Error description and standardization of Modern Greek based on a context-sensitive grammar: Implications for the teaching of mother tongue

Gakis Panagiotis[1], Kokkinos Theodoros[2], Tsalidis Christos[3]
gakis@sch.gr, theokokkinos@yahoo.gr, tsalidis@neurolingo.gr
[1]University of Peloponnese, [2]Hellenic Open University, [3]Neurolingo Company

Abstract

The current study presents an analysis regarding the standardization of grammatical phenomena as well as corresponding templates for the description of grammatical phenomena / errors in order to develop an advanced, innovative and state-of-the-art computing environment for the creation of digital educational games for students. These templates were based on studies on the definition of linguistic error and focus on cases of common errors and deviations in the field of morphology / inflection, spelling and syntax. The implementation formalism used is Mnemosyne's Kanon, a complete complex natural language processing system that is applied both for information retrieval and extraction in free texts. At the same time, the grammar checker that has been developed based on the aforementioned process, was put to the test during the correction of authentic texts of primary school students (N = 100), which were produced in the context of teaching scenarios in order to investigate its usability and contribution to their linguistic competence. Main findings are that a context sensitive grammar can address the problem of correcting spelling mistakes that result in legitimate words. Moreover, it was found that the proposed formalism, the templates and, consequently, the grammar checker were overall functional in detecting most of the participants' usual errors.

Keywords: error description, error standardization, grammar checking, language teaching, digital technologies

Introduction

The aim of this research is the evaluation of the friendly electronic tool (grammar checker) by students of primary education. This tool carries out the morphological and syntactic analysis of sentences, phrases and words in order to correct syntactic, grammatical and stylistic errors. The foundation of these issues is the settings of Grammar (adaptation of Little Modern Grammar of Manolis Triantafyllidis), which is the formal, since 1976, constituted codified grammar of Modern Greek. The absence of this tool for the Greek, the product's development is based on the detailed record, the analysis and the formulation of the errors of writing speech and then on the choice of the software that will describe the grammatical errors.

This research presents the formalism that was used (the Mnemosyne) and presents the particularities of Greek language which hinder the computational processing. The formalism has already been used to identify multi-word terms and to phrasing grammars, aiming to the automatic information extraction. In this way, all speakers (native or not) will be able to understand better not only the function of various parts of the system of language but the way the mechanisms of linguistic analysis operate in the conquest and more broadly in the linguistic realization.

The basis for the implementation is the electronic morphological lexicon (Neurolingo Lexicon), a 5-level lexicon, consists of, at least 90,000 entries, which produce ~1,200,000 inflection types. These types carry information: a) spelling, b) morpheme information, c)

Panagiotakopoulos, C., Karatrantou, A., & Armakolas, S. (Eds.), *Proceedings of 7th Panhellenic Scientific Conference "Integration and Use of ICT in the Educational Process"*, pp. 42-54, Dept. of Educational Sciences and Social Work, University of Patras, Greece, 16-18 September 2022. ISBN 978-618-83186-7-0.

morpho-syntactic information, d) stylistic information and e) terminology.

A major problem in natural language processing was the lexical ambiguity, a product of the highly morphology of the Greek. Given that the major problem of modern Greek was the lexical ambiguity we design the Greek tagger grounded on linguistic criteria for those cases where the lexical ambiguity impede the imprint of the Greek language errors.

Theoretical Background

Language Errors

The concepts of "right" and "wrong" generally cause concern to all languages' native speakers and to the speakers of the Modern Greek language as well (James, 1988; Iordanidou, 1999, 2004). Language error is described by the majority of scholars as a deviation from the applicable rules and norms, so that it is considered an element of differentiation from what is not acceptable (Ellis, 1994). The linguistic error is a speech deviation that did not occur deliberately, with the speaker not being able to correct it even if its existence is pointed out. According to researchers, the only way of doing this is by being exposed to further language teaching. Errors cannot always be accurately identified, because there is not a unique system of linguistic expression nor a category of spelling, grammar and syntax rules (Athanassiou, 2001). Thus, when reference is being made to "errors" one may distinguish between systematic errors (errors) and random errors (mistakes). Specifically, random errors (mistakes) are those that the speaker can perceive almost immediately and correct, while systematic ones are those, which happen mainly due to ignorance or misunderstanding of a language system by the speaker (Corder, 1974). Linguistic errors are strongly related to the concepts of acceptability and grammaticality, with the latter - according to Chomsky (1965) - not being the only factor determining acceptability. Moreover, inference errors occur when language learners are in contact with native speakers who have systematically incorrect formulas in their language repertoire.

Additionally, it is argued that language standardization is necessary in order to facilitate communication, to make the establishment of an agreed orthography possible, and to provide a unified form for school books. Language standardization practically means that there are predefined guidelines and vocabulary for writing content. However, there are plenty of other text types that would benefit from some standardization — especially in teaching.

Regarding the Modern Greek language, the basis for dealing with all the aforementioned issues is the regulations of school grammar (Holton et al., 2002), which adequately defines the written -mainly- conventions of the Modern Greek language, such as spelling. At the same time, it sets the fundamental principles on which the regulation of the language code is based and suggests specific practical solutions (James, 1998). Naturally, there are some special cases that are not fully described and standardized, not only because the original writing of these grammar settings dates back to the decade 1940-1950 (Triantafyllidis, 1941), but also because only after 1976 they have been widely used in all types of written speech.

