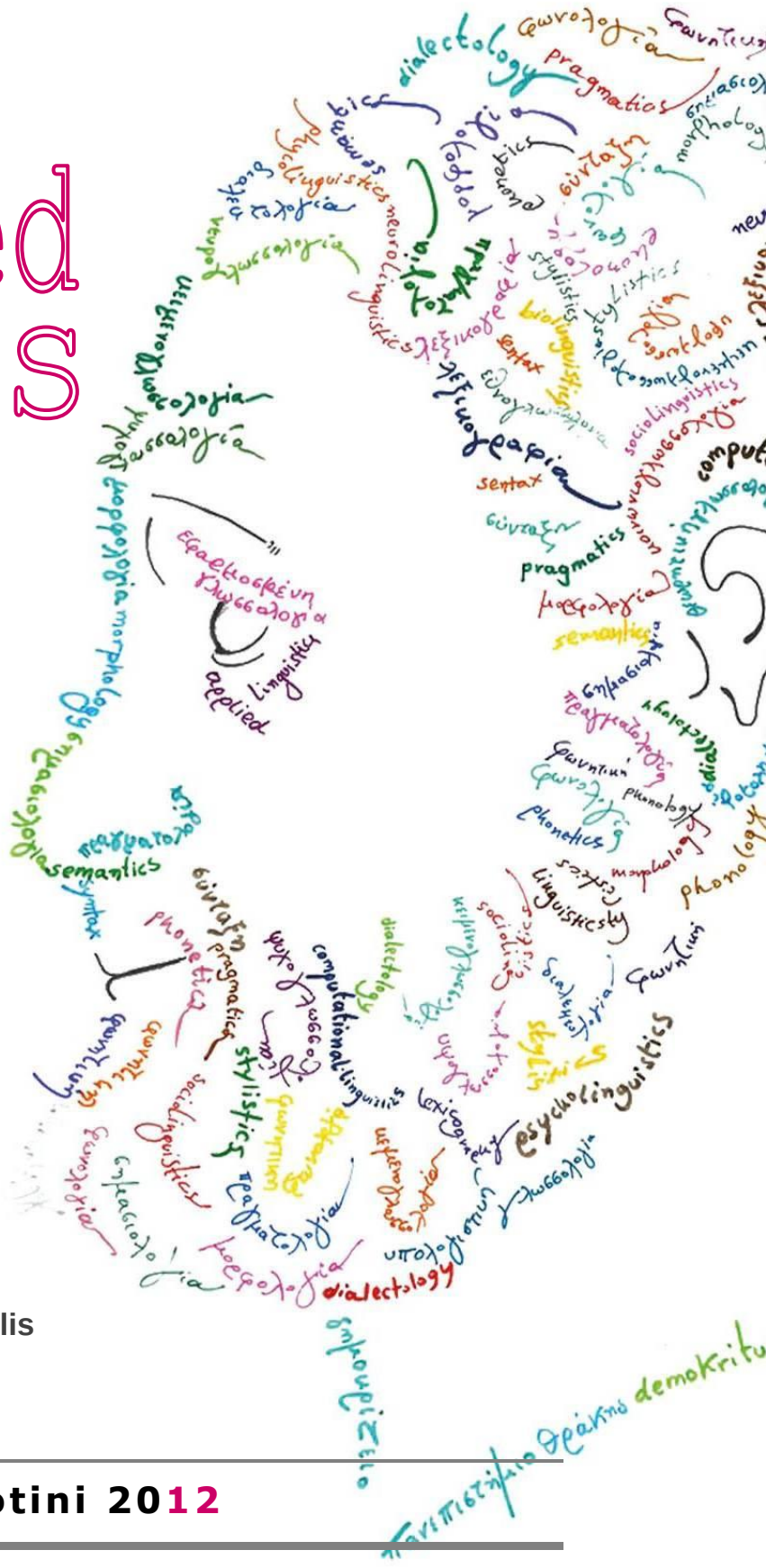
DEMOCRITUS UNIVERSITY of THRACE

selected papers of the 10th International Conference of Greek Linguistics



Edited by

Zoe Gavriilidou
Angeliki Efthymiou
Evangelia Thomadaki
Penelope Kambakis-Vougiouklis

