# The *Greek High School Dictionary*: Description and issues

Maria Gavrilidou
Voula Giouli
Penny Labropoulou
Institute for Language and Speech Processing

*This paper reports on the compilation of a monolingual Greek pedagogical dictionary targeted at young native language learners, namely secondary education students, aged between 12 and 15. The dictionary, which is in printed form, has been designed to be used in the classroom as a supporting tool for language learning, but also as reference work tailored to meet students' needs for language understanding and production both at school and in everyday activities outside school. To this end, considerations on user-friendliness have been accounted for, and the design and implementation of the dictionary content have built primarily on the needs and requirements of schoolchildren pertaining to the specific age group. The dictionary comprises 15,000 lemmas covering general language vocabulary along with terms belonging to subjects taught at the specific level of education. Information that is central to the pedagogical targets of language learning has been encoded for each lemma, i.e., part of speech, morphology-difficult inflectional forms, domain, register, definitions, usage examples, etc. Finally, useful comments focus on interesting aspects of certain words' semantics, usage, register etc. The central feature of the dictionary is the headword organization which employs systematically word formation criteria: derivatives by suffixation are organized in word-families, while prefixes are included in the dictionary as independent headwords accompanied with lists of derivatives or compounds on the basis of derivational and semantic criteria. The paper presents the framework of the project and its specifications, discusses the main methodological principles that underlie its construction and elaborates on the dictionary description, the main problems faced and the solutions adopted in the process of its compilation.*

## 1. Introduction

This paper is concerned with design and implementation issues relevant to the compilation of a monolingual native language learners' dictionary, namely the *Greek High School Dictionary* (GHSD).

The GHSD is the third in a series of dictionaries created in the framework of a national project aiming at the reformation of school curricula and the construction of new teaching material. The series consists of two dictionaries for the primary level and one (the GHSD) for children in the fist three classes of secondary education (12 – 15 yrs old). The aim of the GHSD, as prescribed by the Ministry of Education, is to serve as: (1) a teaching aid in language learning in the classroom, (2) a reference work tailored to meet children's needs for language understanding and production both at school and in everyday activities outside school, (3) a tool for teaching dictionary skills to students by acquainting them with lexicographic meta-language and dictionary usage methods. The GHSD intends to improve the active as well as passive vocabulary of students (active vs. receptive use of the dictionary).

## 2. Current trends in Greek pedagogical lexicography

The GHSD has come to fill an attested gap in Greek lexicography. The last two decades witness an important lexicographical activity, which, however, focuses mainly on large-scale dictionaries for "general use by the average speaker of Greek", i.e. mainly adult native speakers. Although these dictionaries build upon modern lexicographic practices, the load of information they bear and their corresponding size is discouraging for young learners or

schoolchildren. Pedagogical lexicography,[1] however, seems not to have followed the general trend: very few dictionaries exist in Greek for schoolchildren, and these are mainly targeted to pre-school and primary education. Moreover, most of the dictionaries supposedly targeted to younger ages are simplified versions of the existing large-scale dictionaries.

## 3. Dictionary specifications

The specifications set by the Ministry of Education concern macrostructure, microstructure and presentation format. As regards *macrostructure,* the GHSD should contain 15,000 lemmas covering general language vocabulary as well as the technical vocabulary of the school subjects. As regards *microstructure,* information coded for each lemma should cover spelling, part of speech, morphology (inflection), syntax, semantics (senses, synonyms/antonyms), etymology, usage examples, domain, style and register information. As regards *presentation format*, the GHSD will appear only in printed format while specific stylistic details as to fonts, columns etc. were also specified.

Furthermore, the specifications recommended that the whole process of dictionary compilation be corpus-based; this refers to headword selection (in order to identify the appropriate age-graded vocabulary), syntax and senses selection and distinction, and collocations and usage examples extraction. Schoolbooks of subjects taught at the specific level of education were indicated as additional sources for the extraction of the appropriate technical vocabulary.