The need of language use regulation (for both educational and practical reasons) must not be based on subjective views but on a thorough and scientific description of the language system and the variety that characterizes the speakers' linguistic practice (Bartholomae, 1980). Analysis also happens in "errors" of the learners' speech production (written-oral), a process that serves the teaching process. Its usefulness lies in the fact that students tend to avoid specific structures for which they are unsure, and therefore they do

not make mistakes in areas where such linguistic errors would be expected (James, 1998). In addition, errors function as mechanisms through which students discover the rules of language.

Any solutions must be made in accordance with the rules of the official school grammar, be practical and flexible and provide the user of modern Greek with the widest possible range of options, in order to achieve a text production that respects some general rules while, at the same time, highlights the "individuality" of personal expression and the speech genre to which it belongs (Ancker, 2000).

Implementation Grammar

Historically, grammar formalisms are the result of separate research in linguistics, computational linguistics and natural language processing (Gakis et al., 2016). The formalism of grammar unification can be traced in many studies.

Three parameters serve as important criteria of these formalities: a) linguistic "happiness" (the degree to which descriptions of linguistic phenomena can be stated directly or indirectly by linguists), b) expressiveness (what category of analysis can be declared) and c) computational efficiency (Shieber, 2001).

It should be noted that there are plenty online grammar checkers through which try to offer error standardization through specific formalisms. It is beyond the scope of a grammar checker (Greek and other checkers) to identify mistakes along with missing fragments, run-on sentences, wrong expressions, and wrong paragraph boundaries. Tense usage and pronominal reference are equally beyond their ability to correct. The other formalisms identify certain specific types of grammatical errors in the proposed domain of application that are more regular than others. Existing grammar checking systems, such as those described in several studies (Genthial & Courtin, 1992), fall into this discipline, addressing the issue with a collection of heuristic rules to decide on subject and object in unclear cases, it might pick the wrong distribution and not flag anything, although it should do so in exactly this case. CorrecText, from Houghton-Mifflin, is a significant advance in grammar checkers, because it uses a full parse of sentences in its analysis (Dobrin, 1990).

The Current Study

Method

The current study's goal is to propose an error standardization method based on context-sensitive grammars and consequently test its feasibility and practicality with native speakers (primary students).

Specifically, the questions that drove the study were:

(1) Can a context-sensitive grammar constitute the basis for an effective error standardization method within the framework of an online grammar checker?

(2) Can such a tool be practical for native speakers whose linguistic competence is still under development (primary students)?

The authors of the current study hypothesize that for both research questions the answer is affirmative.

The errors presented in the study were chosen based on the rationale of sharing high frequency in most European languages. Nevertheless, the method that these errors can be described and standardized can be utilized by most natural languages. The native speakers' sample was primary (fifth grade) students (N = 100), who were asked to write a text in the framework of a teaching scenario related to television. The teaching scenario was

conceptualized and implemented by postgraduate students of the University of Patras in Western Greece in 2018. Authentic students' texts were transcribed to ".doc" files and inputted in the online grammar checker created based on the formalism that is described in the next section.

Results

Research Question 1

Regarding the first research question of our study, errors were described and standardized based on the implementation of Mnemosyne's Kanon, which is a combination of Context Sensitive Grammars and Unification Grammars. Unification-based grammar formalisms have recently received great interest in computational linguistics because they allow for a very high-level, declarative, and modular description of grammatical relations over linguistic objects. These formalisms have the status of very high-level programming languages that are especially well-suited to encode linguistic theories. They provide the means to represent linguistic objects using feature structures and to encode additional theory-specific principles (Seiffert, 1987). The grammar formalisms used to produce the language analyzers that consist of a parser with actions executed for the grammatical constructs recognized. The parser applies a surface parsing method i.e. does not try to recognize and reduce the full structure of the input text but only parts of it defined in the grammar rules. In our case the analyzer created is the grammar checker. Through this grammar it is possible not only to describe polysyllabic terms but also to define the phrase - word type with the wrong information in an automated way utilizing at the same time the morphological - stylistic characteristics of the word types described in the electronic morphological dictionary/lexicon[1]. The standardization of grammatical errors as well as their categorization/classification was done by creating special templates. This formalism is also used in the Greek Grammar Checker (Gakis et al., 2015; 2016).

More specifically, the grammar analyzer arose from the Kanon grammar rules based on templates and is responsible for checking the text stylistically and morpho-syntactically. Terminal rules are defined as rules that discover a semantic entity (name entity, action, or event). In our case it is the rules that recognize and extract a problematic grammatical error. The result of this morphological analysis is the identification of the various grammatical errors and the recording of the production tree that reflects the part of the syntactic analysis that is problematic. It mainly uses context-sensitive grammar rules, which is a standard grammar where the left and right parts of each rule can be replaced by a set of terminal and non-terminal symbols (Chomsky, 1965). Errors are described through unification grammars that allow context-sensitive grammars to be defined, considering, in other words, their context (McCord, 1987). These grammars are stronger than context-free grammars, which have a similar structure to the grammar of natural language (series of variables: article, adjective, noun, verb, etc.). In each template, in addition to the fields that describe the type of grammatical phenomenon, there are hints/suggestions that justify the correct use of words and provide users with the necessary explanations for their choices.

