Lang Resources & Evaluation
DOI 10.1007/s10579-013-9251-2

ORIGINAL PAPER

Corpus-based vocabulary lists for language learners for nine languages

Adam Kilgarriff · Frieda Charalabopoulou · Maria Gavrilidou · Janne Bondi Johannessen · Saussan Khalil · Sofie Johansson Kokkinakis · Robert Lew · Serge Sharoff · Ravikiran Vadlapudi · Elena Volodina

Abstract We present the KELLY project and its work on developing monolingual and bilingual word lists for language learning, using corpus methods, for nine languages and thirty-six language pairs. We describe the method and discuss the many challenges encountered. We have loaded the data into an online database to make it accessible for anyone to explore and we present our own first explorations of it. The focus of the paper is thus twofold, covering pedagogical and methodological aspects of the lists' construction, and linguistic aspects of the by-product of the project, the KELLY database.

We would like to dedicate this paper to our colleague Frieda Charalabopoulou, who died, following a long struggle with cancer, between its writing and its publication.

A. Kilgarriff (✉) · R. Vadlapudi
Lexical Computing Ltd, Brighton, UK
e-mail: adam@lexmasterclass.com

F. Charalabopoulou · M. Gavrilidou
ILSP/'Athena' R.C., Artemidos 6 & Epidavrou, 151 25 Maroussi, Athens, Greece

J. B. Johannessen
The Text Laboratory & Centre for Multilingualism in Society across the Lifespan, Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo, Norway

S. Khalil · S. Sharoff
Centre for Translation Studies, University of Leeds, Leeds, UK

S. Johansson Kokkinakis · E. Volodina
Språkbanken, Institutionen för svenska språket, Göteborgs universitet, Box 200, 405 30 Göteborg, Sweden

R. Lew
Department of Lexicography and Lexicology, Faculty of English, Adam Mickiewicz University in Poznań, Poznań, Poland

Published online: 14 September 2013

Keywords Corpora · Language learning · Vocabulary · Frequency · Frequency lists

1 Introduction

Word lists are much-used resources in many disciplines, from language learning to psycholinguistics. A natural way to develop a word list is from a corpus. Yet a corpus-derived list on its own usually has grave shortcomings as a practical resource. In this paper we explore a substantial effort to generate word lists for nine languages, as far as possible in a corpus-driven, principled way, but with the overriding priority of creating lists which are as useful as possible for language learners.

The goal of the KELLY project[1] was to develop sets of bilingual language learning word cards in many different language combinations. For this we needed to know which words to include, and we wanted them to be the 9,000 most frequent words in nine languages. We then added a research goal: to use as principled a corpus-driven method as possible. The lists needed to be ordered, so learners could learn the more common words first. Four of the languages were 'more commonly taught' (Arabic, Chinese, English, Russian), the other five 'less commonly taught' (Italian, Swedish, Norwegian, Greek, Polish). The selection of the languages was dictated by three factors: the company that initiated the idea (Keewords AB, Sweden) and their interests; the EU Lifelong Learning Programme's agenda of improving resources for smaller languages and less obvious language pairs; and participants' research networks.

The KELLY procedure for preparing the list for each language was as follows:

- Identify the corpus
- Generate a frequency list (the 'Monolingual 1' or 'M1' list)
- Clean up the list, and compare it with lists from other corpora and other wordlists
- Make adjustments to give the 'M2' list
- Translate each item into all the other KELLY languages (the 'Translation 1' or 'T1' list)
- Use the 'back translations' to identify items for addition or deletion
- Make further adjustments to give the final, M3 list.

While the process was corpus-based, it was not one in which the corpus was religiously seen as the authority. Every corpus has peccadilloes, and the corpus to which you have access is rarely the ideal corpus for the task at hand. So, at various points, we were happy for expert judgement to overrule corpus frequencies. The paper considers these divergences and what underlies them.

[1] EU Lifelong Learning Programme Grant 505630. Partners: Stockholm University, Sweden (co-ordinators); Adam Mickiewicz University, Poland; Cambridge Lexicography and Language Services, UK; Institute for Language and Speech Processing (ILSP), Greece; Italian National Research Council (CNR), Italy; Keewords AB, Sweden; Lexical Computing Ltd., UK; University of Gothenburg, Sweden; University of Leeds, UK; University of Oslo, Norway.

Once the process was complete, the translations were entered into a database which let us ask questions like "What 'symmetrical pairs' are there, where X is translated as Y, and Y is also translated as X?" and "What word sets of three or more words (all of different languages) are there where all words are in symmetric pairs with all others?". The database is available to all to interrogate.[2]

The structure of the paper is as follows: Sect. 2 discusses word lists and presents an overview of the relevant literature, Sect. 3 gives details of the KELLY procedure for preparing lists, Sect. 4 considers the KELLY database as a resource for linguistic research, and Sect. 5 concludes.

2 Word lists

Word frequency lists can be seen from several perspectives. For computational linguistics or information theory, they are also called unigram lists and can be seen as a compact representation of a corpus, lacking much of the information (being decontextualised), but small and easily tractable. Unigram lists (and also n-gram lists where n = 2, 3, 4) are basic for all language modeling, from speech recognition to machine translation. Systems that use word lists in areas relating to language learning include automatic rating of good corpus examples where the vocabulary is checked for being common (frequent) versus rare (infrequent) (Kilgarriff et al. 2008; Kosem et al. 2011; Borin et al. 2012), and readability analysis where texts are analyzed for their lexical frequency profiles (Heimann Mühlenbock 2012; Volodina 2010).

