (*<there>).

On the other hand, following the precepts of generative grammar (cf. Bosque and Demonte, 1999; Real Academia Española, 2009; Bosque, 2019), these constructions are recategorized as subordinate relative clauses since there is an elided nominal antecedent (*Voy Ø <donde digan>)* (I go Ø wherever you say). This shows us that there is no direct relationship among function, meaning and structure. Instead, we are dealing with structures with locative meaning which are made up by a relative clause.

It is possible to redraft such analysis if we slot in the contributions made by Natural Language Processing (NLP). Using NooJ designed by Silberztein (2016), it was possible for us to represent the first and the second analyses since NooJ's formalisms are compatible with any of the grammars previously mentioned. Thus, in accordance with the dictionaries created by the IES_UNR team in the Spanish Module_Argentina, we designed two possible grammars which adjust to both proposals. We then applied these grammars to a series of randomly selected newspaper articles to validate the proposed grammars. In this way, we managed to extract not only sentences but also syntagms which are compatible with such structures. These results enable us to expand and enrich the resources in the Spanish-Argentina module within the frame of the Centro de Estudios de Tecnología Educativa y Herramientas Informáticas de Procesamiento del Lenguaje (CETEHIPL)

References

Bosque, I. & Demonte, V. (2009). Gramática descriptiva de la lengua española. (1999) *RAE-ASALE Nueva gramática de la lengua española.* Madrid: Espasa-Calpe.

Bosque, I. (2019). *Glosario de términos gramaticales*. Salamanca: Ediciones Universidad de Salamanca.

Gili y Gaya, S. (1951). *Curso superior de sintaxis española*. Barcelona: Spes.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. London: ISTE Ediciones.

Spanish Module_Argentina: https://atishs.univ-fcomte.fr/nooj/resources.html

Localization of passive voice structures in corpus of scientific research articles in Spanish using the NooJ system

Carolina Paola Tramallino
Universidad Nacional de Rosario
carolinatramalllino@gmail.com
Celina Beltrán
Universidad Nacional de Rosario
cbeltran2510@gmail.com
Romina Paola Arnal
Universidad Nacional de Rosario
arnalromina@gmail.com

In this paper, we propose to use the open access software NooJ, available online, to automatically analyze certain syntactic structures characteristic of written academic texts. The work methodology, which is based on corpus linguistics (Parodi, 2004; 2008), consists of comparing three samples of scientific research articles published in Spanish-language journals indexed in the Scopus and Scielo portals coming from different areas of science; on the one hand, Social and Humanistic Sciences and, on the other hand, Health Sciences (Medicine, Biochemistry, etc.), and finally, Engineering. For this purpose, productive syntactic grammars were made from dictionaries and grammars specific to the Spanish module (Argentina). As for the corpus, it is of textual type, gathering three samples of 30 texts extracted from three different journals of each scientific area for a total of 90 articles. Regarding the scientific research article, this is constituted as one of the most researched academic genres in the last two decades (Rodríguez Hernández & García Valero, 2015; Meza, 2016) due to its impact on the transmission of specialized knowledge, for the benefit of the advancement of science. Therefore, it is important to note that the authors of such articles resort to various rhetorical-discursive strategies to persuade their readers. One of these resources is the use of the passive voice to expose the facts in an objective manner. The present research proposes a description of the syntactic structures in passive voice found in corpus texts written by native speakers of Spanish. It is considered of utmost importance to obtain, by means of automatic localization, the index of passive voice structures in academic texts in order to be able to draw conclusions about the type of structures used in scientific writing, taking into account the specificity of the three corpuses analyzed. It is necessary to mention that this work is part of a research project on assistive digital tools in the construction of academic texts in Spanish.

The software tool used was designed for natural language processing and developed by Max Silberztein in 2002. This program has the possibility of performing morphological, syntactic and semantic analysis. It has three types of files: dictionaries, grammars and properties. In order to computationally detect passive voice structures, grammars must be created to recognize them. According to Silberztein (2003, p.55), they are Productive because they use categories, for example: V *ser* ("verb" *ser*) + *ppio* stands for participle and they are called Syntactic because they group two or more words and then label them according to the user's indication, for example: VP stands for "Passive voice" in Spanish. After applying the syntactic grammars, using the Locate function, all the passive voice structures present in the corpus will be located and counted. Finally, the results will be measured according to the statistical analysis tool provided by the software. These will show significant differences in the comparison of the samples in favor of the area corresponding to Engineering.

References

Meza, P. (2016). El posicionamiento estratégico del autor. *Artículos de Investigación: una propuesta para su estudio. Forma y Función.* 29(2), 111-134.

Miguel, E. (2004). La formación de pasivas en español: análisis en términos de la estructura de qualia y la estructura eventiva. Available at: https://www.researchgate.net/publication/298806850_La_formacion_de_pasivas_en_espanol_analisis_en_terminos_de_la_Estructura_de_Qualia_y_la_Estructura_Eventiva

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. *RLA. Revista de lingüística teórica y aplicada.* 46(1), 93-119. https://dx.doi.org/10.4067/S0718-48832008000100006

Rodríguez Hernández, M. & García Valero, M. (2015). Escritura de textos académicos: dificultades experimentadas por escritores noveles y sugerencias de apoyo. *CPU-E, Revista De Investigación Educativa.* 0(20), 249-265.

Silberztein. (2003). *NooJ Manual.* Available at: https://atishs.univ-fcomte.fr/nooj/downloads.html

Tramallino, C. & Arnal, R. P. (2021). Reconocimiento de sintagmas nominales construidos con indefinidos a través del sistema NooJ en corpus de español como segunda lengua. *Revista IRICE.* (38), 129-162. Available at: https://ojs.rosario-conicet.gov.ar/index.php/revistairice/article/view/1310

Finding English Phrasal Verbs in Large Corpora

Peter A. Machonis
Florida International University
machonis@fiu.edu

Very large corpora can be statistically analyzed in NooJ to determine fluctuations in English phrasal verb usage for individual authors, find the most frequently used phrasal verbs within a given corpus, and shed light on how author bias, subject matter, use of dialogue, and other stylistic features may be contributing factors to phrasal verb usage. This paper examines four NooJ corpora of British and American writers from the mid-19th century to early 20th century. All of the novels and some novellas of Charles Dickens, Herman Melville, Thomas Hardy and Edith Wharton (from 12 to 20 for each author) were downloaded from Project Gutenberg and used to produce author corpora in NooJ. These corpora thus give an accurate idea of each author's tendency to use phrasal verbs across a broad spectrum of their work.

As per the analysis of the novels of Dickens and Melville (Machonis 2020), we limited phrasal verb searches to six typical particles representing three levels of phrasal verb frequency: high (*out*, *up*), mid (*down*, *away*), and low (*back*, *off*). This method previously used by Hiltunen (1994), attempts to give an accurate picture of phrasal verb usage, without creating excessive noise with prepositions and idiomatic expressions.

Although we previously showed that the British novelist Dickens uses more phrasal verbs than his American counterpart Melville, preliminary results show that the early 20th century American author Edith Wharton uses more phrasal verbs per 1,000 words than the late 19th century British author Thomas Hardy. Furthermore, we find little difference in phrasal verb usage between the three classes of Hardy's novels: Novels of Character and Environment, Romances and Fantasies, and Novels of Ingenuity. Thus, phrasal verb usage may very well be attributed to author predisposition, rather than subject matter or national origin. We conclude with an analysis of the most frequently used phrasal verbs by each author to shed further light on the rise and fall of certain phrasal verbs in the history of the English language.

References

Hiltunen, R. (1994). On Phrasal Verbs in Early Modern English: Notes on Lexis and Style. Dieter Kastovsky (Ed.). *Studies in Early Modern English.* Berlin: Mouton de Gruyter. (129-140).

Machonis, P. A. (2021). Where the Dickens are Melville's Phrasal Verbs? Bekavac B., Kocijan K., Silberztein M. & Šojat K. (Eds.). *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. Communications in Computer and Information Science,* 1389, 99-110. Cham: Springer.

Lexicon-grammar tables for discontinuous Arabic frozen expressions

Asmaa Kourtin
Computer Science Research Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
asmaa.kourtin@yahoo.fr
Asmaa Amzali
Computer Science Research Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
asmamzali@hotmail.fr
Mohammed Mourchid
Computer Science Research Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
mourchidm@hotmail.com
Abdelaziz Mouloudi
EDPAGS Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
mouloudi_aziz@hotmail.com
Samir Mbarki
EDPAGS Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
mbarkisamir@hotmail.com

The frozen expressions play a very important role in natural language processing and have attracted the attention of several researchers in the last few years, leading to many researches for different languages. Indeed, different definitions have been provided to the frozen expressions from the syntactic and semantic points of view by Dubois, Maurice Gross, Gaston Gross, etc. [1-2]. From these definitions, we have adopted that the term "frozen" is reserved for expressions whose global meaning is not deduced by joining the meanings of its components; it is a group of words which, in their entirety, form a meaning that is coming from the accord of a group of linguists. The Arabic language is very rich in frozen expressions which it inherited from the pre-Islamic era and early Islam, and whose use has persisted to this day. We use them in the daily communication language of Arabic speakers, and in the works of writers and poets. We can find these expressions dispersed in Arabic books, such as the Quran, the linguistic heritage and literary books, the books of proverbs, etc., which has led some researchers to collect, classify and explain them. Indeed, several classifications have been proposed according to the needs of each linguist, such as continuous or discontinuous expressions, expressions that do not admit variations, expressions allowing variations, etc. This work presents a continuation of our previous work about the creation of lexicon-grammar tables of frozen expressions that are continuous and do not admit variations, such as "مسك الختام" (misku al-khitam; Save the best for last) [3]. Our aim is to create, for the modern Arabic language, lexicon-grammar tables of discontinuous frozen expressions. For that, we start by collecting and studying those expressions, then we transform their lexicon-grammar tables into dictionaries [4] and syntactic grammars in the NooJ platform [5], allowing us to recognize and annotate these expressions in texts and corpora even if they are discontinuous. This recognition will help to solve many problems related to automatic natural language processing.