The format of the template, which will be used to describe grammatical errors, are context sensitive rules with the following syntactic structure:

$[H_1], [H_2], \dots, [H_v] =>$
$<< C_1, C_2, \dots, C_\xi >>$
$[L_1], [L_2], \dots, [L_\kappa]$
$\backslash$
$[M_1], [M_2], \dots, [M_\mu]$

/
[R_1], [R_2], …, [R_λ]
;

The basic unit that forms the rule is defined between square brackets ('[', ']'), called textspan, represents a sequence of one or more tokens in input and consists of zero or more predicates that specify the conditions that must be fulfilled in order to have a match. According to the syntax of the template, we distinguish four types of textspans with names H (Head), L (Left), M (Main) and R (Right), respectively. The L textspans define the left context, the R textspans define the right context and the M, the main context. The application of the rule will replace the M textspans with the H textspans leaving the left and right textspans intact. The Cs in the template defines rule conditions that must be validated before the application. The constituent parts of these conditions are values collected from the L, M, R textspans.

More specifically in a rule there is: a) the corresponding phrase or word or entry that is the subject of identification and is located between the symbols '\' and '/, b) the left part of the expression before the symbol' \ which is a set of words, phrases, generally tokens that are useful for identifying the expression but are not replaced by the rule head and c) the right part of the expression followed by the '\' symbol.

So, if we have the rule:

H1 H2 => L1 \ M1, M2, M3 / R1, R2 ;

and the following sequence of textspans in input

L1 M1 M2 M3 R1 R2

The application of the rule will give the output

L1 H1 H2 R1 R2

The definition in the textspans is done with a sequence of feature value conditions.

Additional information can be extracted by determining the equivalent morphological attributes, in order to avoid identifying the word with another word with the same ending but different morphological attributes. Thus, in the rule [**SUFFIX**="*ουνα*", MORPHOLOGY = {VERB, A_PERSON, SINGULAR}] words are extracted ending in –"*ουνα*" and morphological attributes [Verb and A' Person, Singular].

These terms can also be defined in terms of agreement at many levels:

(1) Agreement in gender, number and case [AGREEMENT (GNC)],

(2) Agreement in number and case [AGREEMENT (NC)],

(3) Agreement only in number [AGREEMENT (N)] or

(4) AGREEMENT only in case [AGREEMENT (C)]

The following rule describes the agreement in gender number and case:

```
\
    [WORD="ο",
        MORPHOLOGY={ARTICLE,ACCUSATIVE, MASCULINE,SINGULAR}
        AGREEMENT={GENDER, NUMBER, CASE} ],
        [MORPHOLOGY={NOUN,ACCUSATIVE, MASCULINE,SINGULAR}
        AGREEMENT={GENDER, NUMBER, CASE} ],
/
;
```

The agreement between the article and the noun is checked, which must agree in the same morphological attributes at the level of gender, number, and case. Control of the agreement may face more specific, rare cases that allow for better management of syntactic rules. Thus, in the phrase: "τις μυστικές τους δυνάμεις" [their secret forces] the sequence of POS is as follows: article, adjective, ambiguous word: (PRONOUN / ARTICLE), noun. Similarly, in the phrase: "την καλή την κοπέλα" [the good girl], the POS sequence has the same form: article, adjective, ambiguous word: (PRONOUN / ARTICLE), noun. In the first

case, however, the ambiguous word is a pronoun, while in the second word is an article. A general and universal implementation of the rule would create problems, while linguistically it is unacceptable to remove the ambiguity with pure statistical criteria. Overall, the lexical ambiguity is properly depicted by using two levels of description of grammar rules. In the first level, the agreement in gender, number and case of the tokens that precede and follow is indicated, while the same happens in the second level regarding the agreement of the ambiguous word.

It is very difficult to draw a precise boundary around the morphosyntactic information associated with POS tags, since it concerns morphology (e.g., verb tense), morphosyntax (e.g., noun/verb distinction), syntax (e.g., identification of the case for pronouns, accusative versus dative), and semantics (e.g., distinction between common and proper noun). Often it is represented by lexical descriptions which make explicit the way linguistic features are organized into a hierarchy and the constraints that exist between them (some features are defined only for some specific morphosyntactic categories, like the notion of tense which is restricted to the category of verbs). Here is an example of a lexical description of the word form "αμφισβητήσεις" (doubts):

[word form = "αμφισβητήσεις"
[category = noun
subcategory = common
morphology = [number = plural**, gender =**female**, case=**nominative**/**accusative**,** vocative**.**
lemma = "αμφισβητήσεις"**]]**
[category = verb
subcategory = main
morphology = [form = indicative**, tense =** present**, number =** singular**, person =** second**, tense=**future**/**past tense**, mood=**indicative**/**subjunctive
lemma = "αμφισβητώ"**]]]**

Levels of analysis run in formalism. By the level's function the application can replace phrases (e.g. entities such as names of individuals and organizations) with a VIRTUAL WORD. This analysis does not apply to the rules concerning specific context.