Komotini 2012

ISBN 978-960-99486-7-8

Τυπογραφική επιμέλεια

Νίκος Μαθιουδάκης
Ελένη Παπαδοπούλου
Ελίνα Χατζηπαπά

Σχεδιασμός εξώφυλλου

Νίκος Μαθιουδάκης

CREATING FREQUENCY-BASED VOCABULARY LISTS FOR L2 LEARNERS

Frieda Charalabopoulou
Institute for Language and Speech Processing /"Athena" R.C.
frieda@ilsp.athena-innovation.gr

Maria Gavrilidou
Institute for Language and Speech Processing /"Athena" R.C.
maria@ilsp.athena-innovation.gr

ABSTRACT

*This paper presents the EU-funded project KELLY (**Ke**ywords for **L**anguage **L**earning for **Y**oung and **A**dults alike) for building corpus-informed vocabulary lists for nine languages: Greek, Arabic, Chinese, English, Italian, Norwegian, Polish, Russian and Swedish. The emerging ranked lists were aligned to the six levels of the Common European Framework of the Council of Europe (A1-C2) and grouped to thematic domains. The methodology employed for constructing the lists is described here, while the focus of the paper is on the application of this methodology in what concerns the Greek language.*

Keywords: corpus-informed word lists, L2 learners, vocabulary learning, CEFR, flash-cards

1. Introduction

Until recently vocabulary was not considered one of the priority areas in the field of second language (L2) teaching and was mainly viewed as a secondary activity with the emphasis placed on the development of communicative and linguistic (i.e. grammatical) competence. Nowadays, the primary role of vocabulary in Second Language Acquisition has been acknowledged and vocabulary knowledge is considered indispensable and a prerequisite for acquiring the four traditional language skills (i.e. listening, speaking, reading and writing) as well as grammar: knowing the words of the oral or written input the learners are exposed to enables understanding of discourse, while developing the L2 learner's lexical competence fosters language production.

Lexical competence is a crucial aspect in L2 learning given that successful communication in L2 is interwoven with knowing and deploying the right words to convey the message. According to Nation (2001), language comprehension and production is heavily dependent on vocabulary size, with 3.000 word families being a crucial threshold. A systematic and principled approach in order to build and expand the L2 learners' mental lexicon, therefore, results to better language learning. The crucial question, however, is which words to teach at different language levels.

In this paper we present the EU-funded KELLY project (**Ke**ywords for **L**anguage **L**earning for **Y**oung and **A**dults alike) as an innovative effort to address the above issue. The goal of the project was to generate corpus-informed word lists for L2 learners in 9 languages: Greek, Arabic, Chinese, English, Italian, Norwegian, Polish, Russian and Swedish. Based on the lists, sets of bilingual language learning flash cards in many different language combinations comprising the above-mentioned languages will eventually be developed and offered as an on-line vocabulary-building tool in the form of digital bilingual flash-cards. The project was multidisciplinary and comprised experts in language engineering, lexicology, corpus processing, CALL and product development. The consortium consisted of the following partners: Institute for Language and Speech Processing/Athena" R.C. (Greece), Stockholm University (Sweden), Adam Mickiewicz University (Poland), Cambridge Lexicography and Language Services (UK), Italian National Research Council (CNR) , Keewords AB (Sweden), Lexical Computing Ltd. (UK), Gothenburg University (Sweden), University of Leeds (UK) and University of Oslo (Norway).

In Z. Gavriilidou, A. Efthymiou, E. Thomadaki & P. Kambakis-Vougiouklis (eds), 2012,
Selected papers of the 10th ICGL, pp. 195-201. Komotini/Greece: Democritus University of Thrace.

The overall procedure adopted to carry out the above tasks is described in the following sections. After a presentation of the global principles and processes governing all nine languages of the project, the paper focusses on the procedure for the compilation of the Greek lists.

