A Morphological Grammar for Modern Greek: State of the Art, Evaluation and Upgrade

Lena Papadopoulou[1(✉)] and Elina Chadjipapa[2(✉)]

[1] Hellenic Open University, Patra, Greece
papadopoulou.lena@gmail.com
[2] Democritus University of Thrace, Alexandroupolis, Greece
elinaxp@hotmail.com

Abstract. The objective of this paper is six-fold. First, a brief review of the state of the art of the Greek NooJ Module is outlined, in which the need for specific primary lexicographical resources is pointed out. Second, a corpus compilation using the entire text databank of the Centre for the Greek Language is described. Third, a dictionary of simple nouns and an integrated inflectional grammar, which are the outcome of the linguistic analysis of the new Greek NooJ corpus, are presented, with an emphasis on the full alignment of the latter with the inflectional codification of the Dictionary of Standard Modern Greek. Fourth, the compilation of a manual comprising guidelines for inflectional grammar editors is presented. Fifth, the validity of the aforementioned work has been tested through the processing of unknown word forms from the corpus. Sixth, future work is proposed in terms of the educational employment of the Greek NooJ Module and its implementation.

Keywords: Modern Greek · Corpus · Inflection · Lexicography · Simple nouns

1 Introduction

The first steps of the Modern Greek NooJ Module were made in 2007 [1] with the compilation of a dictionary of simple words and a corresponding inflectional grammar, which constituted the basis of the present work. Since then, lexicographical data as well as morphological and syntactic grammars have been imported in order to improve the results of Greek language automatic processing.

Among others, local grammars for the automatic recognition of proper nouns have been compiled [2]. A Greek-Spanish NooJ module has been created [3], where the equivalence between these two languages is studied for educational purposes. Enriched versions of the Greek NooJ module have been developed, such as in the case of the lexicographical elaboration of simple and multiword adverbs, acronyms, and borrowed words written using the Latin alphabet [4]; in addition, formalized methods have been proposed for data enrichment, such as for adjectives [5]. Lexicographical data compilation has been conducted on specific classes of objects, such as professional nouns [6] and <material> predicative adjectives [7]. The compounding and derivation of specific categories – such as neoclassical compounds [8], numeral-noun/adjective construction

B. Bekavac et al. (Eds.): NooJ 2020, CCIS 1389, pp. 3–13, 2021.
https://doi.org/10.1007/978-3-030-70629-6_1

[9], and the derivation of multiply complex negative adjectives from verbal stems [10] – have been studied. Furthermore, phraseological units, such as frozen expressions [11] and pragmatemes [12], have been processed.

Although, so far, rich lexicographical data have been produced and a series of linguistic phenomena have been studied, the accomplished work has mainly been based on secondary lexicographical resources. As a consequence, a dedicated corpus was required: a corpus that would meet the quality requirements of our project, a corpus that would comprise representative authentic texts and would be easy to handle as far as size and representativity are concerned. Such requirements seemed to be fulfilled by the text databank of the Centre for the Greek Language.

Therefore, in the present work, first, both primary and secondary lexicographical resources are defined. Afterwards, the procedure that has been followed for the retrieval and processing of simple nouns, as far as their dictionary compilation and inflectional properties attribution are concerned, is described. In addition, the manual for inflectional grammar editing is outlined. Finally, the results of our work plans for future work are presented.

2 Lexicographical Resources

The lexicographical resources that have been defined for our project are primary and secondary. On one hand, the entire text databank of the Centre for the Greek Language was designated as the primary lexicographical resource. On the other hand, a series of previous Greek NooJ data and the Dictionary of Standard Modern Greek [13] were chosen as secondary lexicographical resources.

2.1 Primary Lexicographical Resources

The entire text databank of the Centre for the Greek Language has served as the primary lexicographical resource for our corpus compilation. This choice was dictated by the quality requirements that our project set for itself, and it can be justified based on four main criteria: (a) resource reliability, (b) material purposes, (c) text representability, and (d) corpus size.