References

[1] Dubois, J. (2002). *Lexis: Larousse de la langue française*.
[2] Gross, M.(1993). Les phrases figées en français. *L'Information Grammaticale*, 59, 36-41.
[3] Kourtin, A., Amzali, A., Mourchid, M., Mouloudi, A.& Mbarki, S.(2021). Lexicon-Grammar Tables for Modern Arabic Frozen Expressions. Bigey M., Richeton A., Silberztein M. & Thomas I. (Eds.). Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2021. *Communications in Computer and Information Science.* 1520. Springer, Cham. https://doi.org/10.1007/978-3-030-92861-2_3.
[4] Kourtin, A., Amzali, A., Mourchid, M., Mouloudi, A. & Mbarki, S. (2020). The Automatic Generation of NooJ Dictionaries from Lexicon-Grammar Tables. Fehri H., Mesfar S. & Silberztein, M. (Eds.). Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications. NooJ 2019. *Communications in Computer and Information Science*, 1153. Springer, Cham.
[5] Silberztein, M. (2015.). *La formalisation des langues, l'approche de NooJ*. London: ISTE.

Syntactic analysis of complex sentences containing Arabic psychological verbs in NooJ Platform

Asmaa Amzali
Computer Science Research Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
asmamzali@hotmail.fr
Asmaa Kourtin
Computer Science Research Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
asmaa.kourtin@yahoo.fr
Mohammed Mourchid
Computer Science Research Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
mourchidm@hotmail.com
Abdelaziz Mouloudi
EDPAGS Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
mouloudi_aziz@hotmail.com
Samir Mbarki
EDPAGS Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
mbarkisamir@hotmail.com

Nowadays, the natural language processing NLP is very important and one of the most active research areas in data science. It is used in many applications that make our lives easier. Indeed, the natural language requires a syntactic analysis as one of the basic steps to the advanced natural language processing, because the crucial part of understanding a sentence is to understand all its sequences of ALUs.

In our previous work [1], we realized a syntactic analyzer of simple sentences containing Arabic psychological verbs using NooJ platform allowing to detect sentences with different word orders, such as "كره زيد أحمد " (Kariha Zaidun Ahmadan; Zaid hates Ahmed) and "كره أحمد زيد " (Kariha ahmadun Zaidan; Ahmed hates Zaid) having a different interpretation, or sentences with different structures, such as "أحبّ زيد هندا" ('Ahabba Zaidun &indan; Zaid loves Hind) and "أكنّ زيد حبّا لهند" ('akanna Zaidun hobban li &indin; Zaid has a love for Hind) having a similar interpretation.

The Arabic simple sentences have the same main components: the predicate (al-mosnad, المسـند), the subject (al-mosnad ‘ilayh, المسـند إليه) that are mandatory in the Arabic sentence, and the complement (al-fodla, الفضلة) to reach the meaning of the sentence. In complex sentences, the predicate, the subject, or the complement can be expanded by adding one or more words, phrases, or clauses to the main clause.

The aim of this work is to merge our previous syntactic analyzer to be able to parse the complex Arabic sentences containing Arabic psychological verbs using NooJ platform [2]. For this reason, we will use the dictionary generated from the lexicon-grammar table of Arabic psychological verbs [3-4], containing all the lexical, syntactic, semantic, and transformational information of these verbs. Then, we will extend our previous analyzer to recognize and denote all the grammatical structures of complex Arabic sentences containing Arabic psychological verbs [5]. Then, we will finish by testing the efficiency of this analyzer on texts and corpora.

References

[1] Amzali, A., Kourtin, A., Mourchid, M., Mouloudi, A. & Mbarki, S. (2021). Syntactic Analysis of Sentences Containing Arabic Psychological Verbs. Bigey, M., Richeton, A., Silberztein, M. & Thomas, I. (Eds.). *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*. NooJ 2021. *Communications in Computer and Information Science*, 1520. Springer, Cham. https://doi.org/10.1007/978-3-030-92861-2_5.

[2] Silberztein, M. (2015). *La formalisation des langues. l’approche de NooJ*. London: ISTE.

[3] Amzali, A., Kourtin, A., Mourchid, M., Mouloudi, A. & Mbarki, S. (2020). Lexicon-Grammar Tables Development for Arabic Psychological Verbs. Fehri, H., Mesfar, S. & Silberztein, M. (Eds.). Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications. *NooJ 2019. Communications in Computer and Information Science*, 1153. Springer, Cham.

[4] Kourtin, A., Amzali, A., Mourchid, M., Mouloudi, A., Mbarki, S. (2020). The Automatic generation of NooJ dictionaries from lexicon-grammar tables. Fehri, H., Mesfar, S. & Silberztein, M. (Eds.). *Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications. NooJ 2019. CCIS*, 1153, 65-76. Springer, Cham.

[5] Bourahma, S., Mourchid M., Mbarki, S. & Mouloudi, A. (2019) Expansive Simple Arabic Sentence Parsing Using NooJ Platform. Mirto, I., Monteleone, M. & Silberztein, M. (Eds.). *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*. *NooJ 2018. Communications in Computer and Information Science*, 987. Springer, Cham. https://doi.org/10.1007/978-3-030-10868-7_10.

Formalization of transformations of complex sentences in Quechua

Maximiliano Duran
Université de Franche-Comté, CRIT, Besançon, France - LIG, UGA, Grenoble, France
duran_maximiliano@yahoo.fr

In the case of some languages, such as English, when a complex sentence consists of a main clause and a subordinate clause, these two clauses are joined together by either, a subordinate 'completive' conjunction (that, so that), a circumstantial conjunction (when, before that, while, because, if, in case of, on condition that, for, even if,) or a relative pronoun (who, that, which,) (e.g. *Because Mom said so, I talked to María*). In Quechua, the subordination is induced by a morphosyntactic marker, applied to the dependent verb that composes the sentence (*e.g. chayta mamai nip**tin**, Mariata rimarqani* [it's because my mother said that, I talked to María]), where the suffix *ptin* agglutinated to the verbal stem *ni*, marks the causative circumstance 'because'.

Among the different classes of dependent clauses in Quechua, there are internally headed relative clauses (IHRC) studied by (R. Hastings 2001), headless relative clauses (HRC) presented by (Peter Cole et al., 1982), and participial relative clauses partially studied by (W.F.H. Adelaar, 2010). In this paper, I complement this last study, by proposing some methods to formalize the clause-subordination strategies of the language.

The verbal suffixes {*-pti, -spa, -stin*} are the ones that allow the construction of adverbial dependent sentences (e.g. *tapu**ptii** chayta willawarqa*/ he told me that, because I asked him; *Kay qellqata tuku**spa**i yanukuita qallarisaq* / I will start cooking after I finish writing; a*siku**stin** sipasqa llamkaq richkan* / the girl goes to work laughing).

These suffixes, which I call, subordination suffixes, can be combined with other verbal suffixes. To precise these combinations, I have built a Boolean matrix of the existing syntactic combinations of (*-pti,-spa, -stin*) with the set of interposition suffixes IPS (*-chka, -yku, -paya*,…). Then, I have selected those agglutinations that preserve subordination such as (*-chkapti, -ykuspa, -payachkapti, -payastin*, …) (e.g. *mikuchkaptii chayaramun*/ he arrived when I was eating; *mikuykuspa llamkaiman rin*/ he went to work after eating).

To formalize all this information, I have built NooJ local elementary transformation grammars, serving to implement automatic larger grammars. These allow us to automatically generate all grammatical transformations of a sentence. Then, I have constructed more specific grammars to extract the transformations that represent paraphrases of the original sentence, more specifically, those containing an adverbial subordinate clause. I also present a set of transformations, serving to implement automatic translation grammars for Quechua-French sentences containing subordinate clauses.

References

Adelaar, Willem, F.H. (2010). Participial clauses in Tarma Quechua. Van Gijn, R., Haude, K., Muysken, P. (Eds.). *Subordination in native South American languages.* University of Zurich Main Library.

Cole P. et als. (1982). Headless relative clauses in Quechua. *International Journal of American Linguistics*. 48(2). University of Chicago Press.

Duran, M. (2013). Formalizing Quechua verbs Inflexion. *Proceedings of the NooJ 2013 International Conference*, Saarbrücken.

Duran, M. (2014). *Les verbes du quechua. Une approche matricielle. Communication Semaine NooJ.* Paris: INALCO.

Duran, M. (2016). The Annotation of Compound Suffixation Structure of Quechua Verbs. *Proceedings of the NooJ 2015 International Conference.* Belarus: Minsk.

Guardia Mayorga, C. (1973). *Gramatica Kechwa.* Lima: Ediciones Los Andes.