Finally, the specifications stressed the importance of user-friendliness as a basic feature of the dictionary, since it addresses children's needs: the guidelines recommended the use of special typesetting conventions and simple meta-language.

In the following sections the implementation of these basic guidelines for the compilation of the GHSD will be presented.

## 4. Methodological principles

In accordance to the specifications, the main methodological principles that have governed the GHSD production were:

(1) *user orientation*: the needs, requirements and interests of prospective users as well as their relevant linguistic competence have greatly influenced the decisions as regards dictionary content and presentation;

(2) *corpus use*: a corpus-based approach was adopted for the compilation of the GHSD, i.e. not only for headword selection, but also for sense distinction and selection, collocates extraction, usage examples selection etc. (cf. Sinclair 1991, Bullon 2006). Ideally, given the purpose of the dictionary, the corpus should reflect the language and interests of the target group. However, since there is no such corpus for Greek, we opted for a hybrid approach, whereby exploitation of a general language corpus was coupled with human processing.

The corpus used is the Hellenic National Corpus (HNC, *http://hnc.ilsp.gr*, Gavrilidou 2002), which is a general language corpus of contemporary Greek (currently 47 m. words), comprising written texts of a broad range of text types / genres and topics from various sources and representing the current use of standard modern Greek. The texts composing the HNC material have been structurally annotated according to the Corpus Encoding Standard (CES, EAGLES 2000), which incorporates the relevant EAGLES guidelines (1996). Texts are annotated with bibliographic information (author, publisher, publication date, etc.) and classified as regards the parameters of Medium, Genre and Topic. Through the web interface,

[1] The term "pedagogical lexicography" is taken here in the broader sense, covering not only foreign language learners' dictionaries, as is often the case in English tradition (e.g. Rundell 1998, Bullon 2006), but also native language learners' dictionaries and school dictionaries (e.g. Dolezal and McCreary 1999).

the corpus can be queried for wordforms, lemmas[2] and morphosyntactic tags or any combination thereof; results come in the form of concordances and statistics (frequency information). Moreover, users can search the whole corpus or define a sub-corpus based on the classification and annotation parameters accompanying each text, thus creating sub-corpora of a specific author, or belonging to a specific genre, text type, domain etc.

In the process of the construction of the GHSD, sub-corpora were defined through the interface, which were more appropriate for the target age (e.g. by excluding very official text types or highly specialized scientific texts). The sub-corpora were subsequently used for all stages of dictionary compilation, so that headwords, collocations and expressions included, as well as the information provided for them (definitions, usage examples, etc.) were of a level suitable for children.

(3) *use of additional resources*: complementary to the HNC, a corpus of schoolbooks used in secondary education was compiled. This was used for the extraction of terminology used in the school subjects, as well as for the formulation of their definitions.

For the compilation of the headword list, other dictionaries were also consulted (both Greek and foreign). Consultation of foreign dictionaries was necessitated by the fact that there exist no Greek dictionaries for this age; obviously, the focus was on foreign *children's* dictionaries. The investigation of these resources revealed headwords to be included in the GHSD, contributing, thus, to the completeness of the headword list.

Last but not least, Greek grammar books were consulted; these provided coverage of the function words (prepositions, conjunctions etc.) to be included in the GHSD. As the dictionary is meant to be used in school, it is imperative that the linguistic analysis of these specific classes proposed by the dictionary is in accordance with the one found in the children's grammar books.

## 5. Macrostructure

The term *macrostructure* is used in the meta-lexicographical literature in (broadly) two ways: minimally, as being practically synonymous to the lemma list (e.g. Hartmann and James 1998) or maximally, as covering the overall organization of the dictionary material. According to Hartmann and James (1998), the macrostructure is supplemented by *outside matter*: front (preface, users' guide etc.), middle (illustrations etc.) and back (abbreviations, lists of names, weights and measures etc.). The totality of all these constitute the *megastructure*. Others (e.g. Nielsen (1994)) use the term *macrostructure* as equivalent to megastructure, in the sense that it pertains to the dictionary as a whole.