First, as far as the resources are concerned, they have been retrieved as educational material for the teaching of Modern Greek as a foreign/second language by the Support and Promotion of the Greek Language research division of the Centre for the Greek Language [14]. The Centre for the Greek Language[1] acts as a cooperating, advisory and planning body of the Ministry of Education on matters of language policy. It is an academic institution dedicated to the description and documentation of trends in the Modern Greek language, and therefore, it follows strictly scholarly methods. Consequently, the text databank is considered reliable with respect to its methods of text compilation.

Second, the text databank in use has been compiled for educational purposes, such as to assist students who take an exam for the Certification of Attainment in Greek. Thus, it is in accordance with the aims of the Greek NooJ module, given that the main

[1] https://greeklanguage.gr/en/.

perspective of the latter is to use the NooJ environment as a tool for Greek language learning and teaching.

Third, the text databank is a compilation of originally written and spoken texts from a wide range of sources including different genres and text types. Consequently, the representability criterion is completely fulfilled. This is considered a feature of great importance in view of the beneficial impact of students' exposure to diverse authentic texts [14].

Fourth, the text databank of the Centre for the Greek Language fulfills the size criterion, given that it is feasible to deal with total data volume, considering the above-mentioned qualitative features.

In total, the corpus includes six (6) text files, one for each level according to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) [15], comprising a total of 336 text units and 117,892 word forms (Fig. 1):

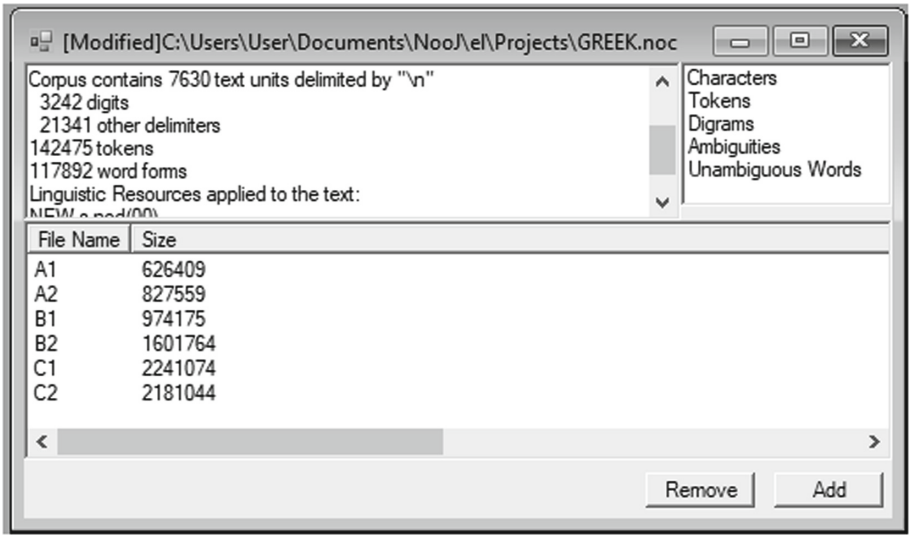

Fig. 1. The GREEK.noc corpus

Table 1 presents thorough information regarding corpus structure, providing information about the number of texts and word forms that each text file comprises.

2.2 Secondary Lexicographical Resources

A series of previous Greek NooJ data, as well as the Dictionary of Standard Modern Greek [13] (hereinafter DSMG), are defined as secondary lexicographical resources for our project.

The DSMG is a monolingual comprehensive definitional, orthographic, and etymological dictionary of Modern Greek published by the Institute of Modern Greek Studies at the Artistotle University of Thessaloniki in both paper and digitalized format. The

Table 1. Corpus data

CEFR level	Number of texts	Number of word forms
A1	78	9,059
A2	77	11,273
B1	45	12,229
B2	45	21,828
C1	49	32,494
C2	42	31,009
TOTAL	336	117,892

DSMG has been selected for two main reasons: (a) it provides an opportunity for online research[2] and (b) it annotates a link between each entry and its inflectional model.

In addition to the DSMG, the Greek NooJ dictionary, from which all semantic and syntactic information was excluded; the inflectional grammar, comprising a total of 757 inflectional rules of which 266 refer to nouns; and the grammar, which processes the double accent in proparoxytones, were applied as resources for the linguistic analysis of the corpus.