Harris, Z. (1951). *Methods in Structutral Linguistics*. Université de Chicago Press.

Hastings, R. (2001). The interpretation of Cuzco Quechua Relative Clauses. *Universite of Massachusetts occasional Papers in Linguistics*. 27.

Parker, G.J. (1969). *Ayacucho Quechua Grammar and Dictionary*. University of Hawaii. Paris: Mouton The Hague.

Rios, A. (2014). https://www.cl.uzh.ch/dam/jcr:ffffffff-d043-9c87-ffff-ffffb112cc62/TR_2014_01.pdf

Silberztein M. (2003). *NooJ Manual*. https://atishs.univ-fcomte.fr/nooj/downloads.html (220 pages, updated regularly).

Silberztein, M. (2016). Formalizing natural languages. *The NooJ approach*. London: ISTE.

Soto Ruiz, C. (1976). *Gramática quechua: Ayacucho-Chanca*. Lima: Ministerio de Educación, Instituto de Estudios Peruanos.

Weber, D. (1983). Relativization and nominalized clauses in Huallaga (Huánuco) Quechua. *University of California Publications in Linguistics.* 103.

Automatic Translation of Arabic Legal Terminology Using NooJ

Khadija Ait ElFqih
UNIOR NLP Research Group - University of Naples 'L'Orientale'
kaitelfqih@unior.it
Maria Pia di Buono
UNIOR NLP Research Group - University of Naples 'L'Orientale'
mpdibuono@unior.it
Johanna Monti
UNIOR NLP Research Group - University of Naples 'L'Orientale'
jmonti@unior.it

Terminology translation plays a critical role in domain-specific machine translation (MT) [4].

Nevertheless, phrase-based statistical MT (PBMT) and neural machine translation (NMT) still present some issues in their results, as proven by several scholars evaluating MT outputs and errors [4, 5, 2]. In fact, current MT outputs do not adhere to the constraints provided by a terminology [2] and this is particularly true for the legal translation context, where MT has typically not been recommended [6].

Furthermore, legal documents may include terms which are characterized by the co-occurrence of phrases defining specific legal aspects of such documents. For instance, in Arabic marriage contracts and divorce provisions the term dowry 'صداق' co-occurs with a phrase that specifies its status, e.g., fully received 'فأبرأته منه فبرئ' .

These types of phrases are context dependent and most MT systems fail in their translation. Example 1 shows the comparison between human translation (HT) and Google Translator (GT).

1. Source: ' قبضت الزوجة جميع الصداق من الزوج قبضا تاما وابرأته منه فبرئ '

AR-EN HT: The wife has fully received the dowry from the husband

AR-EN GT: The wife took all the dowry from the husband and cleared him of him, and he was cured.

In this paper, we propose a set of NooJ grammars [8] to ease the translation of legal terms and their defining phrases from Arabic into English. We choose to work on marriage contracts and divorces provisions, as they pose many challenges due to the terminological, religion-based, culture-specific, system-based asymmetry between Arabic and English [7]. As pipeline, we firstly extract phrases using NooJ Arabic Linguistic Resources (LRs) [3] and concordances based on a list of Legal terms [1] then we proceed with a linguistic analysis of those occurrences and evaluate current MT system outputs, and finally develop a set of Finite State Transducers (FSTs) to propose translations into English.

References

[1] Ait Elfqih, K. (2021).*The impact of the terminological complications on legal translation: From arabic into english with special focus to contracts*. MA Thesis.

[2] Anastasopoulos, A., Besacier, L., Cross, J., Gall´e,M., Koehn, P., Nikoulina, V., et al. (2021). *On the evaluation of machine translation for terminology consistency*. arXiv preprint arXiv:2106.11891.

[3] Bourahma, S., Mourchid, M., Mbarki, S. & Mouloudi, A. (2017). The parsing of simple arabic verbal sentences using NooJ platform. *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ.* (81–95).

[4] Haque, R., Hasanuzzaman, Md. & Way, A. (2019). Investigating terminology translation in statistical and neural machine translation: A case study on English-to-Hindi and Hindi-to-English. *Proceedings of the International Conference on Recent Advances in Natural Language Processing.* RANLP. Varna. INCOMA. (437–446).

[5] Md Mahfuz ibn, A., Anastasopoulos, A., Besacier, L., Cross, J., Gall´e, M., Koehn, P. & Nikoulina, V. (2021). On the evaluation of machine translation for terminology consistency. *Computing Research Repository.*

[6] Killman, J. (2014). Vocabulary accuracy of statistical machine translation in the legal context. *Third Workshop on Post-Editing Technology and Prac-tice.* (85).

[7] S˘ar˘cevi´c., S. (1985). *Translation of culture-bound terms in laws*.

[8] Silberztein, M. (2016). *Formalizing natural languages: The NooJ approach*. John Wiley & Sons.

Corpus Linguistics & Discourse Analysis

Analyzing Political Discourse: Finding the Frames for Guilty and Responsible

Krešimir Šojat
Department of Linguistics,Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
ksojat@ffzg.hr
Kristina Kocijan
Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
krkocijan@ffzg.hr

This paper deals with the analysis of political discourse in Croatia, more precisely, it aims to determine how dissatisfaction is expressed with the attitudes represented by political rivals. We focus on the detection of linguistic means used to show disagreement with decisions or actions taken by parties or individuals considered political and/or ideological opponents. We are particularly interested in the means used by speakers to indicate that someone has failed to do or do something that is under his/her responsibility and is therefore guilty of this omission. In other words, we want to determine how the concept of responsibility is lexicalized, how it is signaled that there is a failure in someone's responsibility and, finally, that someone is therefore to be blamed for that omission or even transgression. For this purpose, we use a large corpus of texts, with over 127 million tokens, consisting of transcripts of plenary debates from the Croatian Parliament that have occurred since the end of 2003 to the end of 2020.

We use NooJ for the construction of a set of rules that aim to detect the usage of the Croatian lexemes *odgovornost* [responsibility] and *krivnja* [guilt] in this corpus. We start from the assumption that it is within the domain of someone's responsibility to conscientiously and properly perform a certain duty or task. Deliberate or accidental omissions in such actions provoke a revolt among speakers and they blame political opponents for these omissions or failures. Since Croatian is very rich in terms of word formation, a subset of rules is designed in order to capture the usage of nouns, verbs, adjectives and adverbs morphologically related to *odgovornost* and *krivnja* as in the following examples:

- As a noun:
  - *On ima **odgovornost** prema ljudima. [He has a responsibility to the people.]*
  - *Oni traže **krivca**. [They are looking for the culprit.]*
- As a verb:
  - *On će **odgovarati** za svoje postupke. [He will be held accountable for his actions.]*
  - *Oni ga **krive** za njegove postupke. [They blame him for his actions.]*
- As an adjective:
  - ***Odgovorna** vlast mora planirati mjere. [The responsible government must plan measures.]*
  - *Za to je **kriv** ministar. [The Minister is to blame for that.]*
- As an adverb:
  - *Oni **odgovorno** vode brigu o imovini. [They take care of the property responsibly.]*
  - *Stranka se ponijela **krivo** prema svojim biračima. [The party has wronged its voters.]*

In data analysis, we take into account the political orientation of speakers, their affiliation with left, right or centrist parties as well as their gender. The preliminary results reveal that, in general, there are more occurrences of *responsibility* than *guilt* in the corpus, but also that the left oriented parties use both terms more than either the right or centrist parties.

References

Kocijan, K. & Šojat, K. (2021). Negation Usage in the Croatian Parliament. Bigey, M., Richeton, A., Silberztein, M. & Thomas, I. (Eds). *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*, 101-113. Switzerland: Springer, Cham. doi:10.1007/978-3-030-92861-2_9.

Silberztein, M. (2016). Formalizing Natural Languages: The NooJ Approach. *Cognitive science series.* London: Wiley-ISTE.

Van Dijk, T. A. (1997). What is political discourse analysis? Bloomaert, J. & Bulcaen, Ch. (Eds.). *Political Linguistics.* 11-52. Amsterdam.

Creation of a legal domain corpus for the Belarusian NooJ module: texts, dictionaries, grammars

Valerii Varanovich
Belarusian State University, Kurchatova str. 5, Minsk, Belarus
Mikita Suprunchuk
Minsk State Linguistic University, Zakharov str. 21, Minsk, Belarus
Yauheniya Zianouka
Tsimafei Prakapenka
United Institute of Informatics Problems, Surhanava str. 6, Minsk, Belarus
Anna Dolgova
Minsk State Linguistic University, Zakharov str. 21, Minsk, Belarus
Yuras Hetsevich
United Institute of Informatics Problems, Surhanava str. 6, Minsk, Belarus
ssrlab221@gmail.com

The current language situation in the Republic of Belarus is characterized primarily as state bilingualism. At the legislative level, two state languages are established – Belarusian and Russian. But despite the state bilingualism, the vast majority of legislative documents are implemented only in Russian. Thus, of the 26 codes of the Republic of Belarus, i.e. texts which are available on the National Legal Internet Portal pravo.by, 25 are officially adopted in Russian and only one in Belarusian. One of the main factors hindering the practical support of bilingualism in the legal sphere of the Republic of Belarus is the unresolved problem of high-quality and fast linguistic processing of large texts, which testifies to the relevance of high-quality machine translation. To handle the question of translating legislative documents into Belarusian, Speech Synthesis and Recognition Laboratory of UIIP NASB, in cooperation with specialists from Faculty of Social and Cultural Communications of BSU, have translated all codes of the Republic of Belarus into the Belarusian language using automatic services of corpus.by.