In this section, entitled *Macrostructure*, the following issues will be discussed: headword selection and organization, types of headwords, index and appendices, in the sense that all these structures, although different in format, simply constitute different types of lemma organization.

### 5.1. *Headword selection*

Dictionary macrostructure has been designed and implemented on the basis of general assumptions about the reference needs and reference skills of the target users. To this end, a number of communicative situations and academic settings were primarily identified; these range from reading comprehension of a variety of texts pertaining to different genres (determined by the curriculum) to oral and writing activities children usually carry out, in and outside school encompassing both language comprehension and production. To this end, the GHSD comprises *15,000* lemmas aimed at covering efficiently a basic vocabulary, i.e. one

[2] The terms *lemma* and *headword* (abstract form grouping together all inflected forms of a word) are used interchangeably in this paper.

that is deemed adequate for secondary education students in their everyday interaction and effective communication.

The adopted headword selection methodology has been a hybrid one, combining corpus statistics with human processing. Lemma selection criteria were: (a) general language corpus frequency, (b) presence in the teaching material, and (c) comprehension and/or production difficulty level.

Initial word lists were derived automatically on the basis of frequency of occurrence in the HNC. Subsequently, candidate lemmas were further processed manually in order to adapt the lemma list to the needs of the specific target group. The processing consisted of two tasks: filtering and enrichment. *Filtering* aimed at removing lemmas not appropriate for the specific age (belonging to highly specialized domains, being very formal or taboo words, etc.), while *enrichment* catered for the inclusion of lemmas necessary for the target group but absent from the HNC list, such as the terminology extracted from the corpus of schoolbooks. Finally, existing dictionaries and grammar books have been consulted, to ensure vocabulary completeness.

### 5.2. *Headword organization*

The GHSD has been conceived as a consulting aid as well as a learning material integrated in the school environment and process; i. e. it will be used in the school classroom as an aid for *teaching students active vocabulary acquisition methods*.

Graves et al. (2004) mention three strategies that help students become independent word learners: using context cues, using word parts and using the dictionary. In designing the GHSD, we have decided to address the second strategy by employing a number of mechanisms that help users familiarize themselves with word production methods. In this way, students will learn how to *comprehend* the meaning of derivative words if they know the base lemma or, at least, understand the meaning of specific word parts, namely prefixes and suffixes. Moreover, they will be better equipped to *produce* words by combining base forms and the appropriate affixes.

For this purpose, we have adopted a different organization of the vocabulary than the alphabetical ordering usually employed in printed dictionaries. In the GHSD, lemmas are grouped together on the basis of the morphological and semantic relation that holds between them; two distinctive types of lexicographical units, namely *word families* and *prefixes connected with lists of compounds and/or derivatives*, have been chosen to represent the two mechanisms of word formation, derivation and synthesis.

More specifically:

(1) *Word families*: The central lexicographic unit of the GHSD is based on the concept of *word family*, i.e. each lexicographical article consists of a *main lemma* and a number of *sub-lemmas*. Lemmas in the word family are connected to each other by derivation processes while preserving the "nucleus" of the lexical meaning; this means that both *morphological* and *semantic relevance* among the members of the family have to be attested. Each word family is restricted to close members and includes only derivatives by suffixation, in order to avoid generating a large and complicated article.

In most of the cases, the main lemma is the *base form* and the ordering follows the derivation process which resulted in their production. For instance, under the main lemma *αθώος* [innocent], we find the sub-lemmas (in that order): *αθώα* [innocently], *αθωότητα* [innocence], *αθωώνω* [acquit], *αθώωση* [acquittal], *αθωωτικός* [exonerative]; in this case, the ordering of the lemmas is: <adjective> → <deadjectival adverb> / <deadjectival noun> / <deadjectival verb> → <deverbal noun> → <denominal adjective>.