Psychologists exploring language production, understanding, and acquisition are also interested in word frequency, as a word's frequency is related to the speed with which it is understood or learned. So frequency needs to be used as a criterion in choosing words to use in psycholinguistic experiments. A number of frequency-based word lists constitute a part of the Psycholinguistic Database[3] with the named resources being used in different experiments, for example Davis (2005) and Aitchison (2012).

Educationalists are interested in frequency too, as it can guide the curriculum for learning to read and similar. To these ends, for English, Thorndike and Lorge prepared *The Teacher's WordBook of 30,000 words* in 1944 by counting words in a corpus, creating a reference set used for many studies for many years (Thorndike and Lorge 1944). It made its way into English language teaching via *West's General Service List* (West 1953), which was a key resource for choosing which words to use in the English language teaching curriculum until the British National Corpus replaced it in the 1990s. More recently, the English Profile project[4] has developed the 'English Vocabulary Profile' which lists vocabulary for each CEFR level[5] (Capel 2010).

[2] http://kelly.sketchengine.co.uk.

[3] http://www.psych.rl.ac.uk/.

[4] http://www.englishprofile.org.

[5] CEFR: Common European Framework of Reference for Languages (Council of Europe 2001).

In language teaching, word frequency lists are used among other things for:

- defining a syllabus
- building graded readers
- deciding which words are used in:
  - learning-to-read books for children
  - textbooks for second language (L2) learners
  - dictionaries
  - language tests for L2 learners

2.1 The pedagogical perspective: learning vocabulary using lists and cards

Vocabulary learning is an essential part of mastering a second language (L2). According to Nation (2001), vocabulary knowledge constitutes an integral part of learners' general L2 proficiency and is a prerequisite for successful communication.

In terms of language pedagogy, there are two generally accepted approaches to vocabulary learning: *intentional*, where activities are aimed directly at learning lexical items, such as using word lists and cards; and *incidental*, where learning vocabulary is a by-product of activities not primarily focused on the systematic learning of words, such as reading (Nation 2001).

Although sometimes seen as opposed to each other (Nation 2001:232), both intentional and incidental vocabulary learning should have a place in language learning and should be seen as complementary to each other (Hulstijn 2001).

From the communicative perspective, incidental or 'contextual' vocabulary learning contributes to successful lexical development, while intentional learning, especially if it involves rote learning such as using word lists and cards, may result in misuse of the vocabulary since words are learned in isolation. Intentional learning may even fail to transfer information contained in chunks of language (e.g. collocations, expressions etc.), seen as essential for communicative fluency (McCarten 2007). Intentional learning methods have therefore largely fallen out of fashion or been dismissed by advocates of the communicative approach.

A substantial body of research, however, lends support to the claim that intentional or 'decontextualised' vocabulary learning using word lists and cards should not be marginalised. In her discussion of L2 vocabulary acquisition, Laufer (2003), for example, has shown that this type of learning may in certain cases prove to be more efficient than incidental/contextualised vocabulary learning, since incidental learning requires exposure to rich L2 input environments as well as extensive reading and listening, which delays the whole learning process. She estimates that learners may need to read a text of 200,000 words in order to learn 108 words from context, which seems unrealistic given classroom limitations. If a learner has limited exposure to the L2 outside the classroom, then intentional, word-focused activities should complement contextual vocabulary learning (Hulstijn 2001; Laufer 2003; Nation 2001). List learning in particular can be of particular benefit for lower-level L2 learners and prove to be an efficient way to achieve vocabulary mastery.

A key issue for vocabulary learning is retention, and a key aim of vocabulary learning activities and materials should be long-term retention. There are a number of studies that have indicated the usefulness of lists in word-learning, such as Schmitt and Schmitt 1995; Waring 2004; and Mondria and Mondria-de Vries 1994; as well as Hulstijn 2001 and Nation 2001, who found that the use of word lists seems to exhibit good retention and faster gains. In fact, "there are 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).

Using lists and cards also facilitates self-directed learning and learner autonomy, as learners may work at their own pace. It does, however, require motivated and disciplined learners, who should also be able to deploy the right metacognitive strategies for self-monitoring, planning their own learning, etc., since "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" (Nakata 2008:7).

2.2 What word lists are there?

If using word lists and cards can be a useful tool for dedicated L2 vocabulary learning, the next question is if such lists are already available. And if so, how good are they? Might the KELLY lists improve on what is currently available? In this section we review the lists in existence for the languages of the project, except English, which has been mentioned above.

Arabic

At the time of the start of the KELLY project, no Arabic word lists or corpora could be found and so a new, internet-based corpus was produced for the purpose of the project. However, during the course of the project, *A Frequency Dictionary of Arabic: Core Vocabulary for Learners* was published (Buckwalter and Parkinson 2011). An excellent resource for learners, it contains the 5,000 most frequently used words in Arabic. It is just over half the size of the final 9,000 word KELLY list for Arabic, but also contains dialectal Arabic words, which were largely removed from the KELLY list in line with most programmes teaching Arabic as a foreign language, which teach Modern Standard Arabic (MSA). In terms of structure, the frequency dictionary is strictly ordered by word frequency, containing smaller thematic lists and an alphabetical index. In the KELLY list, the word frequency order has largely been kept, but in line with the wider KELLY project aim, relevance to L2 learners overrode frequency and irrelevant items were omitted or moved within the list. For example, numbers were included as a category, irrespective of individual numbers' frequency in the corpus. Vocabulary items seen as essential to language learning with few or no occurrences were added through comparison with other language lists—for example names of foods and items of clothing that appeared on several of the other language lists, but not in the Arabic list. Conversely, vocabulary items that did not fit into the CEFR levels and would

seem out of place in a language learning environment were omitted, such as heavily religious vocabulary items.