Thus, in cases where the input is found as both male and female (e.g. "υπουργός" minister) it is replaced by a virtual word at the first level, for which the user doesn't get informed. If that is not the case, the second level includes entries that are only female and due to their ending are treated - incorrectly - as masculine.

Levels of analysis are applied to the tagger, which is oriented especially to the removal of lexical ambiguity in Greek. For the removal of lexical ambiguity, Mnemosyne examines both previous words—up to 4 tokens—and/or the following word—up to 4 tokens. The context will determine if the word [το] is an article or a pronoun, knowledge absolutely necessary at a later level of analysis in the grammatical errors.

Template 1:
[**RULE**="TAGGER_ART_PRONOUN$_1$",
LEVEL: 1st
MORPHOLOGY = {ARTICLE, MASCULINE}] =>
[**MORPHOLOGY**={PREPOSITION}],
(
[**MORPHOLOGY**={*ADJECTIVE, MASCULINE*}
AGREEMENT={*NUMBER, CASE, GENDER*}] |
[**MORPHOLOGY**={PARTICIPLE, MASCULINE}
AGREEMENT={NUMBER, CASE, GENDER}] |
[**MORPHOLOGY**={PRONOUN, MASCULINE}
AGREEMENT={NUMBER, CASE, GENDER}]
)
\
[**MORPHOLOGY**={*ARTICLE, PRONOUN,* MASCULINE }
AGREEMENT= {*NUMBER, CASE, GENDER*}]
/
[**MORPHOLOGY**={*NOUN, MASCULINE*}

```
            AGREEMENT= {NUMBER, CASE, GENDER} ]
    ;


    [ RULE="AGREEMENT_ARTICLE-NOUN1",
     LEVEL: 2nd
    STATUS=WRONG,
    MESSAGE= "There is no agreement in this noun phrase. Replace the masculine adjective with the female."] =>
            [ MORPHOLOGY ={ARTICLE, MASCULINE}
            AGREEMENT= {NUMBER, CASE} ]
    \
            [ MORPHOLOGY={ADJECTIVE, MASCULINE}
             AGREEMENT= {NUMBER, CASE} ]

    /
            [ MORPHOLOGY={NOUN, FEMALE}
             AGREEMENT= {NUMBER, CASE} ]
    ;
    In more complex cases the level's function and the use of virtual word ensures the best description of the language.
    For example, in the first level we define the noun phrase (VIRTUAL WORD = noun phrase) and in the second level the agent
(VIRTUAL WORD = agent), which will be used in other rules (e.g. passive syntax).
    [ RULE="SIMPLE ADJECTIVE PHRASE ",
     LEVEL: 1st
    VIRTUAL WORD=SIMPLE ADJECTIVE PHRASE] =>
    \
            [ MORPHOLOGY={ADVERB}]?,
            [ MORPHOLOGY={ADVERB}]?,
            [ MORPHOLOGY={ADJECTIVE}],
            [ MORPHOLOGY={NOUN}]?
    /
    ;

    [ RULE="SIMPLE NOUN PHRASE ",
     LEVEL: 2nd
    VIRTUAL WORD=SIMPLE NOUN PHRASE] =>
    \
    (
            [ MORPHOLOGY={ARTICLE},
            AGREEMENT= {GENDER, NUMBER, CASE} ],
            [ MORPHOLOGY={NOUN},
            AGREEMENT= {GENDER, NUMBER, CASE} ]
    )                                                    |
(
            [ MORPHOLOGY={PRONOUN},
            AGREEMENT= {GENDER, NUMBER, CASE}
            MORPHOLOGY={NO A PERSON, NO B PERSON} ]
)                                   |
    [VIRTUAL WORD=SIMPLE ADJECTIVE PHRASE]
    /
    ;

    [ RULE="AGENT1",
     LEVEL: 3rd
    VIRTUAL WORD=AGENT =>
    \
            [ WORD ==("από")],
            [ VIRTUAL WORD= SIMPLE NOUN PHRASE
            MORPHOLOGY={ACCUSATIVE} ]
    /
    ;
```

We can also have more complex logical expressions as presented in the following examples:

- ([T_1], [T_2], [T_3]) define a sequence of three textspans, T_1, T_2, T_3 that must present in the input. The corresponding expression in Boolean algebra is (T_1 & T_2 & T_3) with operator '&' to represent the logical AND. All three expressions (T_1, T_2, T_3) must hold in order to match the sequence expression.
- ([T_1] | [T_2] | [T_3]) define a choice condition and the meaning is that one of T_1, T_2, T_3 must be true in order to match the expression. The corresponding Boolean expression

is (T_1 | T_2 | T_3) where operator '|' represents the OR.