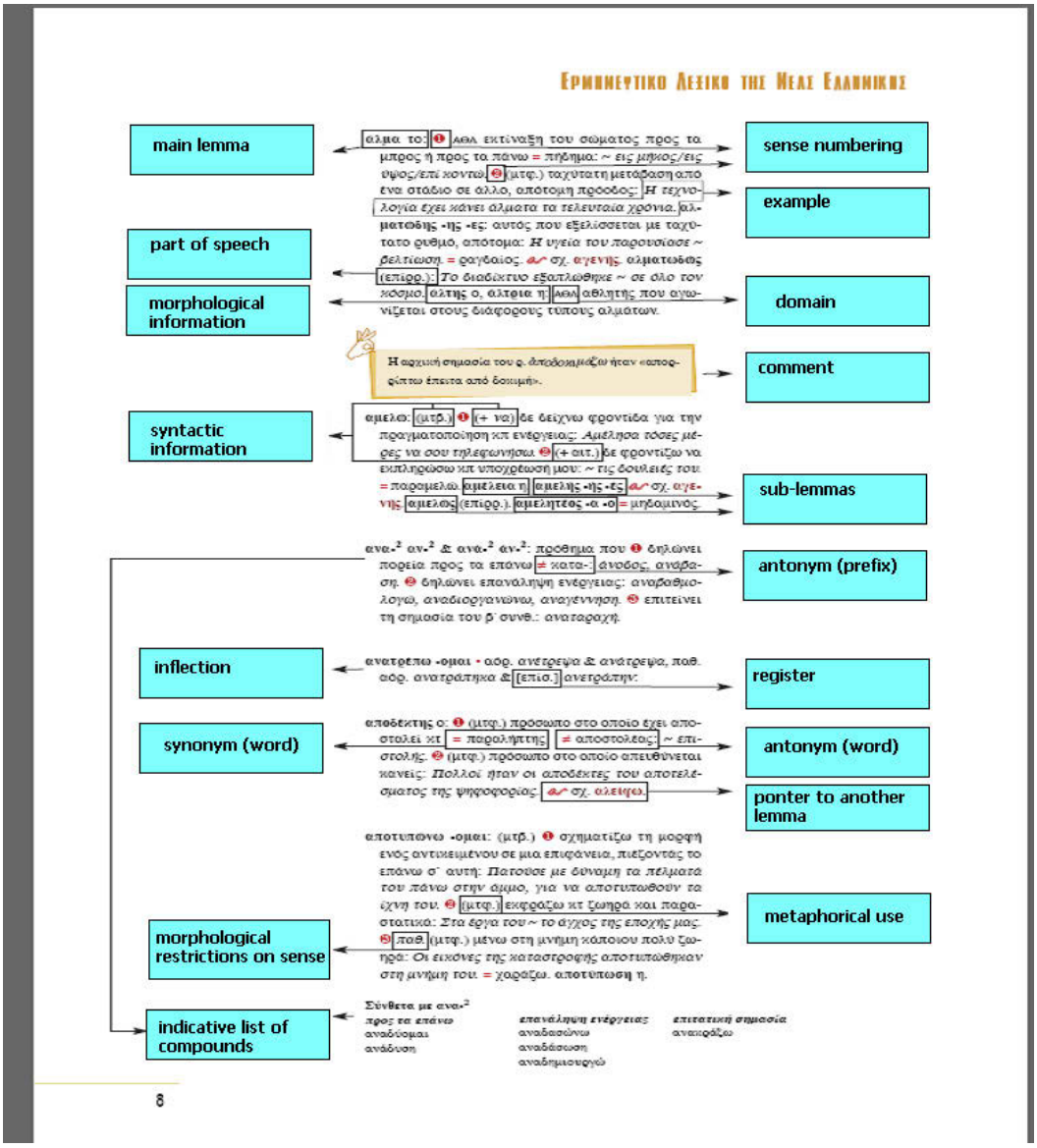


Figure 1: User's guide to the GHSD lemma structure

However, when deemed necessary, the *"semantically simpler" word* (cf. Mel'čuk et al. (1995: 80)), i.e. the word whose content is easier to define in a user-friendly way and which can then be used for defining subsequent members of the family, is chosen as the main lemma; for instance, *δημοσιογραφία* [journalism] is the main lemma in the relevant article because it can easily be defined independently of the word *δημοσιογράφος* [journalist] which is then defined as "person professionally occupied in journalism".