The next step is to collect all legislative codes of the Republic of Belarus in the Belarusian language in order to create a unified text corpus. This task is relevant for solving the following tasks. First of all, it is very important to perform primary analysis of legal domain corpus to find out the main linguistic peculiarities of this kind of corpus in comparison with Belarusian literary corpus. Secondly, we will be able to compose different types of dictionaries (Belarusian-Russian, Belarusian-English, Belarusian-English-Russian dictionaries). This question is very actual for Belarusian since there are very few translated Belarusian-foreign and foreign-Belarusian dictionaries of legal terminology – we are aware of 6 dictionaries that are different in their merits and significance for the ordering and development of Belarusian legal terminology. And the last – but not least – task is to develop special morphological and syntactical grammars for further prosodic analysis of legal texts. Automated syntagmatic delimitation is still not solved for the Belarusian language. That is why developing NooJ grammars will assist the process of creating a system of prosodic marks (including punctuation and intonation marks) and further automatic segmentation of Belarusian texts of the legal domain.

References

Budějovice (2017), Czech Republic. Barone, L., Monteleone, M. & Silberztein, M. (Eds.). *Revised Selected Papers.,* June 9-11 (101-111).

Computational platform for electronic text and speech processing corpus.by (2019). [Electronic source]. https://corpus.by/. Last accessed July 12 2018.

Drahun A. (2019). Semi-Automatic Proofreading of Belarusian and English texts. Drahun, A., Hetsevich, Y., Bakunovich, A., Dzenisiuk, D. & Shynkevich, J. *International Conference NooJ 2019: Book of Abstracts*. Tunisia: Hammamet.

Hetsevich Y. (2016). Semi-automatic Part-of-Speech Annotating for Belarusian Dictionaries Enrichment in NooJ. Hetsevich, Y., Varanovich, V., Kachan, E., Reentovich, I. & Lysy, S. *Automatic Processing of Natural-Language Electronic Texts with NooJ: 10th International Conference*, České.

Hetsevich, M., Silberztein, H. *Stanislavenka, Springer International Publishing* (3-15).

National Legal Internet Portal of the Republic of Belarus (2019). List of legal acts [Electronic source]. https://pravo.by/document/?guid=3871&p0=H11800130. Last accessed July 18 2019.

Reentovich I. (2016). The First One-Million Corpus for the Belarusian NooJ Module. Reentovich, I., Hetsevich, Y., Voronovich, V., Kachan, E., Kozlovskaya, H., Tretyak, A. & Koshchanka, U. *Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015*, Minsk, Belarus, June 11-13. Okrut, Y. *Revised Selected Papers.*

Prosodic segmentation of Belarusian texts in NooJ

Yauheniya Zianouka
United Institute of Informatics Problems, Minsk, Belarus
evgeniakacan@gmail.com
David Latyshevich
United Institute of Informatics Problems, Minsk, Belarus
david.latyshevich@gmail.com
Yuras Hetsevich
United Institute of Informatics Problems, Minsk, Belarus
yuras.hetsevich@gmail.com
Mikita Suprunchuk
United Institute of Informatics Problems, Minsk, Belarus
ms@philology.by

The paper represents some syntactic grammars and rules to Belarusian prosodic delimitation. To date, there are no general rules or mechanisms for an unambiguous definition of syntagmas in a written text or speech flow. The study of the prosodic speech organization is conducted on the basis of auditory and experimental analyses, with the help of which the parameters of supersegmental means are distinguished. They are the limits of the speech flow segmentation, types of intonation constructions (IC), tonal, dynamic and quantitative signals of the IC center, changes in the speed and intensity of sound.

This work is a continuation of previous research where analyzed syntagmas were separated by punctuation [1-3]. Now the authors have expanded the study. We applied a technique for automated phrase segmentation not only at the punctuational level but also at the semantic. The keystone is the number of syntagmas in a sentence that can significantly exceed the number of punctuation marks in the text. The main core of the research is the morphological and syntactic principle. The approach is confined to the ability of a particular speech part to match with words of other lexical and grammatical classes and occupy a certain syntactic position. The concept is grounded in a superficial syntactic analysis of a text with an emphasis on grammatical characteristics of speech parts that combine accentual units.

Developed grammars will be used for further research in phrase delimitation of Belarusian. Identified prosodic rules for dividing speech flow at punctuation and syntactic levels estimate the value of intonation peculiarities of a certain language. It will be also wholesome to create an algorithm for segmenting textual information in the Belarusian speech synthesis systems and may also serve to improve the Belarusian NooJ module with so-called prosodic transcription at different levels.

References

[1] Dzenisiuk, D. (2019). Automatic Generation of Right Intonational Marks and Speech for Medical domain in Belarusian. Dzenisiuk, D., Hetsevich, Y., Drahun, A., Bakunovich, A. & Shynkevich, J. *International Conference NooJ 2019: Book of Abstracts*. Hammamet, Tunisia.

[2] Hetsevich, Y. (2016). Grammars for Sentence into Phrase Segmentation: Punctuation Level. Hetsevich, Y., Okrut, T. & Lobanov, B. *Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015.* Minsk, Belarus, June 11-13, 2015. Okrut, T., Hetsevich, Y., Silberztein, M. & Stanislavenka, H. (Eds.). *Revised Selected Papers* (74-82). Springer International Publishing.

[3] Zianouka, Y. (2022). Automatic Generation of Intonation Marks and Prosodic Segmentation for Belarusian NooJ Module. Zianouka, Y., Hetsevich, Y., Latyshevich, D. & Dzenisiuk, Z. *15th International Conference, NooJ 2021.* (231-242). Besançon: Springer, Cham.

A linguistic approach for automatic analysis, recognition and translation of Arabic predicative nouns

Cheikhrouhou Hajer
University of Sfax, LLTA, Tunisia
cheihkkrouhou.hager@gmail.com
Imed Lahyani
University of Sfax, LLTA, Tunisia
lahiani.imed@gmail.com

Our project consists in the design and assembly of a digital dictionary of Arabic predicative nouns, of the type "…ضربات, أخطاء, أمراض, جرائم, مساعدات, نصائح ".

In this project, we tried to select the Arabic support verbs which make the actualization of the Arabic predicative nouns, relying on the principles of Gaston Gross's object class theory. In fact, according to this theory, the verbs are classified into three basic classes: the predicative verbs " les verbes prédicatifs:أفعال إسـنادية ", the support verbs " les verbes supports:أفعال ناقلة / عماد ", the frozen verbs " أفعـال متكلسة : les verbes figés " without forgetting the auxiliaries " les auxiliaires:الأفعال المساعدة ". The same goes for nouns, where we find the predicative nouns " أسماء إسناديّة : les noms prédicatifs", the non-predicative nouns "les noms non-prédicatifs: أسماء غير إسنادية " and the fixed arguments " معمـولات متكلّسة : les arguments figés".

In this study, we are interested in the study of Arabic predicative nouns which select their own arguments and their actualizer which is the support verb Vsupp+NPRED. Ouerhani Bechir defines the Arabic support verb as: "*The Support Verb is a non-predicative verb which is associated with the predicate which it accompanies and whose actualization it ensures. As a result, the selection of the arguments is ensured by the predicate. On these arguments, the supporting verb only imposes constraints relating to the arrangement of the elements of the sentence such as their order and their case endings, in addition to gender and number agreements*"[1].

At the semantic level, support verbs carry an aspectual or modal value. In this framework we find:

¬ **The basic support verbs: أفعال ناقلة عامّة : Les verbes supports de base**

¬ **Appropriate supporting verbs: أفعـال ناقلة مخصّصة : Les verbes supports appropriés**

¬ **Aspectual support verbs: أفعال ناقلة مظهريّة : Les verbes supports aspectuels**

- Inchoative aspect: شروعيّ : Aspect inchoatif
- Progressive/continuative aspect: تطوّريّ : Aspect progressif / continuatif
- Iterative aspect: تكراريّ : Aspect itératif
- Final aspect: انتهائيّ : Aspect terminatif

¬ **Metaphorical supporting verbs: ناقل مجازي : Les verbes supports métaphoriques**

We note that the majority of aspectual support verbs are translated into French by verbal periphrases with aspectual values.

For the realization of this project, we have started the first step which consists in studying all the classes of predicative nouns proposed and fixing for each class the verbs' appropriate supports and delimit the appropriate arguments for each semantic-syntactic construction. We recall that the theory of object classes is based on the analysis of argument schemes and on the principle of the elementary sentence, i.e. the predicate and its arguments.

We have grouped, so far, thirty classes of predicative nouns such as the example of:

Classe of predicative nouns <help>: صنف - < مساعدات > : مساعدة، دعم، إعانة، مساندة.

The basic support verbs did (fait): أفعال ناقلة عامّة: قام

]ف ن | 0]بشر[حرف | 1 إس | 2]بشر

[Hum]2ا N1 Prep [Hum]0 N

قام سامي ب)مساعدة + دعم + إعانة + مسـاندة(سـلمى Sami (helped/supported) Salma. (Sami a aidé/soutenu Salma).