2. The notion of wordlist

Wordlists are considered as a compact (and flat) representation of a corpus in computational linguistics or information theory: they lack much of the information resident in a corpus, but through the simple listing of all the words found in the corpus with their corresponding frequencies, we get a glimpse at the profile of the corpus. Domains such as language production, understanding and acquisition are interested in word frequency, as a word's frequency is related to the speed with which it is understood or learned. In lexicography, frequency lists constitute basic material for the construction of the macrostructure of the dictionary, i.e. the decision which words are to be included. In language teaching, word lists are commonly used in the process of syllabus design and of deciding which words should be included in books teaching children to read, in textbooks for non-native learners, in language tests etc.

From a language learning perspective, word lists and flash cards may be considered as powerful vocabulary-building tools in the context of intentional L2 vocabulary study. A substantial body of research indicates that dedicated vocabulary learning (as opposed to incidental) should have its place in the L2 learning and instructional context, as "there is a very large number of studies showing the effectiveness of such learning (i.e. using vocabulary cards) in terms of the amount and speed of learning." (Nation, 1997). Such studies favour using list and word learning (see for example Schmitt & Schmitt, 1995; Waring, 2004; Mondria & Mondria-de Vries, 1994; Nation, 2001), as the use of the above-mentioned tools exhibits good retention rates (Hulstijn, 2001; Nation, 2001) and faster learning gains. Fostering learner autonomy is an additional argument in favour of the use of word lists and flash cards by L2 learners, since using such tools allows them to work on their own at their individualized pace. Schmitt (1997: 215) notes that "One main advantage of flash cards is that they can be taken almost anywhere and studied when one has a free moment. Another that they can be arranged to create logical groupings of the target words." However, it should be noted at this point that the use of such devices requires motivated and disciplined learners, who should also be able to deploy the right metacognitive strategies required for self-monitoring, planning their own learning etc. As Nakata (2008:7) notes: "If they (learners) cannot monitor their learning accurately and plan their review schedule accordingly, they cannot make the most of word cards and may run the risk of inefficient learning, e.g. over-learning (devoting more time than necessary) of easy items or under-learning of hard items".

3. The methodology adopted

The initial goal was to identify for all nine languages the words that have the highest frequency but at the same time are the most useful for L2 learners. The number of words to be included in the final lists was set to 9,000 per language and the method was decided to be corpus-driven as far as possible. The procedure for preparing the list for each language was as follows:

- Identify the corpus (or corpora) to be used
- Perform a frequency count based on the corpus
- Generate a monolingual frequency list
- Use additional resources to enhance the monolingual corpus-informed lists with words considered essential for language learners per language level
- Translate each item into all the other project languages
- Compare lists and identify items for addition or deletion
- Enrich the monolingual lists by adding words as a result of cross-language comparison
- Provide the final bilingual lists, which were then ranked and aligned to the six CEFR levels (A1-C2)

The above steps were common for all languages; however, as in this paper the focus is on the Greek language, the next sections present this common methodology as applied to the development of the Greek KELLY list.

3.1 Corpus identification

The objective of the endeavour dictated the specifications for the corpus to be selected: it should contain general, everyday language and it should be large with a variety of texts, so that it would not be biased towards any particular text type or topic and would not miss basic vocabulary. An additional specification was for all corpora across languages to be, as far as possible, 'comparable'; in this way, all the lists produced would represent the same kind of language. For some of the languages of the project there was a good choice of corpora available, but not all project languages were equally well served. Spoken corpora were only available for a minority of the languages.

A web corpus provides large bulks of data of general language in a variety of topics and genres. This can be created for any language, using methods as presented in Sharoff (2006). These methods of corpus creation result in corpora that serve the purpose of KELLY better than the BNC-type corpora, which typically have large components of newspaper and fiction, while the predominant language features are past tense verbs, third person pronouns and other prototypical written language features. On the other hand, web corpora are more personal, action-based and future-oriented, and they include more present and future tense verbs, first and second person pronouns, i.e. prototypical spoken language features also. According to Ferraresi et al (2008) there is a better match between CEF *can-do* statements and web corpora, than *can-do* statements and BNC-type corpora.

As regards Greek, two different corpora were used in comparison, with the aim to conduct a comparative study of the frequency wordlists extracted from each type. Specifically:

- Greek Web as a Corpus (GkWaC): a random collection of Greek texts over the web of a size of 41 Million Words, produced in 2010.
- The Hellenic National Corpus (HNC, http://hnc.ilsp.gr: 50 Million Words, published between 1990 and 2010, consisting of newspapers, magazines, books (fiction & non-fiction) and miscellaneous texts.