Chinese

Interest in producing Chinese frequency lists is amplified by the unique need to arrange a very large inventory of characters in a way that is useful for language learners. One of the first corpus-based frequency lists for Chinese was produced in the 1920s from a corpus of more than 500,000 words (Xiao et al. 2009). This research line continued in the 20th century culminating in *A Frequency Dictionary of Mandarin Chinese* (Xiao et al. 2009). Like the Arabic dictionary from the same series mentioned above, it is a very useful resource for language learners, although it is based strictly on frequency and does not group words into thematic categories.

Greek

There are some word lists available for Greek, mainly created and used for language learning purposes (Charalabopoulou and Gavrilidou 2011). The first, provided by the Center for the Greek Language, which has exclusive responsibility assigned by the Greek Government for the organisation, planning, and administration of examinations for the Certification of Attainment in Modern Greek, includes two word lists, simply described as "Indicative Vocabulary for Levels A & B" (Efstathiadis et al. 2001). The lists are not corpus-based and the number of lemmas is not specified.

The second wordlist is found in an appendix to the curriculum for teaching Modern Greek as an L2 to adults published by the University of Athens, and is based solely on the authors' intuition and teaching experience. The authors believe the words are "representative vocabulary", and comply with the communicative needs and learning goals specified in the curriculum in relation to particular notions and functions, speech acts and thematic domains. The number of words is not specified (University of Athens 1998).

Thirdly, a dictionary of Greek as a foreign language[6] has recently been produced as part of the Education of the Muslim Minority Children in Thrace project, as part of the Programme for the Education of Muslim Children 1997–2008.[7] The dictionary includes 10,000 lemmas arrived at through combining existing monolingual dictionaries for Greek schoolchildren, representing basic/core vocabulary items, and e-corpora, including school textbooks.

Lastly, three different but complementary corpora were created as part of the research project 'Corpora in Modern Greek Language Research and Teaching', co-funded by the European Social Fund and National Fund (EPEAEK I) (Mikros 2007): a general corpus of Modern Greek, a special corpus for teaching Modern Greek as a foreign language, and a corpus of material produced by learners. Various word lists were produced from the corpora in order to study high and low frequency vocabulary usage in various Natural Language Processing applications.

[6] http://www.museduc.gr/docs/gymnasio/Dictionary.pdf.

[7] http://www.museduc.gr/en/index.php.

Italian

The *Lessico di frequenza dell'italiano parlato (LIP)* [Frequency Lexicon of Spoken Italian] is one of the most important collections of texts of spoken Italian and one of the most widely used in linguistic research. It was composed by a group of linguists led by Tullio De Mauro who used it to build the first frequency list of spoken Italian (De Mauro et al. 1993). Its 469 texts, containing a total of approximately 490,000 words, were collected in four cities (Milan, Florence, Rome and Naples), and comprise face-to-face and mediated dialogues and monologues.

The *Vocabolario di Base della lingua italiana (VdB)* [Basic Vocabulary of Italian], also by De Mauro, is a 7,000 wordlist drawn up with mainly statistical criteria and appears in the *Guida all'uso delle parole* [Guide to the Use of Words] (De Mauro 1997). It represents the part of the Italian language used and understood by most Italians. It includes the first 4,700 words in the *LIP* (Bortolini et al. 1972) with a further 2,300 frequently used words mainly sourced from widely-used Italian dictionaries. The words in the *VdB* are grouped into three levels: fundamental vocabulary (from the *LIP*), high-use vocabulary (also from the *LIP*) and high-availability vocabulary (those words sourced from dictionaries).

The *VdB* was the first work of this kind in Italy and is now widely used, for example to monitor and improve the readability of a text according to scientific criteria.

Two centres for teaching Italian as a foreign language, the Università per Stranieri di Perugia and the Università per Stranieri di Siena, were contacted and replied that there are no official word lists for assessing students' knowledge of Italian or for preparing teaching material. However, the most used frequency lists for deriving lexical syllabi are the *LIP* and *VdB*. Both centres have developed lists of words most used by learners based on speech produced by L2 students of Italian at different levels.

Norwegian

Although no official word list could be found, several word lists exist for Norwegian in textbooks for learning Norwegian as a foreign language. However, it is unclear how these word lists were formed.

There is also *Lexin*,[8] the online series of bilingual dictionaries (Norwegian-minority languages) with 36,000 entries, based on the Swedish version (see below). It includes a series of illustrations divided into 33 topic areas such as family and relatives, our bodies outside, the human body inside, mail and banking, and school and education.

Polish

No official or otherwise widely-used word list was found.

Russian

Early modern frequency lists from the 1950s and 1960s are available for Russian (Josselson 1953; Shteinfeld 1963), as well as a later dictionary (Zasorina 1977) produced from a one-million-word corpus. However, Russia's turbulent history in the past 50 years has resulted in substantial changes in the Russian lexicon, which are not reflected in these early lists.

[8] http://decentius.hit.uib.no/lexin.html.

Corpora since then have expanded significantly with the increase in the number of texts available in electronic form.

Further development of the KELLY list for Russian led to a frequency dictionary in the same series as those referred to above for Arabic and Chinese (Sharoff et al. 2013), with corpus examples and their translation into English, topical word lists, and information on the frequency of multiword units.

Swedish

For Swedish there are a number of word lists available. The oldest and most famous is Sturé Allen's *Tiotusen i topp* [Top ten thousand; Allen 1972]. It was produced using newspaper texts collected around 1965, and has not been updated.
Other leading resources include:

Svensk skolordlista [Swedish wordlist for schools], with 35,000 words, is the outcome of a collaboration between the Swedish Academy and the Swedish language board. It is aimed at pupils in the 5th grade and higher, and contains short explanations in simplified Swedish for most words. It is a selection from the SAOL (Swedish Academy's Wordlist of Swedish Language) and is updated regularly, with approximately 125,000 words. It reflects the most frequent vocabulary in modern newspapers and books, and includes a number of colloquial words. However, no frequency information is provided.