1 Ouerhani, B, *Les verbes supports dans les dictionnaires arabes en lignes: étude d'échantillons.* https://www.academia.edu/32040441/Ouerhani_Cahiers_Dict_Vsup?email_work_card=view-paper

In the second step, we tried to integrate these linguistic and theoretical data into an electronic dictionary using the NooJ platform for the creation of an Arabic-French machine-translation (MT) application.

For the realization of this TA application, we have built:

1/ A bilingual electronic dictionary of support verbs for each class of predicative nouns “Verbe supportar”.

2/ A bilingual electronic dictionary of the predicative nouns of each class.

3/ Grammars for the automatic analysis and recognition of predicative nouns.

EXP: إِسْـــــــــــــــــــــــتَمَ ر ,V+Supp+CONS=V+N0Hum+PREP+NPRED+N1Hum+FLX=V_estakhafa10a+DRV=N_estakhafa10a:FlxDRV+FR= “continue”

References

Cheikhrouhou H. (2016). Arabic Translation of the French Auxiliary: Using the Platform NooJ. Barone L., Monteleone M. & Silberztein M. (Eds). *Automatic Processing of Natural-Language Electronic Texts with NooJ. NooJ 2016. Communications in Computer and Information Science.* 667 (74-86). https://link.springer.com/chapter/10.1007/978-3-319-55002-2_7

Cheikhrouhou, H. (2017). The Automatic Translation of French Verbal Tenses to Arabic Using the Platform NooJ. Mbarki S., Mourchid M. & Silberztein M. (Eds). *Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications. NooJ 2017. Communications in Computer and Information Science.* 811 (156-167).

Gross, G. (2008). Les classes d’objets. *Lalies, Presses de l'ENS, Editions rue d'Ulm.* 111-165. https://halshs.archives-ouvertes.fr/halshs-00410784

Lahyani, I. (2014). *Les verbes transitifs par une préposition à un seul complément*, Thèse de Doctorat, FLSHS, Sfax, Tunisie. 5. Ouerhani, B, Les verbes supports dans les dictionnaires arabes en lignes: étude d'échantillons. https://www.academia.edu/32040441/Ouerhani_Cahiers_Dict_Vsup?email_work_card=view-paper

Silberztein, M. (2015). La formalisation des langues l’approche de NooJ. *Collection Science Cognitive et Management Des Connaissances*. ISTE.

Creation of parallel medical and social corpora for the machine translation and speech synthesis

Mikita Suprunchuk
Minsk State Linguistic University, Zakharov str. 21, Minsk, Belarus
ms@philology.by
Nastassia Yarash
United Institute of Informatics Problems, Minsk, Belarus
anyarosh62@gmail.com
Yuras Hetsevich
United Institute of Informatics Problems, Surhanava str. 6, Minsk, Belarus
yuras.hetsevich@gmail.com
Valerii Varanovich
Belarusian State University, Kurchatova str. 5, Minsk, Belarus
gamrat.vvv@gmail.com
Siarhey Gaidurau
United Institute of Informatics Problems, Minsk, Belarus
Yauheniya Zianouka
United Institute of Informatics Problems, Surhanava str. 6, Minsk, Belarus
evgeniakacan@gmail.com
Palina Sakava
United Institute of Informatics Problems, Minsk, Belarus
polina.sakova.work@gmail.com

In the Republic of Belarus both Belarusian and Russian have the status of official language. Since there is bilingualism at the legislative level, it is necessary to provide the availability of texts for different purposes in both official languages. One of the important spheres is public education in medical and social fields.

Medical and social texts contain essential information. The first are intended to convey information between health professionals and scientists (for example, in articles, research papers) or to inform the population about public and personal health. The second group represents the historical and cultural heritage of Belarus and is aimed at acquainting country visitors with it.

Therefore, the need for a system of machine translation and speech synthesis in Belarusian and Russian is still relevant. There is a data layer of mentioned domains poorly covered with such systems, which, in turn, require large parallel corpora for their training. There are no large Belarusian-Russian parallel corpora on open access. So, the first step in the machine translation and speech synthesis development may be the creation of corresponding corpora. Using proofreading and word processing services and NooJ processor [1; 2], it is possible to work with big data to further use it for training machine systems.

Currently, we work at creating a corpora of medical and social domains. A medical corpus of 848 parallel texts in Belarusian (303,469 word forms), English (373,709 word forms) and Russian (330,996 word forms) has been compiled. A trilingual parallel corpus of social domain (based on parallel texts in Belarusian, English and Russian from the “KrokApp” travel audio guide [3]) is being created. Development of specialized syntactic grammars for the Belarusian NooJ module, which will allow the analysis of set phrases specific for medical and social texts, is planned.

References

[1] Hetsevich, Y. [et al.] (2021). Creation of a legal domain corpus for the Belarusian NooJ module: texts, dictionaries, grammars. Bigey, M. [et al.] (Eds.). *15th International Conference NooJ 2021: book of abstracts.* Besançon. (36-37).
[2] Drahun, A. (2019). Semi-Automatic Proofreading of Belarusian and English texts. Drahun, A. [et al.]. *International Conference NooJ 2019: Book of abstracts*. Hammamet, Tunisia.
[3] KrokApp – personal audio guide in Belarus. https://krokapp.by/about/.

Processing the discourse of insecurity in Rosario with the NooJ platform

Andrea Rodrigo
CETEHIPL, Facultad de Humanidades y Artes, UNR, Argentina
andreafrodrigo@yahoo.com.ar
Silvia Reyes
CETEHIPL, Facultad de Humanidades y Artes, UNR, Argentina
sisureyes@gmail.com
Mariana González
CETEHIPL, IES N° 28 “Olga Cossettini”, Rosario, Argentina
marianagonzalez826@gmail.com

The CETEHIPL (Centro de Estudios de Tecnología Educativa y Herramientas informáticas de Procesamiento del Lenguaje) has been working with the pedagogical application of computer tools to language teaching. Today we will take a small turn towards discourse analysis and choose to analyze a partly exacerbated and recurring topic in post-pandemic Argentina: insecurity. Here we intended to record what impact insecurity had and still has on the linguistic domain. Consequently, we cannot ignore the mediatization of information and how journalism discourses on insecurity. We built a corpus of journalistic texts published in December 2021 in the main newspapers of Rosario: *La Capital*, *Rosario12* and *El Ciudadano*. However, we will not discuss the very complex phenomenon of insecurity, since we consider that the media are exposed to a specific logic, as it is well stated by Guemureman et al. (2010):

The structure the media system takes in a country is partly determined by a correlation of forces between various political actors. A tour around the “media map” in Argentina today allows us to affirm that the media system in our country pursues a commercial logic, not a public service one.

The concept of news is influenced by the construction of a piece of merchandise generating profits and participating in consumption. This always occurs within the process of dismantling a protective state, where the notion of punishment is often erased in order to focus on "the victim":

The news cycle as merchandise reproduces the merchandising cycle of any product, the news- merchandise is consumed, its consumption fosters demand, demand fosters production, circulation and consumption, all merchandise generates demand and a consumer market.

From our role as researchers and taking a NLP perspective, we can provide an interesting analysis of insecurity and its linguistic impact. We tackle this issue with the Rioplatense Spanish resources developed by the IES_UNR team with NooJ. Neologisms such as "gatillero" (shooter) or "balacera" (shooting), terms such as "toy" (meaning “weapon"), or even words from Rioplatense Spanish slang (*lunfardo*) such as "luca" (a thousand, referring to money) or "gilada" (gullible people) are introduced in our dictionary. To account for this discourse, we created tags in order to distinguish three different types of expressions referring to the victim, to the role of the state and to the perpetrator. In this first approach, we can observe that victim related expressions occupy a predominant place, while state related expressions are very scarce. This comes into line with what is called the weakening of the role of the state as an immediate consequence of the “media construction of insecurity”. Perpetrator related expressions are not preponderant either, since stress is always placed on the victim.

To complete our analysis, we developed grammars to show how the impact of insecurity is made visible from a syntactic viewpoint. As it is our first approach to this phenomenon, we can only offer partial conclusions. For this reason, our research will go on by enlarging the sample in order to progressively achieve a greater coverage of insecurity expressions with the resources of the Spanish Module Argentina designed by our team on the NooJ platform

References

Guemureman, S., Fridman, D., Graziano, F., Jorolinsky, K., López, A., Pasin, J., Salgado, V. (2010). Rol de los medios de comunicación en el despliegue de los mecanismos de control social, proactivos y reactivos. Legitimación de la violencia estatal contra los jóvenes pobres y su vinculación discursiva con la "delincuencia" [online]. *VI Jornadas de Sociología de la UNLP.* Ensenada, Argentina. Memoria Académica. https://memoria.fahce.unlp.edu.ar/trab_eventos/ev.5699/ev.5699.pdf

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. London: ISTE.