(2) *Prefixes and lists of compounds*: In order to familiarize students with the mechanisms of derivation and synthesis by prefixation, a different kind of grouping has been chosen: prefixes (mainly grammatical but also lexical morphemes which are very productive in compounding) have been included in the GHSD as headwords; a subset of derived or compound words formed from these prefixes are listed at the bottom part of the same page, grouped according to the senses of the prefix (see Figure 1). No further information (word senses, examples etc.) is provided for these words, which have been carefully selected so that their meaning is easily derivable from its parts; thus, they serve as examples upon which the students will learn how to analyze words in the appropriate parts.

The above described types of vocabulary organization are not new to dictionaries; ten Hacken et al. (2006) discuss the various ways dictionaries employ headword organization to represent word formation mechanisms and their respective advantages and disadvantages. In general, the approaches are: (1) strict alphabetical ordering of all words, which hinders the user from grasping the relation between the various derivatives; (2) inclusion of a small subset of derivatives as run-on entries (usually adverbs, diminutives and some deverbal nouns); (3) inclusion of all derivatives under the base lemma, which makes the task of looking up a derivative word (especially words with prefixes) more difficult, since users must be able to analyze it correctly and recognize the base lemma. In addition, the inclusion of affixes as independent headwords is also employed in various dictionaries; however, it is questionable whether users look up affixes (especially suffixes) instead of looking up the whole word.

The approach we have implemented in the GHSD tries to combine these features in order to maximize their advantages and minimize their drawbacks: derivatives by suffixation are grouped under the base lemma and derivatives by prefixation under the prefix (in cases of highly productive prefixes). In this way, the semantic and morphological relation is highlighted and, at the same time, derivatives are found close to where they would normally appear if strict alphabetical ordering was adopted, so that their lookup is facilitated.

### 5.3. *The alphabetical index*

The described lemma organization contributes also to a much desired lexicographical goal, namely economy in space (see also sections 6 and 8). However, it also poses a burden on users, who find it difficult to spot the word they are looking for and might be, therefore, discouraged from further use of the dictionary. For this reason, the *alphabetical index* has been introduced to the dictionary: it appears on the side of each page (Figure 2), includes all the lemmas *and* sub-lemmas in alphabetical order, and directs the user to the appropriate lemma where these are located. It also contains difficult inflected words (distinctively marked), common misspellings and spelling variants, in order to facilitate dictionary lookup.


Figure 2: Sample page of the GHSD

### 5.4. *The appendix*

Finally, as regards the macrostructure, an *appendix* is included in the GHSD which contains words that belong to the so-called closed sets (names of months, days, measures etc.), person names, names of geographical entities, numerals etc. These entries are deemed useful for schoolchildren, yet they do not require proper definition or they can all be defined according to the same formula (e.g. numerals).

## 6. Microstructure

In designing the GHSD, we have tried to cover all types of information required by the original specifications in a way that would accommodate the target students' needs and knowledge. However, given the target user group, the microstructure had to be as informative but as light as possible.

For this reason, not all lemmas include all types of information; otherwise, the dictionary would become overburdened, difficult to use, unfriendly and the inclusion of the required number of lemmas in the specified volume of pages would not be feasible. Each specific information type is included for each headword (main or sub-lemma) only where deemed necessary and appropriate. Moreover, the word family structure has been exploited in order to achieve economy in entry compilation: for instance, *sense* information is detailed only for the main lemma and this knowledge is "inherited" to the sub-lemmas, unless new or different senses are attested: this means that the sub-lemmas either include no definition at all or their definitions include the main lemma in order to be shorter and easier to comprehend.

As regards the form of the headword, the *basic form* may be followed by spelling and morphological *variants* as well as older forms still in use in modern Greek; when appropriate, additional information (style, register and/or sense restrictions) is encoded for variants.