Lexin Svenska ord med uttal och förklaringar[9] [Lexin Swedish words with pronunciation and explanations] contains 28,500 words and is aimed at immigrants. The vocabulary has been selected using frequency studies, vocabulary from course books, words specific to social studies (partly manually selected and partly from specific interpreter lists), and colloquial and/or 'difficult' vocabulary items taken from a range of sources (Gellerstam 1978). It is regularly updated from corpus studies, though there are no frequencies or information on the vocabulary appropriateness for different learner levels.

The Base Vocabulary Pool[10] (Forsbom 2006) is a frequency-based list constituting central vocabulary derived from the SUC (Stockholm Umeå Corpus). The base vocabulary pool is created on the assumption that domain- or genre-specific words should not be in the base vocabulary pool. The core of this list is constituted by stylistically neutral general-purpose words collected from as many domains and genres as possible. Out of 69,371 entries in the lemma list based on SUC, 8,215 lemmas are included in the base vocabulary pool.

3 Preparing the KELLY lists

The KELLY lists aim to reflect the contemporary language, constitute the most frequent core vocabulary and are based on objective selection unless dictated by pedagogical needs.

[9] http://lexin.nada.kth.se/.

[10] http://stp.lingfil.uu.se/~evafo/resources/basevocpool.

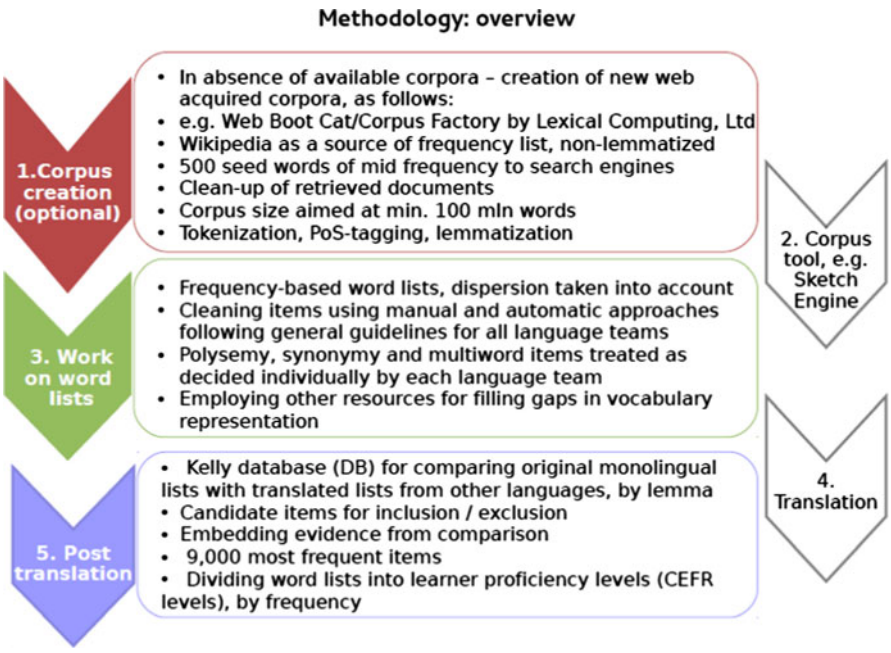

Fig. 1 Methodology overview

The corpora they are based on should be large enough, and comprise enough different documents from a range of domains, to minimise the risk of words of specialised vocabulary appearing in the lists. We used the same methodology to create the corpora for each of the nine languages, so that the respective word lists could be, as far as possible, comparable.

Work on the lists was divided into five distinct phases, as outlined in Fig. 1.

We will now walk the reader through these phases, step by step.

3.1 Identify/create the corpus

For each language, we needed a corpus. We wanted it to be a corpus of general, everyday language and we wanted it to be large, with enough different texts so that it would not be skewed by particular texts or topics, and so that it would not miss any core vocabulary. Moreover, we wanted the corpora of the different languages to be, as far as possible, 'comparable': we wanted all the lists to represent the same kind of language, so we could make connections between them.

For some languages there was a good choice of corpora available, but not for others. Spoken corpora were only available for a minority of the languages.

One corpus type that is available or can be created for most languages, and which does provide a large general corpus, is a web corpus, using methods as presented in Sharoff (2006) and Baroni et al. (2009). These papers also show that web corpora can represent the language well—in some regards, better than a corpus such as the

Table 1 Main corpora and processing tools for each language

Language*	Name	Size in tokens (m)	Processing tools
Arabic	Internet-AR	174	Sawalha and Atwell (2010)
Chinese	Internet-ZH	277	From Northeastern University, China
English	UKWaC	1,526	TreeTagger
Greek	GkWaC	149	ILSP tools
Italian	ItWaC	1,910	TreeTagger
Norwegian	NoWaC	700	Oslo–Bergen tagger
Polish	Polish web corpus	128	TaKIPI, Piasecki (2007)
Russian	Internet-RU	188	Sharoff et al. (2008)
Swedish	SwedishWaC	114	Kokkinakis and Johansson Kokkinakis (1997)

* The corpus was, as far as possible, Modern Standard Arabic only

BNC, which has a heavier weighting of fiction, newspaper, and in general the more formal and less interactive registers. For each of the languages, we had access to or created a web corpus using the methods described by Sharoff and Baroni et al.