Spanish Module_Argentina: https://atishs.univ-fcomte.fr/nooj/resources.html

IMF´s Impact on Argentina: using NooJ to automatically analyze the print media literature on Argentina´s debt affairs

Carmen González
Facultad de Ciencias Económicas y Estadística, UNR, Argentina
caar.gonzalez26@gmail.com
Andrea Fernanda Rodrigo
CETEHIPL, Facultad de Humanidades y Artes, UNR, Argentina
andreafrodrigo@yahoo.com.ar
Mariana González
CETEHIPL, IES N° 28 "Olga Cossettini", Argentina
marianagonzalez826@gmail.com

This work seeks to examine the automatic processing of a corpus of newspaper articles within the working frame of the Centro de Estudios de Tecnología Educativa y Herramientas Informáticas de Procesamiento del Lenguaje (CETEHIPL). The articles were selected from three Argentine newspapers with a prominent focus on economic, financial and business matters: El Cronista, Ámbito Financiero and El Economista. These articles are written in Rioplatense Spanish and were published the following days after Argentina's 2021 midterm elections. Our goal is to fully analyze and understand where it is placed their focus of attention, and how the post pandemic effects on the Argentine economy are reflected at the discourse level. We are acquainted with the fact that our corpus is certainly intersected by a media-related logic. According to Hjarvard (2016), media play a relevant role in transforming culture and society[1]. In effect, the so-called "mediatization" becomes a global phenomenon in such a way that:

Mediatization should be viewed as a modernization process on a par with urbanization, globalization and individualization[2]

However, we do not intend to go into the aspects that involve the mediatization of information nor the role played by the media in shaping the public opinion. Instead, our attention is placed on how the discourse is used and which linguistic impact these articles have on the audience. In this way, we used NooJ, a linguistic development environment constructed by Max Silberztein (2016). We worked with the dictionaries and grammars created on the NooJ platform by our IES-UNR team to automatically process our corpus. We set out two semantic fields of study: one concerning to the COVID- 19 pandemic and the other one concerning to social and economic issues. We observed that topics related to Argentina's debt affair with the IMF ranked first whereas other topics such as the COVID-19 pandemic, poverty and devaluation were relegated to a back seat. In this way, we chose two sample statements to show how this hot topic is syntactically expressed:

1. *El FMI sostuvo a Macri, benefició a inversores y endeudó a la Argentina (*The IMF upheld Mauricio Macri´s administration benefited investors and drove Argentina into debt)*,*
2. *El FMI reconoció que la plata se utilizó para pagar deuda insostenible a acreedores privados* (The IMF admitted the loan was used to pay an unsustainable debt to private creditors)*.*

We then used the dictionaries and grammars created with NooJ platform to analyze them. The resulting output enabled us to extend and enrich the resources in the Spanish Module_Argentina. We are planning to add other Argentine newspapers of record such as La Nación and Clarín to our corpus. Although they do not specialize in economic affairs, they have a huge impact on public opinion and will contribute towards further strengthening the conclusions we reach here.

References

Hjarvard, S. (2021). Mediatización: La lógica mediática de las dinámicas cambiantes de la interacción social. *La Trama de la Comunicación.* UNR, 20(1), 235-252. Available at: https://www.redalyc.org/pdf/3239/323944778013.pdf Last accessed January 20 2021
Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. London: ISTE.
Ámbito financiero. https://www.ambito.com/ Last accessed January 20 2021.
El cronista. https://www.cronista.com/ Last accessed January 20 2021.
El economista. https://eleconomista.com.ar/ Last accessed January 20 2021.
Spanish Module_Argentina. https://atishs.univ-fcomte.fr/nooj/resources.html

1 Hjarvard (2016). p. 238.
2 Hjarvard (2016). p. 239

Natural Language Processing Applications

The digital text workshop cloud, new solutions for super calculation environments

Ilaria Veronesi[1], Rita Bucciarelli[2], Francesco SaverioTortoriello[3], Andrea Rodrigo[4], Marianna Greco[5], Colomba La Ragione[6], Javier Julian Enriquez[7]

[1] University of Salerno, Italy, [2] University of Siena, Italy, [3] University of Salerno, Italy, [4] University of Rosario, Argentina, [5] Ministry of Education, Italy, [6] University Pegaso of Naples, Italy, [7] Polytechnic University of Valencia, Spain

Quantum physics is the basis of scientific thinking and influences the mind. Wendt (2015) in *Quantum Mind and Social Science* argues that the mind and social life are macroscopic phenomena quantum mechanics and "...*that the quantum consciousness hypothesis" (QCH)* is the cognitive basis of a quantum social science[1]. In these new scientific parameters of knowledge, our visualization matrix is essentially a holographic projection. Research innovation lies in seeking the point of convergence to describe a formal process that, as a multi-code process, goes from a sentence assembly phase to a data implementation phase.

The work is included in the *Quantum computing* project, which includes SSD: Area 01 - Mathematical and Computer Sciences; Area 10 - Ancient, Philological-Literary and Historical-Artistic Sciences. The research is supported by the scientific collaboration with experts in quantum computing such as M. Planat of the Institute of Femto -ST Dep. of Micro Nano Sciences and Systems (MN2S), Besançon, France. The goal is to find points of contact between quantum physics, computational linguistics, and quantum computing. M. Planat affirms that the quality that the linguist must possess beyond courage is ... *compétence de transcribe les concepts mathématiques dans les structures grammaticales.*[2] This assumption is at the basis of our study. Planat, in Quantum Gravity Research, Los Angeles, and in Raymond Aschheim, and Marcelo M. Amaral, Fang and Klee Irwin in Graph Coverings for Investigating *"Non-Local Structures in Proteins, Music and Poems"* states that: "*a remarkable analogy between the pattern structure of bonds between amino acids in a protein (the protein secondary structure has been pointed out firstly and that non-local structures are observed in tonal music and in poems. The origin of these analogies has been explained with finitely generated groups and graph covering theory'*.[3] This study focuses on the narrative text of Charles Baudelaire in seq, 5-13, in Les Petits Poèmes en prose in Spleen de Paris, 1869,[4] to describe the point of contact that emerges from the proposed assumptions, that perhaps *mathematical linguistics is the proper frame for making progress and artificial intelligence (AI) may help in the classification of languages*.[5] The focus and point of convergence that unites mathematics to linguistics emerges, in the form of a narrative sequence described in a mathematical theory in which numbers and letters and terms are inserted in a different grammar, with a transformation into a synthetic code. The tools are 1) To identify new parameters for producing fixed structures in equations, recursiveness, Fibonacci sequence; 2) NooJ system by Max Silberztein (2015),[6] for producing analysis and paraphrase of sentences, tools to develop formal dictionaries and grammar, and NLP applications such as semantic annotators; 3) The implementation of advanced integrated Machine Learning algorithms for knowledge creation from obtained data; 4) BuViTeMS (© 2020) Digital Intelligence AW model for producing sentences drafted, reformulated, and translated using a blended system. Hence, the whole formal process will make use of: 1 Lexicon-grammar to transfer the numerical quantum language into a lexicon-grammar code; 2 NooJ's environment to produce statistical analysis.

References

Baudelaire, C. (2019). *Le Spleen de Paris (Petits poèmes en prose): Un recueil posthume de poésies de Charles Baudelaire*. BoD-Books on Demand.

ChromaticScale. Available: https://en.wikipedia.org/wiki/Chromatic_scale Last accessed April 1 2021.

Planat, M., Aschheim, R., Amaral, M.M., Fang, F. & Irwin, K. (2020). *Complete quantum information in the DNA genetic code. Symmetry. 12-1993.*

Planat, M., Aschheim, R., Amaral, M.M., Fang, F. & Irwin, K. (2020) *Quantum information in the protein codes, 3-manifolds and the Kummer surface*. *Symmetry* . 13-1146.

Planat, M., Aschheim, R., Amaral, M.M., Fang, F. & Irwin, K. (2021). Coperture di grafici per lo studio di strutture non locali in proteine, musica e poesie. *Sci.* 3(4), 39.

The Protein Data Bank. https://pdb101.rcsb.org/ Last accessed January 1 2021.

1 Wendt , A. (2015). Quantum mind and social science. Cambridge University Press.

2 Planat, M., Aschheim, R., Amaral, M.M., Fang, F. & Irwin, K. (2020). *Quantum information in the protein codes, 3-manifolds and the Kummer surface*. Symmetry. 13-1146. [CrossRef]

3 Planat, M., Aschheim, R., Amaral, M.M., Fang, F. & Irwin, K. (2021). Graph Coverings for Investigating Non Local Structures in Proteins, Music and Poems. *Ski,* 3(4), 39.

4 http://poetes.com/textes/baud_spl.pdf

5 See *Ibid.,* Planat, et al (2021). Graph Coverings..., p. 2.

6 Clifford group dipoles and the enactment of Weyl / Coxeter group W (E8) by entangling gates Michel Plana.

Construction of an educational game “CONJ_NooJ”

Hela Fehri
MIRACL Laboratory University of Sfax
hela.fehri@yahoo.fr
Nizar Jarray
ISG Gabes University of Gabes
nizar.jarray1998@gmail.com

Traditional teaching and learning methods are not always effective and adapted to all children. That is why, some users resort to use the game as a learning support. In fact, Game-based education [1] allows to the children to stay motivated and more engaged, to verbalize their thoughts, to argue their choice and to perfect their language and learn from their mistakes. However, the development of an educational game is not a trivial task because the game should be simple and straight. Moreover, the interface must be intuitive so the child does not try to understand how the game works.

The aim of this paper is to propose a game developed with NooJ platform [2]. This game allows how to master Arabic, English and French Verb conjugation. It allows also how to master French Nouns and adjectives inflection. There are two essential steps for each language in this game: review and evaluation. These two steps are independent but complementary. In fact, the first step can help the gamer succeed in the second step.