As regards the *grammatical information* encoded for each headword, the school grammar and language books have been extensively used to ensure compatibility of terminology but also to avoid duplication of information: the GHSD comes to supplement these books, by specifying the appropriate information for each headword in as far as this cannot be inferred from them. Thus, *inflectional* information is limited to irregular forms and codes for the absence of specific types (e.g. plural, particular verb tenses etc.), given that regular inflection is taught at grammar. *Part of speech* information is either implicitly (e.g. through the presence of the definite article for nouns) or explicitly encoded for all headwords. *Syntactic information* (complementation) is supplied mainly for verbs through the use of the meta-language used in the grammar (*transitive/instransitive* distinction); where deemed necessary, additional information is provided for some complements in the form of the required morphosyntactic features they are realized with (e.g. sentential complement introduced with *που* [that]).

*Additional information* is provided for headwords and/or word senses, where appropriate: *domain* codes are mainly used for the technical vocabulary extracted from the special subject schoolbooks; *register* (style) codes have been selected so that students can easily infer from them the appropriate usage of particular words or senses (e.g. in formal/informal settings, slang etc.).

The main information attributed to *word senses* is of course the *definition*: the linguistic structure of the definitions is kept as simple as possible while the vocabulary used is semi-controlled, in the sense that the *definiens* does not contain words more difficult than the *definiendum*; cyclical definitions and definitions through synonyms are excluded, as well as through words possibly unknown to the target users.

*Usage examples* have been adapted from the HNC concordances in order to save space and make them more suitable to students' needs and interests.

*Synonyms* and *antonyms* are included for word senses and/or usage examples (where most appropriate) in order to help students enrich their vocabulary.

*Morphosyntactic constraints* and *selectional restrictions on complements* are encoded when

they characterize specific word senses.

Finally, *etymological information* appears in the GHSD in two forms: derivation is obviously inferable from the above described vocabulary organization. In addition, for specific lemmas presenting etymological interest further information has been encoded, hinting at (but not insisting on) various language evolution mechanisms (loans, calques, diachronic evolution, etc.); this information is given through the mechanism of *comments*.

*Comments* constitute an important feature of the GHSD; they are used for selected entries (placed below the relevant article and referring to the whole word family) in order to provide additional information on a range of phenomena regarding morphology (e.g. irregular formation of comparatives, etc.), syntactic patterns or lexical semantics, etymology etc. (see Figure 1). More specifically, usage notes have been elaborated for lemmas or sub-lemmas with senses that are difficult for children to discriminate (e.g. the meaning of the lemma *αιτιολογώ* [give the reasons for something] vs. *δικαιολογώ* [justify]) or for commonly confused words, such as synonyms or near synonyms or simply words similar in meaning and/or in form (e.g. *ζήλια* [jealousy] vs. *φθόνος* [envy], or *ζευγάρι* vs. *ζεύγος* [couple]). The communicative situations these words are used and their different uses and/or registers are explained where applicable.

Information on the etymology of lemmas designating the origins of a certain word has been provided with the focus being on frequent, very productive words or on words with interesting history. To this end, compounding or synthesis as a word formation mechanism is being exploited. For example, the article of the lemma *έργο* [labour, work, task, project or undertaking], which is a highly productive word in modern Greek, is accompanied by a comment on (a) the word's origin from the ancient Greek language, and (b) an extensive (yet not exhaustive) list of compounds related to it. The advantage is two-fold: word formation is being exemplified, and, at the same time, a number or words that were not included in the headword list due to space limitations are also cited. It remains then to the teacher to show students how meaning can be worked out for words lacking proper inclusion in the dictionary on the basis of the meaning of their parts. Word histories for lemmas that have survived from previous phases of the Greek language or that have been imported from foreign languages either with the same or with new senses have also been recorded in the comments section.

Finally, remarks on frequent mistakes (misspellings, syntactic errors, erroneously inflected forms, etc.) have also been provided for.