A central question was: what should the list be a list *of?* The most basic option was word forms, so *invade invading invades* and *invaded* would all be separate items. This was at odds with usual practice, and not useful for learners (especially for highly inflectional languages like Russian, Polish, Greek and Arabic), so we needed to lemmatise the corpus: to identify, for each word, the lemma. We also decided that the list items would all be associated with a word class (*noun, verb* etc.) with *brush (noun)* and *can (noun)* treated as distinct items from *brush (verb), can (verb)* and *can (modal).* For this we needed a part-of-speech tagger.

Table 1 shows that the corpora are comparable in terms of the source of texts (web-acquired), and all very large. Some random sample analysis of corpus texts and the most frequent nouns/verbs/adjectives, as well as an overview of hapax legomena in the Swedish corpus, SwedishWaC, indicated that its text constitution is very much like that of the English corpus, UKWaC, and that the majority of texts are made up of newspaper texts, Wikipedia articles, forums, chats and blogs (Volodina and Johansson Kokkinakis 2012). It also allows us to hypothesise about the dominating text genres in other web-acquired corpora collected in the same way.

3.2 Generate a frequency list

The processed corpora were then loaded into corpus tools, such as the Sketch Engine (Kilgarriff et al. 2008) or the University of Leeds installation of the Corpus WorkBench. These tools both support the preparation of word lists, lemma lists, or, as we wanted here, lists for lemma + word class, all with frequencies attached. They also allow the user to easily view the underlying data, the 'corpus lines' i.e. the context in which each word originated, for any item in the list, to check for, for example, lemmatisation and POS-tagging errors and other anomalies.

For each language, we took the 6,000 most frequent lemma + word-class pairs, and this was the M1 list, as the input to the next process. (This number is lower than

the target 9,000 because we expected the next steps to add many more items than they deleted, as they largely did.)

3.3 Clean up the list, and compare with lists from other corpora and other wordlists

3.3.1 Clean up

This step consisted of a series of procedures to 'clean up' the list, delete anomalies, correct errors (in particular word class errors) and to check against other lists for omissions. The process would make each team aware of the idiosyncrasies of their corpus so that, where possible, these could be mitigated by the integration of other data. The cleaning process included the following:

- Checking unexpected inclusions to see whether they were errors. For instance *top* as an English verb appeared in the list because of numerous mis-tagged examples of 'back to top' in our internet-derived corpus. Similarly, various lemmatisation errors were identified, for example the entry *ty*, which turned out to be an incorrect formation from *ties*, which should have been *tie*
- Checking unexpected verb uses which are more usefully coded as adjectives, e.g. English *neighbouring* rather than the verb *neighbour* or Polish *zróżnicowany* ('various') which was lemmatized as the verb *zróżnicować* ('vary')
- Amalgamating variant spellings such as *organise* and *organize*, and the Greek *αυγό* and *αβγό* ('egg'), so that their frequency is not distorted by being divided
- Merging and splitting, as necessary, aspectual variants of verbs and reflexive verbs, often mis-lemmatised, such as Polish *opłacać się* ('be worthwhile') versus *opłacić* ('pay for')

To promote consistency between language teams, a list of word types for inclusion was drawn up at the outset. This included decisions on abbreviations, proper nouns, dialect words, affixes, inflections, hyphenated words, trademarks and others. The guidelines are attached as Appendix 2.

3.3.2 Polysemy, multi-word units

Two central issues for creating word lists are polysemy and multiword units. The problem with polysemy is this: if a word has two meanings, for example the word *calm* in 'a calm mind' and 'calm water', then it is not useful for a learner (or translator) to include the word in a list without indicating which meaning is intended. An immediate response might be "let's make it a list of word senses". This strategy has two difficulties, one theoretical and the other practical. The theoretical one is that there is no agreement, and is never likely to be, about what the word senses for each word of a language are (Kilgarriff 1997). The practical one is that we cannot count word senses: 50 years of research in automatic Word Sense Disambiguation has not delivered programs which can automatically say, with a reasonable level of accuracy, which sense a word is being used in.

It is appealing to make a distinction between homonymy, where two words share the same form (and are likely to have different translations), such as a linguistic *sentence* and a prisoner's *sentence*, and polysemy. For homonyms, learners have two words to learn; for polysemous items, usually one. The difficulty is in drawing the line. Because of this difficulty, we largely adopt Lyons's 'maximise polysemy' position (Lyons 1977: 554), as also taken in recent English learners' dictionaries (Rundell 2007; Turnbull et al. 2010).

The problem with multi-word units like *according to,* is similar. It certainly makes more sense for learners and translators to see *according to* in the list than to see a high frequency for the word *according* (or, worse, the verb *accord).* But *according to* is a clear case; what about the many hundreds of compounds, phrasal verbs, idioms and other fixed expressions? The first problem, again, is the theoretical one: what is the list of items we should count? The second is the practical one: how do we count them, without getting many false positives and distortions where, for example, we do not know what frequency to give to *look* because so much of the *look* data is taken up by *look at, look into, look up, look for, look forward to,* etc.?

Different language teams took different strategies on these two issues. Some, including the ones for English and Swedish, took a hard line: we cannot count word senses or multiword units reliably, so we shall have a plain list of simple words (in all but the most vivid cases, such as the English *according to, united* in *United States*).