- [T]{n,k} means that T can presented in input from n to k times. That is, [MORPHOLOGY={NOUN}]{2,4} means that we can have 2 to 4 nouns in a row.
- [T]? means that [T] may or may not appear. It is equivalent to write [T] {0,1}.
- [] means any word (or textual parts), no restrictions.
- The description may include complex structures such as (([T_1] | [T_2]), ([T_3] | [T_4])){1,3} meaning that a T_1 or T_2 may occur followed by a T_3 or T_4 and repeat this 1 to 3 times. Some of the sequences that the above expression recognizes are:
- T_1 T_4
- T_2 T_4 T_1 T_3
- T_1 T_3 T_1] T_3 T_1 [T_4
- The GROUP_N is displayed as a spelling of words referring to words beginning with one of 'κ', 'π', 'τ', 'ξ', 'ψ', 'α', 'ε', 'υ', ' ι ',' ο ',' η ',' ω ',' μπ ',' ντ ',' γκ 'while GROUP_WITHOUT_N for the other words (which do not belong to the GROUP_N).

Research

Question 2

As regards the second research question, it was deemed appropriate that the aforementioned proposal would be tested in a formal native speakers' setting and especially of speakers whose linguistic competence is under development. Those speakers could be primary students, who can be engaged in text production activities in order to see whether the proposed formalism and, consequently, the grammar checker deriving from it, has any effect on their linguistic competence whatsoever.

The grammatical errors were collected, classified and then created special templates through which a detailed and systematic description of each grammatical phenomenon and each category of error was made. Collecting grammatical errors through a specialized corpus is an important resource for grouping rules (Edje, 1989).

If the error categories do not include all the wrong types, it is either because the grammar parser is designed to focus on the problematic types that are the most distinctive or the most common, or because these error categories go beyond the syntax parser derived from the creation of the template-based rules. The grammar editor deals only with the grammaticality of a sentence and uses the most common (school) version of Modern Greek as a basis. Each group of templates also shows the recognizability or not of the anti-grammatical structures by the students. In this way the necessity of the tool for the classroom will be revealed.

Punctuation and spelling errors are related to the reliability of the written word but also to their convenience of reader's understanding (which may be the person or the machine), as the wrong description can affect the wider meaning of the text. Vocabulary errors affect the communicative ability of the

Deviant cases' semantic identification

Many languages are endowed with several types of easily confused words. Some of these words are homophones (homo 'same, ' phone 'sound')-words that are pronounced alike but are different in spelling, meaning, or both. Examples of common homophones include sail and sale; their, there, and they are; and knight and night (Adrian, 1988). In the Greek language words have now prevailed in a form deviating from the correct one (in grammatical production and etymology), as the verb "καταχωρώ" [input] in parallel with the correct lemma "καταχωρίζω" [register]). In other cases it seems that things are not so

clear and that different verbal variations are sometimes identical in meaning and sometimes not (e.g. "άνεργος" [unemployed] – "άεργος", [idle] "διαπραγματεύομαι" [negotiate] – "πραγματεύομαι" [deal]).

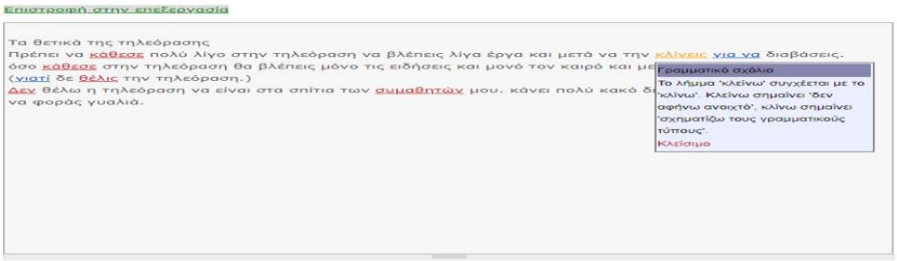
Rules of semantic identification

Description: There is a labeling that the lemma "*λήμμα*" [lemma]' is confused with "*λύμα*" [waste] and the user is informed of the meaning of both entries.

[**RULE**="SAME SOUND$_{30_3}$",
STATUS=WARNING,
MESSAGE="The lemma 'λήμμα' [lemma] is confused with 'λύμα' [waste]. Λήμμα means 'a lexicon entry', λύμα is used usually in plural and means 'waste'."] =>

\
[**LEMMA**= "*λήμμα*" | "*λύμα*"]
/

Students are asked to choose according to the note that the lemma "κλείνω'" [close] is confused with the entry "κλίνω" [incline] and they are informed about the meaning of the two lemmas (Picture 1).



Picture 1. Written text of the participating students in the environment of the Grammar Checker

Stylistic rules

A learned type is defined as either a morphological suffix that refers to a type of learned either an oral style or a word or entry with the corresponding stylistic characterization. The grammar checker manages these types in many levels. At the initial level, the learned phrases (~ 359) that are still in use in the written or spoken speech have been collected. The management of learned phrases is controlled by another template that points out to the user the misspelling of the standard spoken phrase he uses. This pattern is completed by highlighting learned or oral symphonic complexes. The absence of accentuated increase in verbs is an element of oral speech and is pointed out, as well as the misspelling of abbreviations.

Certain morphological types with archaic or oral characterization are acceptable in certain consolidated expressions but not acceptable when used individually. This functionality is addressed by operating at the levels supported by Mnemosyne software. So in the first level, when the specific type is found in a context (learned phrase) it is replaced by a virtual word (VTEXT) that is not visible to the user, while in the second level, if the entry is not in the specific context, the corresponding information is pointed out to the user.