## 7. Sense discrimination

The problem of discriminating the different meanings of a polysemous word is common to all lexicographic projects. Dictionaries aimed at students, however, pose another restriction, that is, avoiding information overload that would distract students from the quest at hand. Moreover, unlike dictionaries for adults (either native speakers or learners), where the listing of meanings is exhaustive and the dictionary value lies upon the inclusion of new or highly technical senses, children's dictionaries need to maintain a simple and clear to follow content. We therefore, have avoided listing too many senses per lemma opting for a *less granular approach* rather than a fine-grained one. Core senses have been identified as a result of grouping sub-senses together, yet without collapsing together meanings that cannot be described in a uniform albeit simple and clear way.

Moreover, based on the assumption that children should not be overloaded with information that is inappropriate or beyond their interests, senses and usages that are deemed sparse, dated or which belong to a highly technical domain (beyond subjects taught at school) have been omitted.

Sense discrimination and inclusion have been performed on the basis of corpus evidence. Similarly, the ordering of senses has been performed on the basis of frequencies attested in the corpus rather than on historical or etymological criteria.

## 8. Presentation format

Given (a) the target users' age, (b) the requirement for user-friendliness but also (c) the need for economy of dictionary space, specific stylistic choices and typesetting conventions have been exploited in the GHSD, in order to render the material more attractive and easier to follow by the prospective user (Figure 2).

These include:

- introduction of the alphabetical index to facilitate lookup;
- the use of different typesetting styles for different types of information so that they become prominent inside the word family article (eg. boldface for the main lemma and sub-lemmas);
- use of different formatting of the comments to attract users' attention;
- layout techniques (e.g. usage of the right and bottom margins) to provide information and gain space at the same time;
- usage of easy to understand symbols instead of lexicographical meta-language (e.g. the symbol "=" to identify synonyms, tiny glasses in place of "cf.", etc.).

## 9. Conclusions

We have hereby presented a monolingual Greek dictionary for secondary education students tailored to not only meet young native learners' needs and requirements but also to serve as an aid in the language teaching context. The GHSD is informative and user-friendly and is intended to serve as an intermediate level between children's dictionaries and dictionaries for adults and act as a means for teaching reference skills.

## References

Atkins, B. T. S. (ed.) (1998). Using dictionaries: Studies of dictionary use by language learners and translators. Tübingen: Max Niemeyer Verlag.

Bullon, S. (2006). "The use of corpora in pedagogical lexicography". In *Language Corpora: Their Compilation and Application* (Proceedings of the 13th NIJL International Symposium). Tokyo. 1-8.

Dolezal, F. T.; McCreary, D. R. (1999). Pedagogical lexicography today. A critical bibliography on learners' dictionaries with special emphasis on language learners and dictionary users. Tübingen: Max Niemeyer Verlag.

EAGLES (1996). *Preliminary recommendations on corpus typology* [online]. EAGLES. *http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html* [Access date: 27 March 2008].

EAGLES (2000). Recommendations on corpus encoding [online]. EAGLES. *http://www.cs.vassar.edu/CES/* [Access date: 27 March 2008].

Gavrilidou, M. (2002). "The Hellenic National Corpus on-line". *Revue Belge de Philologie et d'Histoire* 80. Société pour le progrès des études philologiques et historiques. 1003-1015.

Graves, M.; Juel, C.; Graves, B. (2004). *Teaching reading in the 21st Century*. 3rd ed. Boston: Allyn & Bacon.

ten Hacken, P.; Abel, A.; Knap, J. (2006). "Word Formation in an Electronic Learners' Dictionary: ELDIT". *International Journal of Lexicography* 19 (3). 243-256.

Hartmann, R. R. K.; James, G. (1998). *Dictionary of lexicography*. Routledge.

*Hellenic National Corpus* [online]. Institute for Language and Speech Processing. *http://hnc.ilsp.gr* [Access date: 27 March 2008].

Mel'čuk, I.; Clas, A.; Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvaine-la-Neuve: Duculot.

Nielsen, S. (1994). *The Bilingual LSP Dictionary*. Tübingen: Gunter Narr Verlag.

Rundell, M. (1998). "Recent trends in English pedagogical lexicography". *International Journal of Lexicography* 11 (4). 315-342.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.