Others, notably the Polish team, took a more translator-friendly position, splitting homonyms and giving sense indicators for each. For example the Polish noun *agent* was split into two senses: (1) 'representative', glossed for translators in Polish as 'przedstawiciel', and (2) 'secret operative', glossed as 'wywiadowca'. A sense indicator was also added even if only one sense was included, but we wanted to make sure translators would not get sidetracked by another, rarer sense. So, although the original meaning of the Polish *izba* is 'room', this sense is quite rare in contemporary Polish, and we did not want it covered. Instead, we wanted the dominant contemporary sense of 'parliamentary chamber', so a gloss was added saying 'parlamentu, urzędu'. In addition, multi-word items were included as separate entries as long as their frequencies (estimated manually in each case from the reference corpus) met the threshold criteria of simplex items. For example, another common occurrence of *izba* was in the combination *izba przyjęć* 'hospital admissions unit', and so this multi-word item was entered separately.

Similarly the Arabic team's approach was to separate homonyms in the Arabic list that could have multiple, unrelated meanings depending on their vocalisation, either by adding as separate items and vocalising to distinguish their meanings, or adding as separate items with a comment describing the word as, for example, either a noun or verb. For example the Arabic word شعر which appeared with no vocalisation in the Arabic corpus, was added as the three separate vocalised items: شَعْر (hair), شِعْر (poetry) and شَعَرَ (to feel). The order that the vocalised words appeared in the list was determined by the frequency of their respective occurrences, which was determined by looking at the contexts in which the unvocalised شعر appeared in the corpus. On the other hand, verb/noun forms such as the word شك (to doubt/doubt) for example, were left unvocalised and instead a comment was added to clarify whether it was to be used in the noun or verb form. If one form had a high frequency and the

Table 2 Translation equivalents across languages

Language	Translation of Swedish *rom*	Meaning in English
Arabic	الروم: مشروب كحولي يقطر من عصير قصب السكر؛ بطارخ السمك	(1) rum (drink); (2) caviar
English	rum; roe	(1) rum (drink); (2) caviar/roe deer
Greek	αβγοτάραχο	roe deer
Italian	uova di pesce;, rum	(1) caviar; (2) rum (drink)
Norwegian	rom	rum (drink)
Polish	ikra	roe
Russian	ром	rum (drink)

other a low frequency, the high frequency sense would be highlighted and the low omitted. Obvious multiword units with high frequencies such as الشرق الأوسط (the Middle East) were lemmatised as such.

The hard line approach taken by the English and Swedish teams was motivated by two considerations: firstly, the process becomes more automatic, faster and more reliable; and secondly, it makes it easier to identify one-to-one mappings between different languages and to expand polysemous items after translation into the different target languages. Some of the disambiguation decisions were therefore left to the translators. An example is the word *rom* in the Swedish list, which can mean rum, caviar, gypsies, roe deer, or Rome. In all cases the noun is of a non-neutral gender and, except for the 'roe deer' meaning, is used without articles.

The rule of thumb for translators was to use the most frequent alternative and to keep in mind that the lists are intended for language learners. On that basis, translations were provided for the *rom* as in Table 2.

According to the given translations, the most common equivalents for the Swedish "rom" in the other languages are rum, caviar and roe deer; none of the translators offered Rome or gypsies. The translators into Norwegian and Russian have shown a good sense of humor in choosing the alcoholic drink as the most relevant sense for language learners. Clearly the translated items cannot be used as translations of each other without human processing.

3.3.3 Points of comparison

We quickly realised that everyday items (e.g. *mummy*, *bread*) were under-represented or sometimes missing in the first list, while administrative and technical items (e.g. *sector*, *review*) were over-represented.

For a subset of the languages (English, Norwegian, Italian and Polish) we were fortunate in having at our disposal spoken corpora (or subcorpora), including records of everyday informal speech, against which we could run comparisons. For English, for instance, we used the conversational-speech part of the British National Corpus (BNC-sp). We ran a comparison to identify all the words which had at least 50 occurrences in BNC-sp, and were either not in the M1 list or had much higher normalised frequency in BNC-sp than M1.

We wanted the final list to be ordered by usefulness for language learners. In straightforward cases we could simply use UKWaC frequency for sorting, but it was not clear how words which were added in would be sorted, or how any other manual interventions would interact with the sorting. We decided to use a points system, as follows:

The original list was divided into six equal groups and allocated points, with six for the most frequent group descending to one for the least frequent. BNC-sp words were added on the following principles:

- The most frequent 100 words from BNC-sp were given 5 or 6 points
- 100–200: 4 or 5 points
- 200–400: 3 or 4 points
- 400–600: 2 or 3 points

The variance in points allowed a small amount of judgment as to the overall generality and usefulness of the word. Points were then deducted: (1) for informal, (2) for taboo or slang, (3) for old fashioned. Any words on the UKWaC list that did not occur at all in BNC spoken had one point deducted.

We then looked at a keyword comparison between UKWac and BNC spoken, in which words were sorted according to the ratio of their frequencies in the two corpora (Kilgarriff 2009). For keywords of BNC-sp versus UKWaC and *vice versa,* adjustments were made using a points system, so that words such as *sector* and *review*, which originally had 6 points, were demoted, and words such as *bread* were promoted.

For a number of very restricted sets, such as numbers, compass points and days of the week, points were assigned to ensure consistency. This is because it would be unhelpful to language learners to see such items at different levels. Some proper nouns were also included, based on the corpus, but it was felt necessary for teams to use some judgment. In particular, teams were asked to privilege words which did not come from their own geographical area, since these were more likely to be of universal importance. So, for instance, for the English list, a word such as *Mediterranean* would be deemed to be of more importance than *Cornwall.* The additional resources (corpora and word lists) used for each language are listed in Appendix 1.

3.4 Translate each item into all the other KELLY languages

Once each team had prepared its updated M2 lists, these were sent to a team of translators. Each of the nine lists was translated into each of the eight other languages, in 72 translation tasks giving 72 translation (T1) lists.