3.2 Generation of the Greek frequency list

The differences in corpus constitution are reflected in the wordlists produced. Thus, since the HNC contains samples of written language exclusively (mainly texts from high circulation newspapers, best-selling books etc., with no idiomatic language), the profile that emerges is that of a higher register, whereas daily/informal speech is largely underrepresented. On the other hand, the GkWaC comprises a variety of sources (news, blogs, chats, etc.) and represents mainly the informal register of everyday speech, approximating spoken language. The steps that were taken to generate the monolingual Greek word list are described below:

Step 1
The corpora were processed with the aim to extract the most frequent words, or rather, the most frequent lemmas, i.e. the canonical forms representing all inflected types. In view of this objective, all three corpora were automatically tokenized, tagged and lemmatized with the respective ILSP tools. The process was fully automatic, with no subsequent manual correction.

Step 2
The subsequent phase included the extraction of frequency lists of the so-called *lemgrams*, i.e. the combinations of lemma and part of speech. The inclusion of the grammatical information on the part of speech was considered necessary in order to have the full identity of a word and to differentiate between cases of the type *ασθενής* (noun) and *ασθενής –ής –ές* (adj.)

Step 3
Unwanted items such as foreign words, hyperlinks, proper names, toponyms etc. were filtered out, since they do not consitute material appropriate to be included in vocabulary learning.

Step 4

At this stage, the lists extracted from each corpus were merged in one.

Step 5
The unified list was checked for duplicates and errors in the lemmatizer and tagger output. In the first case, where a lemma was found in both lists with differing frequency, the item originating in GkWac (and its frequency) was selected; this decision was due to the fact that the web corpus, as already discussed, is more suitable for L2 learners. In the second case of erroneous lemma or part of speech assignment by the automatic procedure, manual inspection of the list and correction of errors was deemed necessary.

For Greek, as for all other languages, the 6,000 most frequent *lemgrams* (lemma+part of speech pairs) were selected during this stage; this was the monolingual list, which served as input to the next process. The inclusion specifications for these 6,000 words comprised the following:

- Spelling variants were amalgamated, so that e.g. αυγό and αβγό [egg] were counted as one word for frequency calculations.
- Inflected forms are not included as such in the list, unless an inflected form has a meaning that is not inherent in the base form, or if the specific word does not follow the rule for lemma, e.g. πρόποδες [foot of the mountain], which is a word lacking the singular number, so the lemma is found in the plural.
- Derivational forms, such as adverbs, deverbal nouns etc., were treated as separate lemgrams.
- Affixes, including productive affixes, were not listed as separate entries.
- The most common abbreviations were listed; these were forms of address, weights and measures and the few cases where an abbreviation is the normal way to refer to an item, e.g. DVD.
- Multi-word units were included, when appearing above the set frequency threshold.
- Phrases, idioms, proverbs, quotations were not included.
- Subject-specific vocabulary (terminology) was not included, except for cases where the term has indeed penetrated general language and is commonly used, e.g. some computer terms.
- Dialectal material was not included, given that the target was the production of vocabulary for the teaching of the norm.
- Items marked by register, e.g. very formal, slang or offensive were subject to the frequency rule: if they appeared in the top 6,000 they were included. However, they were marked as such, to be easily removed if need be.
- Beliefs and religions, as well as associated nouns and adjectives were also subject to the frequency rule.
- Toponyms or proper names (names of stars, planets, place names, river names etc., names of imaginary, biblical or mythological creatures, encyclopedic info such as names of wars, treaties, names of ancient peoples, names of organizations, etc. were not included.

3.3 Enhancing the Greek KELLY list with additional resources

In the KELLY project the idea was to build a list from a web-corpus, which would be a corpus of general everyday language consisting of different texts, so that it would not be skewed by topic-specific texts and thus miss any core vocabulary (taking into consideration the language needs of our target group, i.e. L2 learners). Moreover, given that the objective was to build common lists for 9 languages, an additional constraint was that the corpora in the nine languages should be comparable, i.e. represent the same kind of language, which would allow for making connections between them and thus end up with a common word list for all.