Rules of stylistic information

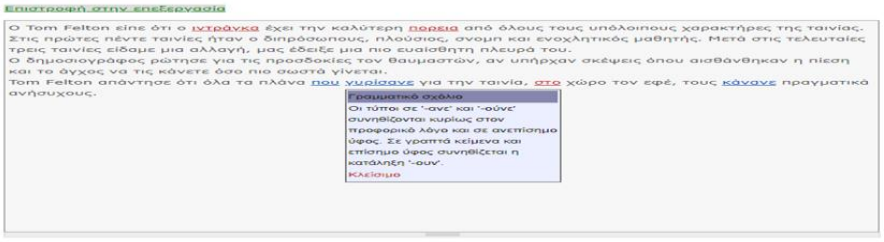
Description: In the 1st level the word '*αδείας*' [license] in the corresponding context (after the

entry ‘επίδομα’ [bonus] or ‘φύλλο’ [sheet] will be replaced by the VIRTUAL WORD = "__ LEARNED PHRASE__"], an element that will not be visible to the user. In another context, described in the second level, the user is informed that the stylistic characterization of these words refers to learned use and it is at his discretion to replace this word with another of the modern common Greek or to preserve it.

```
[ RULE="Stylistic Rules1",
 VIRTUAL WORD="__LEARNED PHRSE__"] =>
1st level
      (
      [ LEMMA=("επίδομα"  |  "φύλλο") ]
      ),
      \
                  [ WORD="αδείας"]
      /
      ;
 [ RULE="Stylistic Rules1",
 STATUS=INFO
 MESSAGE= "The word is used in learned speech. Stress in antepenult".] =>.

 \
      [WORD=("αδειας" | "ακριβειας")
      /
      ;
```

Students are asked to choose according to the note that the suffix “*ανε*” is used in oral speech (Picture 2).



Picture 2. Written text of the participating students in the environment of the Grammar Checker

Rules for stress, spelling marks and punctuation

Punctuation is the sign on writing which makes the meaning of sentences become clear. Punctuation is a standard set of marks which are used in written speech to clarify the meaning and to separate sentences, words, and parts of words (Asayeh & Kumar, 2016).

Additionally, the incorrect absence or presence of stress is described in this template. It also describes instances where there is a wrong stress position. More specifically, it manages monosyllabic types that are stressed in a special context.

More complex is the management of the comma by the grammar checker. At the first level, the dependent sentences that function as adverb phrases "require" a comma before their conjunction. On the contrary, the dependent sentences that function as an object or as a subject do not "require" it.

Also, this template manages cases of very common use of abbreviations (e.g. *κ.λ.π*) that have an incorrect presence of a dot (eg: *κ.λ.π.* instead of the correct: *κ.λπ.*). Finally, this template describes the cases of misprinting (the two spaces, the four dots instead of three, the presence of a question mark, parentheses or comma before or after the dot, etc.).

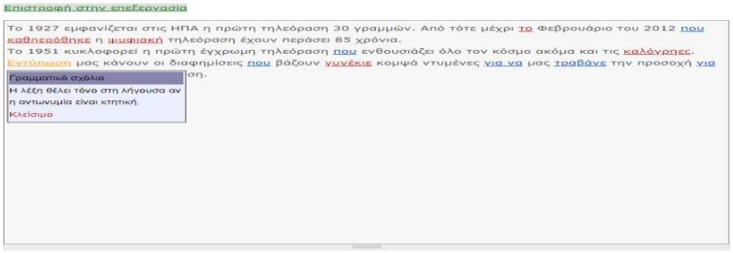
Rules for stress

Description: When a noun or adjective or participle in a passive voice is stressed in the

antepenult and then followed by one of the feeble types of the personal pronoun in continuous sequence (without stress), the user is informed that the stress is incorrect and that the word expects an enclitic one in the suffix.

```
[ RULE="STRESS RULE1",
STATUS= WRONG
MESSAGE= "
The word wants an enclitic stress in the last syllable."] =>
\
        [ STRESS={ANTEPENULT},
        MORPHOLOGY={NOUN}] |
        [ STRESS={ANTEPENULT},
         MORPHOLOGY={ADJECTIVE}] |
        [ STRESS={ANTEPENULT},
        MORPHOLOGY={PARTICIPLE,PASSIVE VOICE }]
/
        [ WORD= "μου" | "σου" | "του" | "της" | "μας" | "σας" | "τους"],
        [ WORD= "μου" | "σου" | "του" | "της" | "μας" | "σας" | "τους"]]
;
```

Students are asked to choose according to the note that the word "*εντύπωση*" (impression) needs an ecliptic stress if the next pronoun is possessive (Picture 3):



Picture 3. Written text of the participating students in the environment of the Grammar Checker

Rules for agreement

Agreement is among the most widely researched issues in theoretical linguistics. With regard to the features relevant for agreement, most of the typological literature has focused on person (Cysouw, 2003), but agreement may also involve gender, number, case, and definiteness. The two major agreement domains are the noun phrase, and the clause in this template, issues of agreement are classified in the verb phrase or in the noun phrase. In oral speech, however, many nouns are essentially used with the wrong choice of the gender (e.g. "***ο*** *ψήφος*" [the vote] instead of the correct one: "***η*** *ψήφος*"). In this way the wrong gender is generalized to all nouns with a similar suffix. The error, however, can be extended to the predicate or to possible aggressive definitions to identify this noun: [e.g. "*οι μέθοδοι είναι καλ****οί***" [the methods are good] instead of right: "*οι μέθοδοι είναι καλές*", or "*αντ****οί*** *οι μέθοδοι*" [these methods] instead of right: "*αντ****ές*** *οι μέθοδοι*