3 Procedure

The procedure that has been followed consists of two major stages. The first consists of the retrieval of nouns, while the second consists of the parallel compilation of a dictionary of nouns and an inflectional grammar as well as the redaction of a manual for inflectional grammar editing.

3.1 Noun Retrieval

In the noun retrieval process, through which a validation test of noun lexicographical data was carried out in parallel, three major steps were involved: (a) corpus linguistic analysis within NooJ, (b) noun extraction, and (c) database compilation.

Within the first step, the Greek NooJ dictionary (GLE), the inflectional grammar, and the grammar that processes the double accent in proparoxytones were applied as resources for the linguistic analysis of the corpus (Fig. 2).

The results that the linguistic analysis produced are considered quite satisfactory. They conclude that there were only 2,660 unknowns, that is, out of the total number of 117,892 word forms encountered in the corpus, 2.25% were unknown.

In the second step, nouns were extracted in a semiautomatic way with the aid of ambiguities and unambiguous word annotations. Once we got the output of the annotations of the ambiguous and unambiguous words, the information regarding the lemma, the grammatical category, and the corresponding inflectional paradigm was filtered out.

[2] At https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/search.html?lq=.

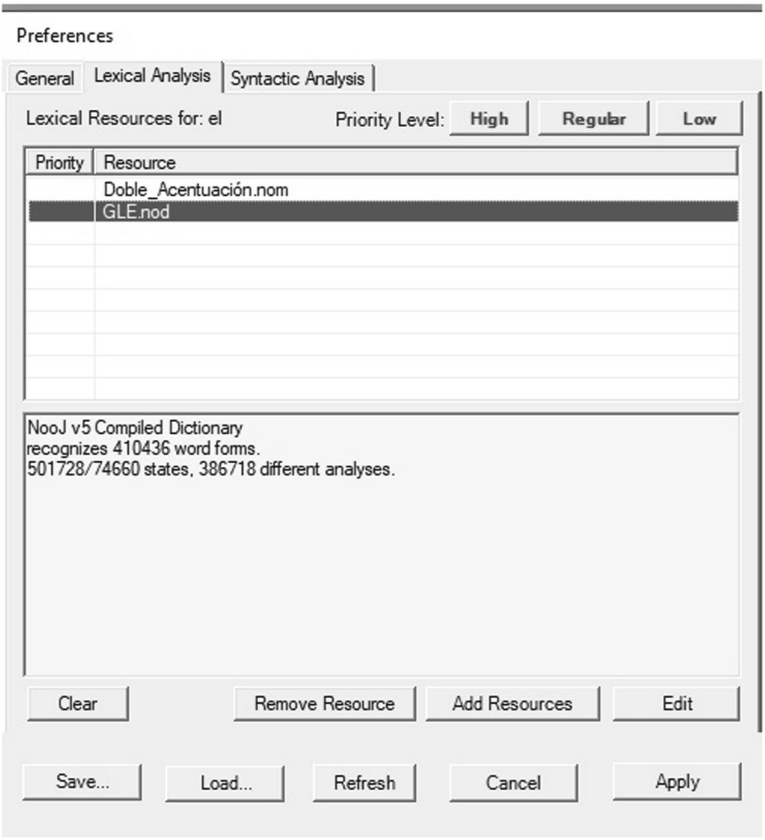

Fig. 2. NooJ data

Subsequently, 3,464 lemmas were annotated as nouns along with their corresponding inflectional property codification. These lemmas consisted of the base on which the manual validation process was grounded.

3.2 Nouns Dictionary and Inflectional Grammar Processing

Once the database was set up, a four-step procedure was followed, comprising (a) the exclusion of word forms, (b) the correspondence of inflectional codes, (c) the simplification of inflectional rules, and (d) the elimination of non-active paradigms.