Yet, a purely corpus-derived list on its own may have shortcomings, especially when it addresses L2 language learners: including the most frequent words is not enough. Some words may not exhibit high frequency rates, still they may be necessary to know in the context of L2 learning. "For example, such words as pencil, eraser, and blackboard do not occur in the most frequent words of general English, though they are frequent and essential in a classroom context". (Schmitt, 2000: 83). Therefore, the additional requirement for the KELLY lists was that they should include the most useful words according to the learner's language level and, furthermore, these should be in alignment with the CEFR-specific domain vocabulary. In order to meet this additional requirement, available educational resources for the target languages were also consulted.

As far as Greek is concerned, at present there are relatively few vocabulary lists available for Modern Greek, which describe modern vocabulary as well as being adapted to different CEFR levels. The available word lists for Modern Greek that were consulted for the project were the following:

- In a publication issued by the Center for the Greek Language (see Efstathiades et al., 2001) an annex is included with a word list addressing L2 learners of levels A and B (beginners and intermediate) as an annex. The authors label the annex as "Indicative vocabulary for levels A & B" without providing any further information. These lists are not corpus-based and the exact number of lemmas is not specified.
- In the curriculum published by the University of Athens for teaching Modern Greek L2 to adults (University of Athens, 1998), a vocabulary list is included as an appendix. The authors state that the list has been created based solely on their intuition and teaching experience and that the vocabulary listed (which they call "representative vocabulary") complies with the communicative needs and the learning goals specified in the curriculum and relates to particular notions and functions, speech acts and thematic domains. The number of words is not specified.
- A dictionary for Greek as L2 has recently been released as support material within the framework of the EU-funded programme MiNERA - O.P. Education II – "Educational Project for Muslim Children 2005-2008" (for information see http://www.museduc.gr/en/index.php). The dictionary (available at http://www.museduc.gr/docs/gymnasio/Dictionary.pdf) includes 10,000 lemmas, which emerged by a processing combination of (i) existing dictionaries of Modern Greek addressed to pupils of primary and secondary education in Greece (as representative for the definition of the "basic/core vocabulary") and (ii) e-corpora, in which school textbooks were also included. No further information about the corpus is provided.

The Greek corpus-based list was enhanced by manually adding words from the above lists. In particular, the added words belonged to the following thematic domains: animals, fruits, vegetables, colours, family relations, clothes, rooms, furniture, means of transport, shops, doctors and weather.

The fact that certain words that are considered necessary for L2 learners (such as *alphabet*, *elbow* or *orange)* were missing from the corpus-based lists and had to be manually added was not suprising: the lists were derived from general language corpora and not domain-specific or pedagogy-oriented texts. Language coursebooks and L2 curricula, on the other hand, are structured and organized on the basis of communicative situations (e.g. at the super market, shopping, going to the doctor etc.) and relevant vocabulary (e.g. fruits, vegetables, clothes, colours etc.) in compliance with the CEFR-specific domain vocabulary. The overall process for building the Greek monolingual list is depicted in Figure 1 below:

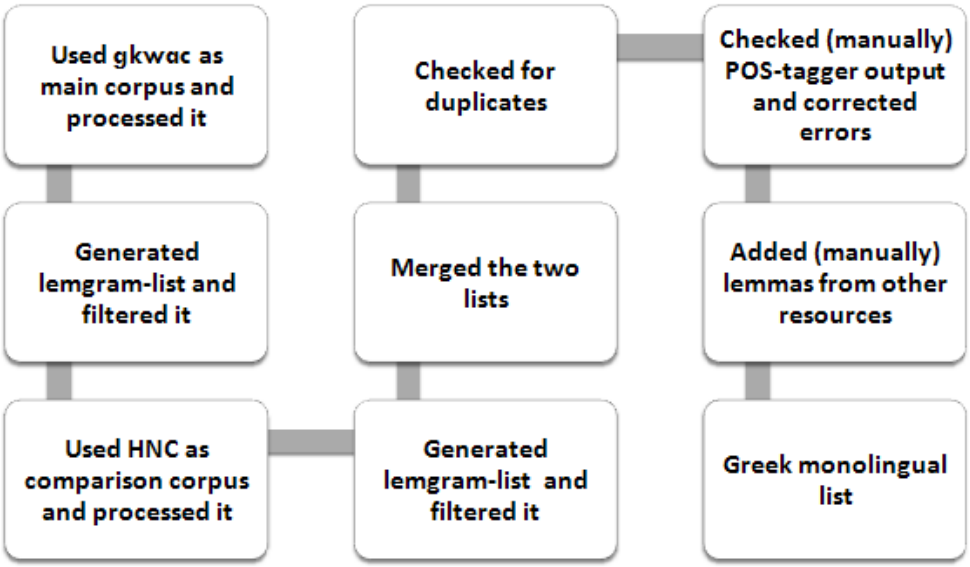


Figure 1 Towards building the Greek monolingual list

3.4 From monolingual to bilingual lists and the KELLY database

The process described above was applied to all other languages and was followed by a cross-language list comparison. The emerging lists were given for translation to all other languages, hence resulting in 72 bilingual lists. Following cross-language comparison, the next step involved handling **"**back translations" (i.e. words used by translators when translating into a language and not appearing in the monolingual lists of this language) in order to decide which of these should be added to the bilingual lists and which should be deleted or demoted.

The completion of this process rendered the final lists, which consist of approximately 9,000 words per language. These words were ranked according to their frequency range, and were equally distributed to the six CEFR-based language levels (i.e. circa 1,500 words per language level).

The content of the final bilingual lists is hosted by the KELLY database, which contains translation mappings to one or more words in each of the other eight languages. In total, it contains 74,258 lemmas and 423,848 mappings, which renders it an interesting resource that may be accessed and exploited for various research and/or learning purposes. The translations are divided in the following two basic categories:

(a) Symmetric pairs (*sympairs*): this is a pair of words <*a, b*> of two different languages A and B, such that *a* translates to *b* and *b* translates to *a*. Table 1 below depicts examples of symmetric pairs across all project languages:

Arabic	Chinese	English	Greek	Italian	Norwegian	Polish	Russian	Swedish
مستشفى	医院	hospital	νοσοκομείο	ospedale	sykehus	Szpital	больница	sjukhus
مكتبة	图书馆	library	βιβλιοθήκη	biblioteca	bibliotek	Biblioteka	библиотека	bibliotek
موسيقى	音乐	music	μουσική	musica	musikk	Muzyka	музыка	music
شمس	太阳	sun	ήλιος	sole	sol	Słońce	солнце	sol
نظرية	理论	theory	θεωρία	teoria	teori	Teoria	теория	teori

Table 1 Example of symmetric pairs

(b) Non-symmetric pairs (*non-sympairs*): this is the case where in a translation pair <*a* of language A, *b* of language B>, *a* translates to *b* but *b* does not translate to *a*, although both *a* and *b* exist in the respective monolingual lists. For example, in the Greek-English pair the word *μοιράζω* is translated as *distribute*, while in the English-Greek pair the word *distribute* is translated as *διανέμω*. The main reasons for non-sympairs are grouped as follows:

- Linguistic, such as use of different spelling variants, difference in word classes across languages, translators' choices, etc.
- Technical, such as differences in corpus construction, list compilation approaches, lemmatization/normalization problems with resulting difference in item frequency range.
- Cultural differences, which result in the presence or absence of certain vocabulary items.

Apart from the KELLY database, the project will lead to the creation of bilingual flash cards, which in the form of an on-line tool that can be used to enhance vocabulary skills in L2.

4. Summary and outlook

In this paper we have presented the KELLY project and the work carried out towards developing corpus-derived word lists, monolingual and bilingual, for nine languages that may be used and expoited within the L2 teaching and learning framework. We have described in detail the method for the monolingual list creation as regards the Greek language and the steps taken for the creation of the final bilingual lists comprising all nine languages.

The KELLY project constitutes an experiment in using automatic solutions for language learning. In that respect, innovative work has been carried out with respect to the following:

- Innovative methodology was used in order to build frequency-based vocabulary lists from web corpora in nine languages.
- A vocabulary-building tool was created, which may be employed either for self-study purposes or as supplementary material for enhancing vocabulary skills in the context of guided instruction.
- Word lists and flash cards were developed for less widely taught and learned languages and "unusual" language pairs (e.g. Greek-Norwegian, Polish-Italian, Swedish-Arabic etc.).
- A wide spectrum of L2 learners (i.e. young and adults, from beginners to advanced) and learner types was addressed.
- Words were ranked according to the Common European Framework and organized to CEFR-based thematic domains.

