Building Corpus-Informed Word Lists for L2 Vocabulary Learning in Nine Languages

Frieda Charalabopoulou[a*], Maria Gavrilidou[a],
Sofie Johansson Kokkinakis[b], and Elena Volodina[b]

a. ILSP/"Athena" R.C., Artemidos and Epidavrou, Maroussi-Athens, Greece
b. Språkbanken, Institutionen för svenska språket, Göteborgs universitet, Göteborg, Sweden

Abstract. Lexical competence constitutes a crucial aspect in L2 learning, since building a rich repository of words is considered indispensable for successful communication. CALL practitioners have experimented with various kinds of computer-mediated glosses to facilitate L2 vocabulary building in the context of incidental vocabulary learning. Intentional learning, on the other hand, is generally underestimated, since it is considered out of fashion and not in line with the communicative L2 learning paradigm. Yet, work is still being done in this area and a substantial body of research indicates that the usefulness of incidental vocabulary learning does not exclude the use of dedicated vocabulary study and that by using aids explicitly geared to building vocabularies (such as word lists and word cards), L2 learners exhibit good retention rates and faster learning gains. Intentional vocabulary study should, therefore, have its place in the instructional and learning context. Regardless of the approach, incidental or intentional, the crucial question with respect to vocabulary teaching/learning remains: which and how many words should we teach/learn at different language levels? An attempt to answer the above question was made within the framework of the EU-funded project titled "KELLY" (**Ke**ywords for **L**anguage **L**earning for **Y**oung and Adults Alike) presented here. The project aimed at building corpus-informed vocabulary lists for L2 learners ranging from A1 to C2 levels for nine languages: Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish.

Keywords: intentional vocabulary learning, corpora, CEFR, corpus-informed word lists, digital cards.

1. Introduction

According to Nation (2001), language comprehension and production is heavily dependent on vocabulary size, with 3,000 word families being a crucial threshold. A

* Contact author: frieda@ilsp.athena-innovation.gr

In L. Bradley & S. Thouësny (Eds.), *CALL: Using, Learning, Knowing, EUROCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings* (pp. 49-53). © Research-publishing.net Dublin 2012

systematic and principled approach in order to build and expand the L2 learners' mental lexicon, therefore, results in better L2 learning. Given that vocabulary knowledge constitutes an integral part of general proficiency in L2 and a prerequisite for successful communication, the question is how it should best be taught and learned.

Intentional vocabulary learning (involving focused activities aiming directly at learning lexical items, such as using word cards and vocabulary lists) is often considered out of fashion and dismissed in the context of the communicative approach in L2 teaching and learning. Contextualised and incidental vocabulary learning, on the other hand, where learning vocabulary is considered a by-product of other L2 activities not primarily focusing on the systematic learning of words, seems to fit perfectly within the communicative framework.

While vocabulary learning from context seems to be favoured, a number of studies show that such learning has its drawbacks: it may require learners to engage in large amounts of reading and listening and may be more demanding and slow; it requires exposure to words through reading, listening and speaking, which, however, should be combined with a systematic study of lexical items, collocations etc. In addition, if the L2 learner has limited exposure to L2 outside the classroom, word-focused activities should complement vocabulary learning in context (Hulstijn, 2001; Laufer, 2003; Nation, 2001). On the other hand, considerable amounts of research (Ma & Kelly, 2006; Nation & Waring, 1997; Read, 2000) indicate that intentional vocabulary learning realised by using word lists and cards could be beneficial and should therefore have its place in the instructional/ learning context.

Regardless of the approach, the crucial question is: which and how many words should we teach/learn at different language levels? The aim of the KELLY project was to address the above questions and generate corpus-informed word lists for L2 learners in 9 languages: Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish. The overall procedure adopted to carry out the above task is described in the following section.

2. Method

The main goal of the KELLY project was to identify for all nine languages the words that exhibit the highest frequency rates but at the same time are the most useful for L2 learners. The procedure for preparing the word lists comprised the following steps.

2.1. Corpus identification and corpus enhancement

The objective of the endeavour dictated the specifications for the corpus selection: 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. Moreover, all corpora should be 'comparable' in all languages, so that all the lists would represent the same kind of language.

The main advantage of a web corpus is that it provides large bulks of data of general language in a variety of topics and genres and can be created for any language using various methods (see for example Sharoff, 2006). These methods result in corpora that serve the purpose of KELLY better than the BNC-type corpora, which typically have large components of newspapers and fiction, while the predominant language features are past tense verbs, third person pronouns and other prototypical written language features. Web corpora are more personal, action-based and future-oriented, and they include more prototypical spoken language features (e.g., present and future tense verbs, first and second person pronouns, etc.). According to Ferraresi, Zanchetta, Baroni, and Bernardini (2008), there is a better match between Common European Framework of Reference for Languages (CEFR) can-do statements (Council of Europe, 2001) and web corpora compared to BNC-type corpora. Taking into consideration all the above and in combination with the educational needs of our target group (L2 learners), in the KELLY project we opted for building word lists based on frequency from web corpora of general everyday language and comprising of different texts, thus not skewing the project by topic-specific texts.