Firstly, a series of word forms was excluded. On one hand, these word forms included nominal word forms located exclusively within multiword units. This elimination is due to our particular focus on simple nouns. Within this framework, for example, the nomineme[3] *Αγία Σοφία* (EN: Hagia Sophia) was deleted. On the other hand, ambiguous word forms regarding lexical units and part-of-speech properties that are not used in

[3] See Mel'čuk [16] for more information on lexical phraseme classification.

the corpus were removed. For instance, the form κρατών corresponds both to the nominalized participle κρατών (EN: prisoner) in nominative singular and the noun κράτος (EN: state) in the genitive plural. Given that only the lexical unit κράτος was located, the lemma κρατών was deleted. The same disambiguation process was followed with forms belonging to other forms of speech. For example, lemmas such as the adjective βάρβαρος (EN: barbarian), the verb πιστεύω (EN: believe), and the pronoun εγώ (EN: I) were eliminated given that their form belongs to a non-noun form in our corpus.

In the second stage, the correspondence between lemmas and inflectional paradigms was manually attributed in our database. This way, inflectional codification has been absolutely aligned with the categorization of the DSMG, which comprises 68 broad inflectional models for nouns, which are represented in the following way in the DSMG:

O40	πρόσωπο	προσώπου	πρόσωπο	πρόσωπο	πρόσωπα	προσώπων	πρόσωπα	πρόσωπα

Fig. 3. Inflectional model O40 in the DSMG

The third step, which was the most challenging and the most time consuming, concerned the processing of inflectional paradigms within the inflectional grammar and the database in parallel. At that point, previous paradigms were recodified and simplified as regards their nomenclature (see Sect. 3.3 in the Manual) and their internal structure, respectively. This procedure was significantly facilitated by the supplementary utility of operator <A>, through which an accent can be removed not only in the last letter but also in the entire word form. Such utility has definitely contributed to the reduction of the inflectional grammar's size (Fig. 4 and 5).

```
#φύλακας#
N5 = <E>/nom+m+s + <B>/gen+m+s + <B>/acc+m+s + <B>/voc+m+s +
     <B2>ες/nom+m+p + <L5><A><R2><Á><RW><B2>ων/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
#μεσαίωνας#
N5_a = <E>/nom+m+s + <B>/gen+m+s + <B>/acc+m+s + <B>/voc+m+s +
     <B2>ες/nom+m+p + <L4><A><R><Á><RW><B2>ων/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
#ήρωας#
N5_b = <E>/nom+m+s + <B>/gen+m+s + <B>/acc+m+s + <B>/voc+m+s +
     <B2>ες/nom+m+p + <L4><A><R2><Á><RW><B2>ων/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
#Έλληνας#
N5_c = <E>/nom+m+s + <B>/gen+m+s + <B>/acc+m+s + <B>/voc+m+s +
     <B2>ες/nom+m+p + <L6><A><R3><Á><RW><B2>ων/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
#ελέφαντας#
N5_d = <E>/nom+m+s + <B>/gen+m+s + <B>/acc+m+s + <B>/voc+m+s +
     <B2>ες/nom+m+p + <L6><A><R2><Á><RW><B2>ων/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
#υδατάνθρακας#
N5_e = <E>/nom+m+s + <B>/gen+m+s + <B>/acc+m+s + <B>/voc+m+s +
     <B2>ες/nom+m+p + <L7><A><R4><Á><RW><B2>ων/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
```

Fig. 4. Operator <A> removing an accent in the last letter in paradigm N5

```
#φύλακας#
N5 = <E>/nom+m+s + <B>/gen+m+s + <B>/acc+m+s + <B>/voc+m+s +
     <B2>ες/nom+m+p + <B2><A>ων<L2><Ά>/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
#ελέφαντας#
N5a = <E>/nom+m+s + <L2><B3>/nom+m+s + <B>/gen+m+s +  <B4>/gen+m+s + <B>/acc+m+s
+ <B4>/acc+m+s + <B>/voc+m+s + <B4>/voc+m+s +
     <B2>ες/nom+m+p + <B2><A>ων<L2><Ά>/gen+m+p + <B2>ες/acc+m+p + <B2>ες/voc+m+p;
```