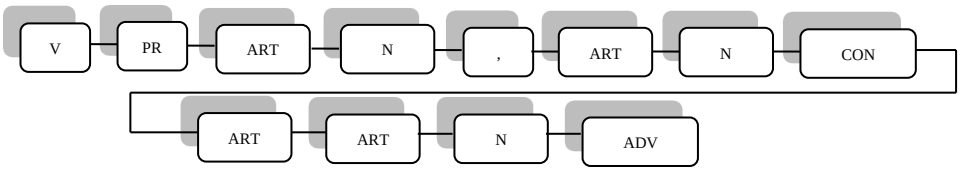
Participles that are used as an aggressive definition in female nouns also have a wrong statement (e.g., "*ισχύοντες διατάξεις*" [current provisions] instead of the right: "*ισχύ****ουσες*** *διατάξεις*").

A product of confusion, due to the similar property, is the phrases that the adjective has incorrect spelling when it identifies specific nouns. Consequently, we have the configuration of the wrong noun sets "*ψ****η****λή κυριότητα*" [high ownership] or "*υψ****η****λή κυριότητα*" instead of the correct: *ψ****ι****λή κυριότητα*.

In the category of this phrase errors that are observed in broader noun sets are also

described: a) with the adjective "*πολύς*" [very], b) in the subject agreement with the predicate, c) in the (semantically) wrong choice of adjective (e.g. "***ραγδαία*** *βελτίωση*" [rapid improvement] instead of: "***θεαματική*** *βελτίωση*" [spectacular improvement]), d) the use of adjectives (instead of adverbs) in cases that identify noun sets (e.g. "*τόσ**οι** πολλοί*" [so many] instead of: "*τόσ**ο** πολλοί*").

In addition, this template includes prepositional phrases that are connected in a row or there is an incoherent shape. Therefore, in the phrase: "*πήγαμε στην πόλη, το χωριό και το γήπεδο σήμερα*" [we went to the city, the village and the stadium today.], the grammar points out the incorrect use of the articles "*το*" [the] and "*το*" [the] and suggests the use of preposition and article "*στο*" [in]. The grammar points out the incorrect use of the articles [the] and [the] and suggests the use of intention and article [in]. The same syntactic analysis, however, appears in the phrase: "*πήγαμε στην πόλη, τη Δευτέρα και την Τετάρτη κάποτε*" [we went to the city, once on Monday and Wednesday]. In both sentences the analysis has the form (Picture 4):



Picture 4. Syntactic analysis of nphrase

In the second sentence, however, the rule described does not apply. Therefore - in another set of rules and in a second level - the nominal set "*την Τετάρτη*" [on Wednesday] -as well as all the dates- is recognized as VIRTUAL WORD = DATE. In this way the rule does not apply to phrases with the same syntactic analysis but with the presence of VIRTUAL WORD = DATE.

This category manages the subject verb agreement only in cases of confusion of the suffix of third singular of the passive voice of the definite present tense "*-ται*" or "*-αται*" and the suffix of the second plural of the active voice of the definite present tense "*-έτε*" ή "*-άτε*".

The description of the rules concerning agreement

[**RULE**="AGREEMENT ARTICLE ADJECTIVE$_1$",
STATUS=WRONG
MESSAGE=" There is no agreement on this noun phrase. Replace the adjective with the corresponding female ending '-ειών."] =>]
=>

\
[**ENDING NOMINATIVE SINGULAR**={*ύς*},
MORPHOLOGY={*ADJECTIVE*},
ENDING={*έων*},
MORPOLOGY={*ADJECTIVE, GENITIVE, PLURAL, MASCULINE]*}

/
[**MORPHOLOGY**={*GENITIVE, FEMALE, PLURAL*}]
;

Discussion

The goal of the current study was to propose an error standardization method based on context-sensitive grammars and consequently test its feasibility and practicality with native speakers (primary students).

As regards the first research question (can a context-sensitive grammar constitute the basis for an effective error standardization method within the framework of an online

grammar checker), it is clear that a context sensitive grammar addresses the problem of correcting spelling mistakes that result in legitimate words. We compared the templates of the Greek grammar checker with the templates of other grammar checker and we noticed that it includes their templates (Gakis et al., 2015, 2016). It is a complete and complicated natural language processing system used for information retrieval and information extraction in free text. The Greek Grammar Checker definitely has advantages for the teaching of mother tongue, mainly concerning its modernization (not focus on drills and text reviews by hand). Greek Grammar Checker offers a modern aspect of language teaching and contributes to a more solid linguistic understanding (Kokkinos et al., 2020).