Translators were asked to choose the core translation for each word and to make sure that the translation was equivalent in word class and register. They were encouraged to give single-word translations, and only one translation, where this was viable, though they should give multiword translations and/or multiple translations if this seemed the only sensible thing to do. Each team prepared instructions to deal with specific aspects of their language: for example, should the

translation include word class (not relevant for Chinese, where word class is a problematic concept) and should the translated noun's gender and declension class be given, and if so, how.

The work was subcontracted to a translation agency. There were, in some cases, several iterations, with KELLY project members who knew both languages for a list assessing the quality and sending it back for re-translation if the quality was not high enough. Translations were returned for re-translation or additional proofreading if any mistakes were discovered in a random sample check of 150 words. Typical errors found at this stage included:

- spelling mistakes, e.g. *ecyklopedi* for *e**n**cyklopedi;* (Eng. encyclopedia)
- lemmatization mistakes, e.g. *dumhet**er*** (plural) for *dumhet* (singular) (Eng. stupidity)
- incorrect translation, e.g. Swe–Rus <*förvåning, сюрприз*> when it should have been <*förvåning, удивление*> (Eng. surprise)

The output of this stage was a rich dataset of 72 T1 lists, each of around 6,000–7,000 translation pairs and additional information relating to word class, frequency, points, sometimes sense indicators, translator notes and so forth.

3.5 Use the 'back translations' to identify items for addition or deletion

By 'back translations' for a language, e.g. Italian, we mean those words used by translators when translating into Italian. It seemed likely that some words that were wanted in the final list but were not in the M2 list, and some high-salience multiword units, would occur frequently as back translations.

We simplified all rows in T1 lists to plain lemma-translation pairs. This involved a number of iterations to ensure all items which should match, as they were essentially the same word although they came from either the M2 list or one of eight translator's files, did match. To support the process we threw away word-class information: word classes often did not match across languages, e.g. Swedish *numerals* versus *determiners* in Norwegian. We then built a database of the resulting pairs.

The database was used to prepare three lists for each language: single-word candidates for inclusion, multiword candidates for inclusion, and candidates for exclusion/demotion.

- **Single-word inclusions:** each team was given a list of items that occurred as back-translations, but were not in their own list. These were incorporated according to a points system based on the number of lists in which they occurred as translations. So, for instance, for English, words such as *wolf*, *torture, mayor, earthquake* and *institute* were not in the original list, but occurred frequently as translations, so they were added.
- **Multi-word inclusions:** phrasal verbs and other phrases had not been included in the original lists because of the difficulty of identifying them automatically. It was hoped that these would emerge as translations of other languages. Items

such as *take out, of course, for example* and *take place* were identified in this way.

- **Demotion/deletion:** conversely, words such as *align, arguably, broker* and *bungalow*, were in the original list but did not occur once as translations from other lists. These were therefore considered for deletion or demotion.

After the inclusions from the translated lists, some key words for language learning still had not appeared on some of the lists—words such as *orange*, *elbow*, *banana* and *alphabet.* So, a set of common key 'domains' was created based on the CEFR themes and 'can do' statements. Each domain was then populated independently for each language. The domains for all languages were:

1. calendar: days, months, time, celebrations
2. city facilities
3. clothes
4. colours
5. computer terminology
6. cutlery, crockery, cooking equipment
7. directions, including compass points
8. emotions
9. family relationships
10. food and drink
11. grammar and punctuation
12. jobs
13. nature: animals/insects/birds/plants
14. numbers
15. parts of the body, as well as health and medicine
16. religion
17. rooms and furniture
18. school life and subjects
19. shapes
20. shop transactions
21. sport and leisure
22. travel
23. weather
24. weights and measures

Ensuring that certain 'closed' sets were included, such as calendar days and months, compass points and numbers for example, resulted in resolving earlier discrepancies in the lists. For example, the previously mentioned high frequency of some of the days of the week but not others meant that some days of the week may have been included in a list while others may not have been. The domain approach allowed each list to be populated with all of the days of the week. This is an instance where learner-centeredness overrode frequency in the lists.

For 'open' sets, such as food and drink, and parts of the body, frequency was referred back to and higher frequency words were chosen over lower frequency ones, even where the overall frequency was low.

Then, after many extra rounds of editing and checking, each word on the new M3 lists was assigned to a CEFR level, using the level descriptions and 'can do' statements as a guide. This allowed the several thousand words on the lists to be broken down and become more useful for language learners.

At last, the final M3 lists were handed over to our commercial partner Keewords who engaged in producing electronic word cards from them.

4 The KELLY database

The KELLY database is an interesting object. For each of nine languages, for each of around 9,000 words,[11] it contains translation mappings to one or more words in each of the other eight languages. With 74,258 lemmas and 423,848 mappings, it is large. We are not aware of any other comparable resources. While it has many limitations, which are apparent from its method of construction as detailed above, it can supply data for many research questions.

We did not want to miss matches between languages because they were given different grammatical labels, or (for the European languages) different capitalisation. So we left out grammatical class information, and the database is a database of lemmas rather than <lemma, word class> pairs, all normalised to lower case.

The database, as discussed here and as accessible on-line, is the version of the data after the various iterations of list-translation but before the processes that then finalised the word cards. Thus errors and problems identified have not, in the database version, been corrected.

4.1 Symmetric pairs (sympairs)

A basic construct for fathoming the database is the symmetric pair (hereafter *sympair*). 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*.

An example of a symmetric pair is English–Swedish <*regard, betrakta*> and Swedish–English <*betrakta, regard*>. One translator chose *betrakta* for *regard* and the other, independently, chose *regard* for *betrakta.* Likewise for the Greek–English pair <*λίμνη, lake*> and the English–Greek <*lake, λίμνη*>.