Fig. 5. Operator <A> removing an accent from an entire word form in paradigm N5

Given that Greek is a heavily inflected language – for example, the inflection of nouns includes four cases (nominative, genitive, accusative, and vocative) and two numbers (singular and plural) within the declension many times the accent is being both removed and moved and there are nouns that have double inflectional forms – high precision is required in inflectional rule editing. Such precision results in the generation of more inflectional paradigms than the DSMG provides. Consequently, the proportion between the DSMG's inflectional paradigms and ours is in total 68 to 122, almost 1 to 2. For example, the inflectional model O40 in the DSMG corresponds to 6 different inflectional paradigms in the NooJ inflectional grammar (Fig. 3 and 6).

```
#πρόσωπο#
N40 = <E>/nom+n+s + <A><L2><Ά><RW>υ/gen+n+s + <E>/acc+n+s + <E>/voc+n+s +
      <B>α/nom+n+p + <A><L2><Ά><RW><B>ων/gen+n+p + <B>α/acc+n+p + <B>α/voc+n+p;
#γυμναστήριο#
N40_a = <E>/nom+n+s + <A><L><Ά><RW>υ/gen+n+s + <E>/acc+n+s + <E>/voc+n+s +
      <B>α/nom+n+p + <A><L><Ά><RW><B>ων/gen+n+p + <B>α/acc+n+p + <B>α/voc+n+p;
#ένστικτο#
N40_b = <E>/nom+n+s + <A><L3><Ά><RW>υ/gen+n+s + <E>/acc+n+s + <E>/voc+n+s +
      <B>α/nom+n+p + <A><L3><Ά><RW><B>ων/gen+n+p + <B>α/acc+n+p + <B>α/voc+n+p;
#περίχωρα#
N40-s = <E>/nom+n+p + <A><L2><Ά><RW><B>ων/gen+n+p + <E>/acc+n+p + <E>/voc+n+p;
#Χριστούγεννα#
N40-s_a =  <E>/nom+n+p + <A><L3><Ά><RW><B>ων/gen+n+p + <E>/acc+n+p + <E>/voc+n+p;
#γενέθλια#
N40-s b =  <E>/nom+n+p + <A><L><Ά><RW><B>ων/gen+n+p + <E>/acc+n+p + <E>/voc+n+p;
```

Fig. 6. Paradigm N40 in the NooJ inflectional grammar

Finally, the fourth step consisted of deleting non-active paradigms, given that they are not related to any lemma of the noun dictionary in use (Table 2).

Table 2. Structure of inflectional grammar

Type	Number of paradigms	Codes in grammar
Masculine	34	*N1-N22*
Feminine	30	*N23-37*
Neutral	42	*N38-N53-s*
Non-inflected	3	*N_INDECm- N_INDECn*
Irregular nouns	3	*NIRRn1-NIRRn3*
Nominalized adjectives	9	*NA1m-NA17f*
Nominalized past perfect participles	1	*NPARPPmm*

In conclusion, a dynamic morphological grammar for simple nouns based both on primary and secondary lexicographic resources has been completed. Such a dynamic applies both to the Greek inflectional grammar as well as to the Greek NooJ dictionary, given that the introduction of new lemmas and their inflectional paradigms correspondence have been performed smoothly in the case of the 1,101 word forms of 947 new nouns.

3.3 Manual

Our aim to align the codification of paradigms with those of DSMG seemed doomed to failure in case of lemmas that either the DSMG does not comprise at all, such as proper names and gentilics, or does not provide them with an inflectional paradigm, such as nominalized words. Such a failure has been avoided by the development of a redaction manual. The manual aims to serve as a guideline for the inflectional grammar compilation for present and future versions by old and new users.

On one hand, useful information about the nomenclature of paradigm is provided. The first number of each paradigm name corresponds to DSMG codification. Meanwhile, information following an underscore (_) refers to accentuation variants and inflectional particularities, while information introduced by a hyphen (-) indicates the exclusion of grammatical categories (Table 3).

For example, the code NA5n-s indicates that the paradigm refers to a nominalized noun (N) which is inflected via adjective inflectional model “1” (A1) according to the DSMG and it does not have any singular forms (-s).