Yet, a purely corpus-informed approach to build word lists addressing L2 learners may have certain shortcomings, including: the most frequent words may simply not be enough. Some words may not exhibit high frequency rates, yet they may be necessary and useful in the context of L2 learning. Therefore, the additional requirement for the KELLY lists was that they should include the most useful words according to the learner's L2 level and, furthermore, these should be in alignment with the CEFR-specific domain vocabulary. In order to meet this additional requirement, available educational resources (i.e., course books, dictionaries, already available vocabulary lists) were also consulted to enhance the original corpus-informed lists. After this enhancement process, the monolingual (M1) frequency lists were built for all 9 languages.

2.2. Building the bilingual word lists and the KELLY database

Each of the nine M1 lists were then translated into the eight other languages, thus rendering 72 translation lists. This process was followed by a cross-language list comparison and 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 emerging lists were translated to all other languages, hence resulting in the final 72 bilingual lists with each translation pair accompanied by word class, frequency, translator notes, etc. These words were ranked according to their frequency range and were equally distributed to the six CEFR-based proficiency levels resulting in approximately 1,500 words per proficiency level after merging two translated lists with each other (for instance Swedish-Greek and Greek-Swedish). The content of the final bilingual list is hosted by the KELLY database (available at http://kelly.sketchengine.co.uk), which contains

74,258 lemmas and 423,848 mappings, hence rendering it an interesting resource that may be deployed and exploited for both research and educational purposes.

3. Discussion

The KELLY project constitutes an experiment in employing automatic solutions for L2 learning. Based on the emerging word lists, the end-product of this endeavour comprises an on-line educational service for vocabulary (self-directed) study in nine languages in the form of bilingual digital cards addressing all language proficiency levels (A1-C2). The cards are divided into subject categories/domains, thus enabling the users to tailor their vocabulary studies to individualised communicative needs and goals. Within KELLY, innovative work has been carried out with respect to the following:

- Innovative methodology for building frequency-based vocabulary lists from web corpora in nine languages;
- Creation of a vocabulary-building tool, which may be employed either for self-study purposes or as supplementary material for enhancing vocabulary skills in the context of guided instruction;
- Development of word lists and digital cards for less widely taught and learned languages and "unusual" language pairs (e.g., Greek-Norwegian, Polish-Italian, Swedish-Arabic etc.), available at http://www.keewords.com/en/;
- Addressing a wide spectrum of L2 learners (i.e., youngsters (-16) and adults, from beginners to advanced, guided and self-directed in L2 learning) and learner types;
- Ranking words according to the Common European Framework and organised 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 questions are: how efficient are corpus-informed word lists as pedagogical tools for L2 learning? Is employing purely lexico-statistical approaches to define vocabulary syllabuses for L2 learners a good enough approach? In other words, can we rely merely on technology and purely on objective strategies when it comes to the selection of relevant vocabulary for L2 learners? Even more so, when do word lists need to cover the CEFR-related thematic domains and topics?

4. Conclusions

In this paper we presented the KELLY project and its outcomes as an example of work carried out in order to develop corpus-derived word lists for nine languages that may be used and exploited within the L2 teaching and learning framework as vocabulary-building tools.

The pedagogical potentials of the lists and the digital cards mainly involve their use directly as a learning tool that may be deployed for vocabulary (self-)study as well as indirectly, e.g., for analysis of lexical complexity of L2 texts. As far as their pedagogical effectiveness is concerned, this needs to be validated by the end-users, i.e., actual L2 learners, in order to overcome existing shortcomings and provide a really useful reference tool for vocabulary learning. Evaluation and validation embraces issues such as content from an L2 learning perspective, its relation to the CEFR scale, coverage of the KELLY lists compared to different corpora and/or L2 course books based on CEFR, etc. One interesting aspect that also needs to be investigated is to what extent the KELLY lists could be considered as key resources and potential candidates for official vocabularies, especially with regard to those languages which lack such valuable resources.

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching and Assessment.* Cambridge: Cambridge University.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff, & S. Sharoff (Eds.), *Proc. 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?* (pp. 47-54). Marrakech, Morocco.

Hulstijn, J. (2001). Intentional and incidental second language vocabulary learning: a reappraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge, UK: Cambridge University Press.

Laufer, B. (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.

Ma, Q., & Kelly, P. (2006). Computer-Assisted Vocabulary Learning: Design and Evaluation. *Computer-Assisted Language Learning, 19*(1), 15-45. doi: 10.1080/09588220600803998

Nation, P., & Waring, R. (1997). Vocabulary Size, Text Coverage and Word Lists. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19). Cambridge University Press.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Read, J. (2000). *Assessing Vocabulary*. Cambridge, UK: Cambridge University Press.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni, & S. Bernardini (Eds.), *WaCky! Working papers on the Web as Corpus* (pp. 63-98). Bologna: Gedit.

Published by Research-publishing.net
Dublin, Ireland; Voillans, France
info@research-publishing.net

CALL: Using, Learning, Knowing
EUROCALL Conference, Gothenburg, Sweden
22-25 August 2012, Proceedings
Edited by Linda Bradley and Sylvie Thouësny

Typeset by Research-publishing.net
Cover design: © Raphaël Savina (raphael@savina.net)
Aquarelle reproduced with kind permission from the illustrator: © Sylvi Vigmo (sylvi.vigmo@ped.gu.se)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-03-2 (paperback)
Print on demand (lulu.com)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Bibliothèque Nationale de France - Dépôt légal: décembre 2012.