Apart from the advantages and innovations, the work carried out within the KELLY project has raised a number of issues which need to be addressed in the future. From a language pedagogy perspective, the crucial question is: how efficient are corpus-informed word lists as pedagogical tools for L2 learning? Is employing purely lexico-statistical approaches to define vocabulary syllabus for L2 learners a safe approach? In other words, can we rely merely on technology and purely objective strategies when it comes to the selection of relevant vocabulary for L2 learners? Even more so, when word lists need to cover the CEFR-related thematic domains and topics?

Future plans for the KELLY outcomes mainly focus on two areas: commercial and linguistic. The commercial dimension involves the creation, sales and marketing of digital bilingual flash cards by the commercial partner of the consortium (Keewords AB). The linguistic part refers to the further exploration of the KELLY database and the evaluation and validation of the KELLY outcomes by the end-users, i.e. actual L2 learners, in order to overcome existing shortcomings and provide a reference tool for vocabulary learning. For less widely used and taught languages (such as Greek) the KELLY lists may prove particularly useful and could be considered, in that respect, a welcome addition.

References

Efstathiadis, Stathis, Niovi Antonopoulou, Dimitra Manavi, and Smaro Vogiatzidou. 2001. *Certificate of Attainment in Greek*. Salonica: Ministry of Education-Center for the Greek Language.

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. "Introducing and evaluating ukWaC, a very large web-derived corpus of English". In *Proc. 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, edited by Stefan Evert, Adam Kilgarriff, and Serge Sharoff. 47-54. Marrakech, Morocco.

Hulstijn, Joris. 2001. "Intentional and incidental second language vocabulary learning: a reappraisal of elaboration, rehearsal, and automaticity". In *Cognition and second language instruction* edited by Robinson, P. Cambridge: Cambridge University Press, 258–286.

Laufer, Batia. 2003. "Vocabulary acquisition in a second language: do learners really acquire most vocabulary by reading? some empirical evidence". *Canadian Modern Language Review*, 59(4): 567–587.

Mondria, Jan-Arjen, and Siebrich Mondria-de Vries. 1994. "Efficiently memorizing words with the help of word cards and 'hand computer': theory and applications". *System, 22*(1): 47–57.

Nakata, Tatsuya. 2008. "English vocabulary learning with word lists, word cards and computers; implications from cognitive psychology for optimal spaced learning". *ReCALL, 20*(1): 3–20

Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge

Nation, Paul., and Michael McCarthy.1997. "Vocabulary size, text coverage and word lists". In *Vocabulary: Description, Acquisition and Pedagogy*, edited by Norbert Schmitt, and Michael McCarthy, 6-20. Cambridge University Press

Schmitt, Norbert, and Diane Schmitt. 1995. "Vocabulary notebooks: theoretical underpinnings and practical suggestions". *ELT Journal, 49*(2): 133–143.

Schmitt, Norbert. 1997. "Vocabulary learning strategies". In *Vocabulary: Description, Acquisition and Pedagogy,* edited by Norbert Schmitt, and Michael McCarthy, 199-227. Cambridge University Press.

Schmitt, Norbert. 2000. *Vocabulary in Language Teaching*. Cambridge University Press.

Sharoff, Serge. 2006. "Creating general-purpose corpora using automated search engine queries". In *WaCky! Working papers on the Web as Corpus,* edited by Marco Baroni, and Silvia Bernardini. Gedit, Bologna.

University of Athens. 1998. *Curriculum for Teaching Modern Greek as a Foreign Language to Adults (Levels 1 and 2: Introductory and Basic).* Athens: University of Athens.

Waring, Rob. 2004. "In defence of learning words in word pairs: but only when doing it the 'right' way!" Accessed September 25, 2011. http://www1.harenet.ne.jp/~waring/vocab/principles/systematic_learning.htm Retrieved 25/9/2011