On the other hand, a series of conventions has been created in order to formalize lemmas that are not included in the DSMG. Such standardization has been followed mainly for proper nouns and gentilics that the DSMG generally does not include. For example, the codes N25a and N35 are proposed for feminine proper names ending in *-ία* and *-ος*, respectively.

Table 3. Indicators in paradigm nomenclature

Indicator	
_a, b, c…	*subclass of paradigm because of accent movement*
-	*exclusion*
A1, 2, 3…	*adjective inflectional model*
ac	*accent*
f	*feminine*
g	*genitive*
INDEC	*non-inflected*
irr	*including an irregular form*
IRR	*irregular inflection overall*
m	*masculine*
n	*neutral*
p	*plural*
PARPP	*past perfect participle*
s	*singular*

4 Results

Once the dictionary and the inflectional grammar of simple nouns were compiled, we proceeded to the processing of unknown forms, which were extracted from the database.

Due to the high number of typographical errors, which concerned mainly the interference of similar Latin characters (Fig. 6) in Greek words, a new file was introduced in order for linguistic analysis results to be optimized. This file has in view the aforementioned interference, by providing the Greek-Latin correspondence of similar characters, so that such word forms will henceforth be recognized, thus reducing the total number of unknown word forms (Fig. 7).

The database of the unknowns comprised 2,660 word forms in total, which were manually analyzed. The total number of noun word forms amounts to 1,101, which corresponds to 947 entries. It has to be pointed out that only simple nouns have been considered, while noun word forms that appear solely in multiword units were excluded.

Through this process our main aim, which is the aim of a dynamic morphological grammar for simple nouns based both on primary and secondary lexicographic resources, has been achieved. Such a dynamic applies both to the Greek inflectional grammar as well as to the Greek NooJ dictionary. The introduction of new lemmas and their corresponding inflectional paradigms has been performed in a systematic and lower time-consuming way.

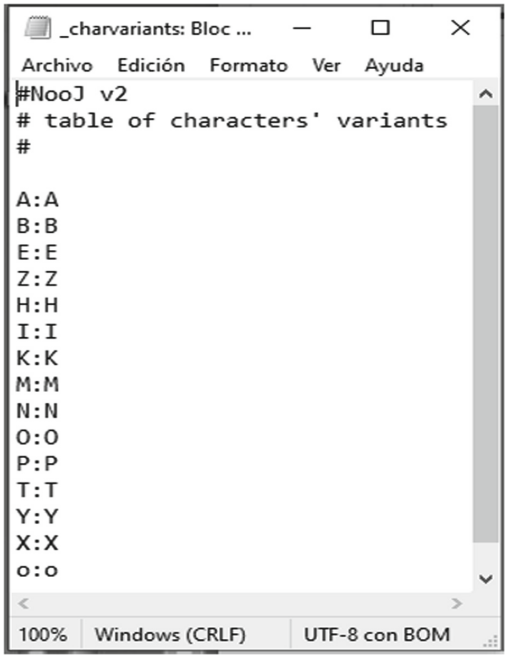

Fig. 7. Variants: Greek - Latin characters

5 Future Work

Within the present work a series of resources has been developed: (a) the compilation of a corpus for educational purposes, (b) a dictionary of simple nouns based on corpus resources, (c) an integrated inflectional grammar for simple nouns, and (d) a manual for inflectional grammar editing.

Undoubtedly, the present work will be the engine for future work on the educational purposes of the Greek NooJ Module. In view of its implementation, the linguistic analysis of the corpus has to be improved. First, the same procedure has to be followed for all parts of speech and for multiword units. This will smoothen the way to proceeding to the syntactic and semantic processing of our corpus in order to reach our ultimate aim: for each lemma of our dictionary to correspond to a lexical unit. Subsequently, lexical richness and core vocabulary could be studied in the future.