The first step, labeled "Review," is to recognize the conjugation of appropriate verb in the chosen language or the inflection of appropriate noun or adjective. This step is named "Review" because it can help the gamer remember the conjugation of a few verbs or the infection of some nouns and adjectives before beginning the "Evaluation" part. This can help succeed in the second step. The second step, labeled "Evaluation", is composed of three exercises the difference being the level of difficulty. A user can only move to the last level when he succeeds the second level.

Concerning the conjugation of the verbs, the first level consists in learning how to conjugate a verb in a specific time. The second level consists in learning how to conjugate a verb in different tenses. The third level consists in filling the crossword puzzle with the appropriate conjugated forms of a set of verbs given randomly.

Regarding the nouns and adjectives, the first level consists in learning the agreement of the appropriate word. The second level consists in recognizing the gender of the noun and how to transform an adjective to an adverb. The third level consists in in filling the crossword puzzle with the appropriate noun or adjective.

The implementation of this game is based on _dm dictionary for the French language, the Verbes Arabes dictionary for the Arabic language and the phrasal verb and _Contractions dictionaries for the English language. It is based also on the transducers that allow the conjugation of the appropriate verb or the infection of the noun or adjective. This game is easy to play and does not require computer skills using Java interface. The obtained results are encouraging and satisfactory.

References

[1] Fehri, H. & Ben Messaoud, I. (2020). Construction of Educational Games with NooJ. Fehri, H. [et al.] (Eds.). *Hammamet Tunisia*. 20, NooJ 2019, CCIS 1153, 173–184.
[2] Fehri, H. et al. (2021). Construction of an Educational Game “VocabNooJ”. Bigey, M. [et al.] (Eds.). *NooJ 2021. (*124–134). Besançon.
[3] Silberztein, M. (2015). La formalisation des langues: l’approche NooJ. *Collection Sciences Cognitive et Management des Connaissances*. ISTE.

From questions in natural language to SPARQL queries

Ismahane Kourtin
ELLIADD Laboratory, Bourgogne-Franche-Comté University, Besançon, France
MISC Laboratory, Faculty of Science, Ibn Tofail University, Kenitra-Morocco
kourtin_ismahane.math@yahoo.fr

The mass of information in the legal field, which is constantly increasing, has generated a capital need to organize and structure the content of the available documents, and thus transform them into an intelligent guide capable of providing complete and immediate answers to queries in natural language, and promoting the development of new forms of collective intelligence. Therefore, the question-answering system (QAS) [1], which is an application of the automatic language processing domain (NLP), perfectly meets this need by offering different mechanisms to provide adequate and precise answers to questions expressed in natural language. The general context of our work is the construction of a Question-Answering System in the legal field based on ontologies [2] [3], allowing users to ask a question on the desired information using natural language without having to browse through the documents. In this article, we will focus on the process that consists of reformulating the question by a SPARQL query(s) with which we can query the ontology and thus retrieve the appropriate answer to the question asked by the user. We have adopted a methodological framework in two steps:

1- Question analysis. The text file containing the question, and the input and answer languages, is transmitted to NooJ platform [4][5] with noojapply, which allows, from a constructed grammar according to the input language, to extract the RDF triplet(s) components of the question. The result of the extraction is retrieved as an « .ind » file.

2- SPARQL queries generation. After having retrieved the file containing the RDF triplet(s) components of the question, we extract these components from the file, and then we generate the appropriate SPARQL query specifying the language of the requested data depending on the response language chosen by the user.

References

[1] Hirschman, L. & Gaizauskas, R. (2001). Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300. http://dl.acm.org/citation.cfm?id=973891

[2] Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition.* 5(2):199–220. http://secs.ceas.uc.edu/~mazlack/ECE.716.Sp2011/Semantic.Web.Ontology.Papers/Gruber.93a.pdf

[3] Willem Nico Borst. (1997). *Construction of engineering ontologies for knowledge sharing and reuse*. Universiteit Twente. http://doc.utwente.nl/17864

[4] Silberztein, M. (2003). *NooJ Manual*. Available at: https://atishs.univ-fcomte.fr/nooj/downloads.html

[5] Silberztein, M. (2016). *Formalizing Natural Languages: the NooJ approach*. Wiley-ISTE. Hoboken.

Integrated NooJ environment for Arabic Linguistic Disambiguation using MWEs

Dhekra Najar
RIADI, University of Manouba, Tunisia
Dhekra.najar@gmail.com
Slim Mesfar
RIADI, University of Manouba, Tunisia
mesfarslim@yahoo.fr

Language resources are a necessary component to language Development in NLP. They are useful for any empirical language study including linguistic analysis, language translation and language disambiguation.

The linguistic development environment NooJ allows formalizing complex linguistic phenomena such as compound words generation, processing as well as analysis. NooJ offers the possibility to use the dynamic library NoojEngine.dll or the command-line program: noojapply.exe. In this study, we will take advantage of noojapply.exe program that is freely available in the Standard edition of NooJ. Noojapply.exe allows users to apply dictionaries and grammars automatically to texts from external environments.

In this paper, we introduce a module for Arabic MWEs recognition that is based on rules grammar. MWEs module allows recognizing several types of morphosyntactic variations that can occur to a MWE. Then, we study the impact of multi-word expressions recognition on Word Disambiguation in Arabic language texts. These linguistic resources are compiled to be used as parameters in the command-line noojapply.exe in order to be integrated within an Arabic language processing environment for linguistic disambiguation.

Our work is divided into three sections. First, we deal with a literature review on disambiguation task in the Arabic language. Then, we give a detailed description of our Integrated NooJ environment for Arabic linguistic disambiguation and the associated grammars. Finally, a set of tests and experiments is carried out to measure the impact of multi-word expressions recognition in Word Disambiguation.

References

1. Ditters, E. (2001). A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic. *In the proceeding of Arabic NLP Workshop at ACL/EACL*.
2. El Jihad, A. & Yousfi, A. (2005). Etiquetage morpho-syntaxique des textes arabes par modèle de Markov caché. *Proceedings of Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues.* (649-654).
3. Silberztein, M. (2015). *La formalisation des langues: l'approche NooJ*. ISTE.
4. Mesfar, S. (2008). *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard.* Doctoral dissertation, Université de Franche-Comté. UFR des Sciences du langage, de l'homme et de la société.
5. Silberztein, M. (2005). “NooJ's Dictionaries”. *In the Proceedings of the 2nd Language and Technology Conference.* Poznan.
6. Najar, D., Mesfar, S. & Ghezela, H. B. (2015, June). A Large Terminological Dictionary of Arabic Compound Words. *International NooJ Conference* (16-28). Springer International Publishing.
7. Najar, D. & Mesfar, S. (2017). Opinion mining and sentiment analysis for Arabic on-line texts: application on the political domain. *International Journal of Speech Technology.* 20(3), 575-585.

The Use of NooJ Platform to Build a Formative Assessment for Arabic language

Ilham Blanchete
Faculty of Science, Sofware Engineering department, ibn Tofail University, Kenitra, Moroco
ilham.blanchete@gmail.com
Mohammed Mourchid
Faculty of Science, Sofware Engineering department, ibn Tofail University, Kenitra, Moroco
mourchidm@hotmail.com

This paper discusses the use of Arabic linguistic resources to build a formative assessment for the Arabic language lessons. The formative assessment helps teachers to conduct in-process evaluations of student comprehension. The formative assessment will be presented as an application that:

- Helps learners to understand and construct different verbal phrases.
- To paraphrase verbal sentences. Grammar (1) allows changing a verb with a "Masdar Moawal" المصدر المؤول.

E.g., (I input: يجب الذهاب الألن- go to have now). (I input: يجب أن أذهب الألن- go to have now).
The verb "to go أذهب " has been changed to the masdar form الذهاب.
For this purpose, we have used:

1. NooJ platform implements the following linguistic resources:
   - NooJ's morphological grammars represent the extracted transformational rules (in-process).
   - NooJ's syntactical grammar represents the simple verbal phrases (in progress).
2. Python programming language to develop the formative assessment as an NLP application. The formative assessment detects the learner's mistakes and provides a linguistic description as a solution.

References

Amzali, A., Mourchid, M., Mouloudi, A. & Mbarki, S. (2020). Arabic psychological verbs recognition through NooJ transformational grammars. *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities,. NooJ 2020. Communications in Computer and Information Science.* 1389. Bekavac, Springer, Cham. https://doi.org/10.1007/978-3-030-70629-6_7

Bourahma, S., Mourchid, M., Mbarki, S. & Mouloudi, A. (2018). The Parsing of Simple Arabic Verbal Sentences Using NooJ Platform. *Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications*. 811, 81–95. Springer, Cham. doi: 10.1007/978-3-319-73420-0_7.

Faryadi, Q. (2007). *Techniques of Teaching Arabic as a Foreign Language through Constructivist Paradigm: Malaysian Perspective.* UiTM Malaysia.

Jendi, T.(1999) بحث في التركيب و الدلالة :المصدر المؤول . Publisher Dar Elhani. ISBN 977-222-183-7.

Ramadhanti Febriani, S. & Widayanti, R. The Evaluation of Arabic Learning based on Multiple Intelligences Classroom. *UIN Maulana Malik Ibrahim Malang*.