As far as the second research question is concerned (can the grammar checker be practical for native speakers), the implementation of the grammar checker that was developed through the proposed formalism and templates was investigated. Native speakers with a developing linguistic acquisition were chosen; therefore primary students were the participants of the current study. It was found that the proposed formalism, the templates and, consequently, the grammar checker were overall functional as it detected most of the participants' usual errors. The grammar checker offers native speakers a "cohesive" environment where text segments can be checked for grammar errors. It can also contribute to students' linguistic competence and within the framework of more complicated teaching approaches e.g. differentiated language teaching (Kokkinos et al., 2020). Additionally, the grammar checker does not correct the errors, but reminds to the user the grammar rule and suggests a possible revision. This element is particularly significant as it can contribute not only to the users' grammar competence but also to their critical literacy skills through the process of reviewing and decision making.

The originality of the current study can be traced in: a) research in finding common (frequent) grammar errors in Modern Greek texts, b) description of the erroneous grammatical usages with a high-level formalism (Kanon) by linguist experts without computer programming knowledge and c) Automatic production of full featured grammar checker. The description of rules and templates of the Greek Grammar Checker in a generic way so that it can constitute a framework for describing grammatical errors and developing grammar checking software. Additionally, the educational impact of this whole process can be seen in the data collected in class as it enhances students' grammatical competence and supports the "text as a process" framework (Yang, 2010).

Limitations of the Study

The construction of the grammar checker for the Modern Greek language is the first collection and coding effort of errors that occur in the specific spoken and written language. The effective software evaluation was with the parallel correction of the same texts by the grammar checker and a human. The human correctors were four philologists, teachers in high schools, with great experience in text correction. More than 100 texts were given for correction to the grammar and human checker.

In a very large percentage, the grammar checker approximates the correction of a human, because the electronic environment of Mnemosyne is closer to the human way of thinking and the natural writing process (Daiute, 1985).

Differentiation between grammar checker and human corrector is noticed in cases referring to the conceptual field which is not described in the grammar checker templates. The human corrector handled all the foreign words found in texts. Grammar checker doesn't describe these cases since the electronic lexicon manages only Greek words.

Acknowledgments

The authors wish to thank Ms. Vassiliki Mavreli for her contribution to the digitization of students' texts utilized in the current study.

References

Adrian, R. (1988). *Dictionary of confusing words and meanings*. Dorset Press.

Ancker, W. (2000). Errors and corrective feedback: Updated theory and classroom practice. *English Teaching Forum, 38*(4), 20-24.

Asayeh, W. & Kumar, P. (2016). A study on the difficulties faced by Libyan university students in using punctuation marks in English writing. *English Language and Literature,4*(3), 60-63, https://doi.org/10.33329/rjelal.

Bartholomae, D. (1980). The *study of error. College Composition and Communication, 31*(3), 253-269.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Corder, S. (1974). *Error Analysis*. In J. P. B. Allen & S. P. Corder (Eds.), *Techniques in Applied Linguistics*. Oxford University Press, 363-366.

Cysouw, M. (2003). *The paradigmatic structure of person marking*. Oxford University Press.

Daiute, C. (1985). *Writing and computers*. Addison-Wesley.

Dobrin, N. (1990). A new grammar checker. *Computers and the Humanities, 24*(1/2), 67-80.

Edje, J. (1989). *Mistakes and correction*. Longman.

Ellis, R. (1994). *The study of second language acquisition.* Oxford University Press.

Gakis, P., Panagiotakopoulos, C., Sgarbas, K., Tsalidis, C. & Verykios, V. (2016). Design and construction of the Greek grammar checker. *Digital Scholarship in the Humanities*, 554-576. http://dx.doi.org/10.1093/llc/fqw025.

Gakis, P., Panagiotakopoulos, C., Sgarbas, K., Tsalidis, X. & Verykios, V. (2015, September 20-24). The construction of a Greek Grammar Checker through Mnemosyne. Proceedings of 17th *International Conference on Speech and Computer, SPECOM, 2015*, Athens.

Genthial D. & J. Courtin, (1992, August 23-28). Prom Detection/Correction to Computer Aided Writing. *Proceedings of the 15th International Conference on Computational Linguistics* (COLING- 92), Kyoto.

Holton, D., Mackridge, P., & Filippaki-Warburton, E. (2002). *Grammar of the Greek Language*. Patakis.

Iordanidou, A. (1999). *The guidebook of Modern Greek*. Patakis.

Iordanidou, A. (2004). *The guidebook of Modern Greek* (2nd part). Patakis.

James, C. (1998). *Errors in language learning and use: Exploring error analysis*. Routledge.

Kokkinos, T., Gakis, P., Iordanidou, A., & Tsalidis, C. (2020, February 11-13). Utilizing grammar checking software within the framework of differentiated language teaching. In the *Proceedings of the 9th International Conference on Educational and Information Technology (ICEIT 2020),* St Anne's College, University of Oxford, United Kingdom.

McCord, M. (1987). *Natural language processing in Prolog.* In A. Walker, M. McCord, J. F. Sowa, & W. G. Wilson (Eds.), *Knowledge Systems and Prolog: A logical approach to expert systems and natural language processing* (pp. 94-99). Addison-Wesley.

Seiffert, R. (1987). *Chart-Parsing of Unification-Based Grammars withID/LP-Rules* (LILOG-REPORT 22). IBM.

Yang, X. (2010). *Modelling text as process*. Continuum.