References

1. Gavriilidou, Z., Chadjipapa, E., Papadopoulou, E., Giannakopoulou, A.: The New Greek NooJ Module: morphosemantic issues. In: Blanco, X., Silberztein, M. (eds.) Proceedings of the 2007 NooJ International Conference, pp. 96–103. Cambridge Scholars Publishing, Cambridge (2008)

2. Gavriilidou, Z., Papadopoulou, E., Chadjipapa, E.: New data in the Greek NooJ module: a local grammar of proper nouns. In: Silberztein, M., Varadi, T. (eds.) Proceedings of the 2008 International Conference (Budapest), pp. 93–100. Cambridge Scholars Publishing, Cambridge (2010)
3. Papadopoulou, E., Gavriilidou, Z.: Towards a Greek-Spanish NooJ module. In: Hamadou, A., Mesfar, S., Silberztein, M. (eds.) Finite State Language Engineering: NooJ 2009 International Conference and Workshop (Touzeur), pp. 301–315. Centre de Publication Universitaire (2010)
4. Papadopoulou, E., Chadjipapa, E.: Version 4 Greek NooJ Module: adverbs, acronyms and words with Latin characters. In: Gavriilidou, Z., Chadjipapa, E., Papadopoulou, L., Silberztein, M. (eds.) Proceedings of the NooJ 2010 International Conference and Workshop, pp. 95–101. Komotini (2011)
5. Papadopoulou, E., Anagnostopoulos, G.: Enrichment of the Greek NooJ module: morphological properties and translation equivalence of Greek adjectives. In: Silberztein, M., Donabédian, A., Khurshudian, V. (eds.) Formalising Natural Languages with NooJ, pp. 182–193. Cambridge Scholars Publishing (2013)
6. Chadjipapa, E., Papadopoulou, L.: Greek professional nouns processed with NooJ. In: Gavriilidou, Z., Chatzipapa, E., Papadopoulou, L., Silberztein, M. (eds.) Proceedings of the NooJ 2010 International Conference and Workshop, pp. 183–191. Komotini (2011)
7. Gavriilidou, Z., Papadopoulou, L., Chadjipapa, E.: <material> predicative adjectives in Greek NooJ module. In: Monti, J., Silberztein, M., Monteleone, M., Pia di Buono, M. (eds.) Formalising Natural Languages with NooJ 2014, pp. 49–54. Cambridge Scholars Publishing (2015)
8. Gavriilidou, Z., Papadopoulou, L.: Greek neoclassical compounds and their automatic treatment with NooJ. In: Gavriilidou, Z., Chadjipapa, E., Papadopoulou, L., Silberztein, M. (eds.) Proceedings of the NooJ 2010 International Conference and Workshop, pp. 73–83. Komotini (2011)
9. Gavriilidou, Z., Papadopoulou, L., Chadjipapa, E.: Numeral-noun and numeral-adjective construction in Greek. In: Silberztein, M., Donabédian, A., Khurshudian, V. (eds.) Formalising Natural Languages with NooJ, pp. 113–122. Cambridge Scholars Publishing (2013)
10. Gavriilidou, Z., Papadopoulou, L.: Derivation of multiply complex negative adjectives from verbal stems in Greek. In: Koeva, S., Mesfar, S., Silberztein, M. (eds.) Formalising Natural Languages with NooJ 2013: Selected papers from the NooJ 2013 International Conference, pp. 63–68. Cambridge Scholars Publishing (2014)
11. Gavriilidou, Z., Papadopoulou, E., Chadjipapa, E.: Processing Greek frozen expressions with NooJ. In: Vučković, K., Bekavac, B., Silberztein, M. (eds.) Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik), pp. 63–74. Cambridge Scholars Publishing (2012)
12. Papadopoulou, L.: Local grammars for pragmatemes in NooJ. In: Monti, J., Silberztein, M., Monteleone, M., Pia di Buono, M. (eds.) Formalising Natural Languages with NooJ 2014, pp. 122–128. Cambridge Scholars Publishing (2015)
13. Dictionary of Standard Modern Greek (1998). https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/index.html. Accessed Sept 2020
14. Κέντρο Ελληνικής Γλώσσας (n.d.). https://www.greek-language.gr/certification/dbs/teachers/index.html. Accessed May 2020
15. Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. https://rm.coe.int/1680459f97. Accessed May 2020
16. Mel'čuk, I.: Clichés, an understudied subclass of phrasemes. In: Buhofer, A. (ed.) Yearbook of Phraseology, pp. 55–86. De Gruyter, Berlin (2015)