Using First Language Grammar to Support Second Language Grammar Acquisition in an Argentinian ESL Classroom by Resorting to the NooJ Platform

Mariana González
Instituto de Educación Superior N.º 28 "Olga Cossettini", Rosario, Argentina
marianagonzalez826@gmail.com
Andrea Rodrigo
Facultad de Humanidades y Artes, Universidad Nacional de Rosario, Argentina
andreafrodrigo@yahoo.com.ar

There have been many pedagogic approaches to Second Language Teaching that discourage teachers from using students´ first language in the classroom, such as the Natural Approach (Krashen & Terrell, 1983) and the Total Physical Response approach (Asher, 1977). The basic assumption is that language acquisition will take place if learners do not depend on their L1, or if they do not translate (Ellis, 1984). However, there are also many authors who recommend the use of mother tongue as a tool for teaching a second language. It is a fact that, in order to learn some structures in English that are similar to their own native language ones, students can make a comparison between the two. The question is: will students who are learning English as a second language benefit from making a comparison of the two languages' grammar?

The purpose of this paper is to determine whether it is more profitable, in order to enhance English acquisition, to compare English structures to their Spanish counterparts or to leave them aside during ESL Grammar lessons.

The first part of this paper will analyze the theories that support the use of native language in a second language classroom and how NooJ can foster the learning of grammatical structures (see Rodrigo and Bonino 2019). We chose to work with the NooJ platform developed by Max Silberztein (2015) (2016) because it enables us to work with both languages in isolation and promote the comparison of the two. Then, based on the data previously gathered, we will develop our tripartite action plan. In part one, we will explain the questionnaire that we asked the students to take as a diagnosis. In part two, we will show how the lessons of the same grammatical features were taught in two similar groups of students. In one class, students will be taught the grammatical feature compared with Spanish grammar and the other class will be taught the new grammatical feature only using English. All the grammatical structures will be explained by using the NooJ platform (see Spanish Module Argentina and English Module). In part three, we will show the test carried out by students assessing the level of understanding of the grammatical feature in question. Finally, we will analyze the results and reach a conclusion.

References

Asher, J. (1977). *Learning another language through actions: The complete teacher's guidebook.* Los Gatos, CA: Sky Oaks Production.

Ellis, R. (1984). *Classroom second language development*. Oxford: Pergamon Press.

English Module. Available at: https://atishs.univ-fcomte.fr/nooj/resources.html

Krashen, S. & Terrell, D. (1983). *The natural approach: Language acquisition in the classroom*. Hayward, CA: Alemany Press.

Rodrigo, A. & Bonino, R. (2019). *Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua*. Rosario: Ciudad Gótica.

Silberztein, M. (2015). *La formalization des langues, l´approche de NooJ*. London: ISTE.

Silberztein, M. (2016). *Formalizing natural languages: The NooJ approach*. London: ISTE-Wiley.

Spanish Module Argentina. Available at: https://atishs.univ-fcomte.fr/nooj/resources.html.

Extraction of phrasemes in the electrical field by NooJ

Tong Yang
North China Electric Power University, Beijing, China
tongyang@ncepu.edu.cn

Our study fits the teaching method French on specific objectives (Mangiante and Parpette, 2004) for Chinese workers who come to work in French electrical companies. In the electrical field, the grammatical pattern is very recurrent (e.g., coupe-circuit; prise électrique; bus-bar; attaché-fil; mise hors tension; bassin de retenue; planche à dessin). A grammatical pattern is "a syntactic configuration, a distribution pattern that integrates certain homogeneous classes of words" (Hunston & Francis, 2000: 33). According to them, a grammatical pattern can determine the meaning of the word and can also contribute to language learning. Before teaching this grammatical pattern to our students, the extraction of this structure becomes an unavoidable task. However, the extraction of sequences of multiple words always poses a problem in the TAL (automatic language processing) (Luka et al., 2006). According to the needs (disambiguation and flexibility) of our extraction, NooJ becomes the most appropriate software, because in theory, NooJ can describe all the natural languages of the world (Silberztein, 2015; 2016). The extraction of the pattern with NooJ is therefore the problematic of our communication in which we will present, first of all, our object of study: the grammatical patterns of the verbs of the electrical field. Then, the modeling and the disambiguation will be highlighted to achieve the automatic extraction. Inspired by our previous projects (Yang, 2018; 2019), our modeling is based on observations found in our corpus and in dictionary LVF (les verbes français, Dubois & Dubois-Charlier, 1997): an exhaustive lexicographic model of French verbs covering. Our disambiguation concerns both the nouns and verbs of recognized expressions. Take the rejected names for example, <ce>, <la>, <être>, <avoir>, <bien>, <si>, <été>, <y>, <g>, <h>, <tout>. Finally, we will implement the data in NooJ by elaborating some grammars in the form of a transducer and compile our data according to the NooJ codes, in particular the variable (prefixed by the character "$"), the global variable (prefixed by the character "@") and the colored nodes (prefixed by the character ":").

References

Dubois, J. & Dubois-Charlier, F. (1997). Les Verbes français. Larousse. Hunston, S. & Francis, G. (2000). *Pattern Grammar: a corpus-drivens approach to the lexical grammar of English*. Amsterdam/Philaelphia: John Benjamins.

Luka, N., Seretan, V. & Wehrli, E. (2006). Le problème des collocations. *TAL. Nouveaux cahiers de linguistique française.* 27 (95–115).

Mangiante, J.M. & Parpette, C. (2004). *Le français sur objectif spécifique: de l'analyse de besoins à l'élaboration d'un cours.* Hachette.

Silberztein, M. (2015). *La formalisation des langues: l'approche de NooJ*. International Society for Technology in Education.

Silberztein, M. (2016). *Formalizing Natural Languages: the NooJ Approach.* Wiley.

Yang, T. (2018). Automatic Extraction of the Phraseology Through NooJ. Mbarki S., Mourchid M., Silberztein M. (Eds). Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications. NooJ 2017. *Communications in Computer and Information Science* 811 (167-177). Springer, Cham.

Yang, T. (2019). Automatic Extraction of Verbal Phrasemes in the Culinary Field with NooJ. Mirto, I.M., Monteleone, M. & Silberztain M. (Eds). Formalizing Natural Languages with NooJ and Its Natural Language Processing Applications. NooJ 2018. *Communications in Computer and information Science.* 987 (83-94). Springer, Cham.

Towards a change of paradigm in language teaching in teacher training with computer tools

Virginia Gonfiantini
Facultad de Humanidades y Artes. UNR, Argentina
vgonfiantini@hotmail.com
Andrea Rodrigo
CETEHIPL, Facultad de Humanidades y Artes. UNR, Argentina
andreafrodrigo@yahoo.com.ar

As teachers and researchers at the Facultad de Humanidades y Artes from UNR – Argentina, we seek to present a series of observations we have made after analyzing a corpus of children's stories written by students attending teacher training courses for primary education[1]. These observations are intended to be no more than preliminary in nature and from a general approach focused on traditional language lessons (decoding reading, centered on repetition; lack of –and even under-stimulated- writing; writing development skills based on question guides generally grounded on the text sequence; poor stimulation of critical discussion since the final say is usually held by the teacher) to then expand upon the subject about the need for a global change of paradigm. This issue is a key factor which leads us to incorporate Edgar Morin's thinking and his approach to complex thinking. In this respect, Gonfiantini (2018)'s assertions have also been a significant contribution to develop this perspective. We believe this is an essential step to develop the desired changes or, at least, a starting point to generate the conditions for such changes.

Following the line of work of the Centro de Estudios de Tecnología Educativa y Herramientas Informáticas de Procesamiento del Lenguaje (CETEHIPL), we take the examples to work with from our corpus in order to show, from an interactive perspective, how students-users can play an active role in managing linguistic resources instead of being passive users of previously created resources. In this sense, the choice of NooJ as a platform of work is an asset: NooJ is a linguistic development environment constructed by Max Silberztein (2016) and is available for free download for academic purposes. Unlike other software, NooJ is not a black box. Instead, NooJ is an empty and totally interactive box in which students can do research work and create their dictionaries and grammars. Languages can also be compared and contrasted since NooJ's architecture features a common-format working environment which allows its users to work with different languages. This, far from being an obstacle, is a definite advantage since, we believe, multilingual perspective is an ideal setting to further enhance language learning.

This work is just a general approach and does not exhaust any aspect of this subject area. Indeed, we believe this work is complementary with some issues previously examined in Rodrigo and Bonino (2019). We believe this work also opens the specter for further discussions and takes account of other perspectives of analysis such as Edgar Morin's current of thought and whose contributions are an added-value to this subject. Thus, we will continue to explore our analysis with greater detail as we progress towards the interdisciplinary work proposed by our team.

References

Gonfiantini, V. (2018). Multiversidad mundo real Edgar Morin. *4. Rupturas epistémicas del Siglo XX y los procesos de cambio en educación. 593 Digital Publisher.* SEIT, ISSN 2508-0705. 3-3 (51-60). Quito.

Rodrigo, A. & Bonino, R. (2019). *Aprendo con NooJ: de la lingüística computacional a la enseñanza de la lengua.* Rosario: Ciudad Gótica.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. London: ISTE.

Spanish Module_Argentina: https://atishs.univ-fcomte.fr/nooj/resources.html

1 Escuela Normal Superior N° 36, "Mariano Moreno" and Escuela Normal Superior N° 35 "Juan María Gutiérrez", Rosario, Argentina