A naïve theory of translation might expect most words to come in symmetric pairs. The actual numbers of sympairs, for each language pair, is given in Table 3 (top right, above the leading diagonal). The percentages, also given in the table, are computed as the number of sympairs for a language pair divided by the maximum number there could have been, which is the smaller of the two numbers for the total number of words for the two languages. The total number of words for each language is given in the last row ("list length").

[11] These are lemmas, as discussed above. As the simpler word *word* will introduce no ambiguity, we shall use that throughout this section.

Table 3 Sympairs (top right triangle) and oto-sympairs (bottom left triangle) by language pair

	English	Polish	Italian	Swedish	Chinese	Arabic	Russian	Greek	Norwegian
English		2,863	2,896	2,983	1,574	822	2,526	2,594	2,298
		37.9 %	42.1 %	39.5 %	20.8 %	10.8 %	33.4 %	34.3 %	30.4 %
Polish	1,147		2,342	2,423	945	1,189	2,614	2,461	2,443
	15.1 %		34.1 %	28.7 %	12.2 %	14 %	29.2 %	32.5 %	28.8 %
Italian	1,331	1,198		2,632	1,015	1,059	2,103	2,164	2,366
	19.4 %	17.4 %		38.3 %	15.4 %	15.4 %	30.6 %	31.5 %	34.4 %
Swedish	1,308	1,253	1,163		1,109	617	2,270	1,954	3,109
	17.3 %	14.8 %	17 %		14.3 %	7.3 %	26.9 %	25.8 %	36.9 %
Chinese	390	284	236	315		608	979	726	600
	5.1 %	3.6 %	3.4 %	4 %		7.9 %	12.6 %	9.3 %	7.7 %
Arabic	383	340	323	247	164		1,451	966	916
	5 %	3.9 %	4.6 %	2.9 %	2 %		16.5 %	12.7 %	10.4 %
Russian	1,050	1,620	1,142	1,308	376	399		2,192	2,114
	13.9 %	19.2 %	16.8 %	15.5 %	4.8 %	4.4 %		9 %	23.6 %
Greek	690	962	1,139	941	206	329	957		1,377
	9.1 %	12.7 %	16.3 %	12.5 %	2.7 %	4.32 %	12.7 %		18.2 %
Norwegian	1,074	1,307	1,148	2,338	217	273	1,128	673	
	14.2 %	15.5 %	16.8 %	27.7 %	2.8 %	3 %	12.6 %	9 %	
List length	7,549	8,459	6,867	8,425	7,730	8,744	8,940	7,553	8,942

These numbers are low. In a simple world, sympairs would account for a large share of translations and percentages would approach 100. In practice, the fractions range between 42.1 % (English–Italian) and 7.3 % (Swedish–Arabic).

Note that the definition of symmetric pairs does not exclude *a* having another translation into B in addition to *b*, or *b*, into A. Thus English *room* translates into Italian *camera,* and Italian *camera* translates back into *room*—but Italian *spazio* also translates into English *room.* <*room, camera*> form a sympair, but not an entirely straightforward one because one of the words has another translation too. A more constrained construct is the one-translation-only (*oto*) sympair, where neither *a* nor *b* has any other translations into the other's language. Thus <*spazio,* 空间> form an oto-sympair, because *spazio* translates into Chinese as 空间, and has no other translations into Chinese, and 空间 translates into Italian as *spazio* and not as anything else.[12] We might expect this constraint to set aside the polysemous words. Numbers for these are in the bottom left triangle of Table 3 (below the leading diagonal).

[12] In the online database at http://kelly.sketchengine.co.uk, words which are oto-sympairs with the input word are coloured red, and other sympairs, green.

4.2 Cliques

A further construct of interest is the clique.[13] A clique is where, for words $<a, b, \ldots n>$ of languages A, B, … N, all pairs $<$(a, b), (a, c), … (a, n), (b, c), … (b, n) … $>$ are sympairs. An example of a three-language, English–Italian–Polish clique is $<$*cat, gatto, kot*$>$, since English *cat* translates into Italian and Polish as *gatto* and *kot; gatto* translates into English and Polish as *cat* and *kot;* and *kot* translates into English and Italian as *cat* and *gatto.*

For cliques as for sympairs, we can have or not have the one-translation-only (oto) constraint. Figures are given, with and without oto, in Table 4.

There are just five nine-language cliques in the whole dataset (Table 5). There are no nine-language oto-cliques and just four eight-language ones (Table 6).

Some of these are cognates, with Greek playing a particular role. *Guitar*,[14] in each language, can be traced back to the Greek original. (The Arabic cognate would be there too except its frequency was not sufficient to put it in the Arabic source list.) For *music* this is true for all but Chinese, and for *theory* and *tragedy,* for all the European languages. For *sun,* the link goes back to Proto–Indo–European (Huld 1986).

Table 4 Numbers of cliques and oto-cliques, for different number of languages

No. of languages	No. of cliques	No. of oto-Clique
3	55,023	14,211
4	35,146	6,413
5	16,048	2,204
6	4,980	520
7	975	71
8	106	4
9	5	0

Table 5 The five 9-language cliques in the dataset

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	musik
شمس	太阳	sun	ήλιος	Sole	sol	Słońce	солнце	sol
نظرية	理论	theory	θεωρία	Teoria	teori	teoria	теория	teori

[13] Terminology from graph theory, where a fully-connected subgraph such as this is called a clique.

[14] We represent each group by its English-language member, as that will indicate the group to most readers.

Table 6 The four 8-language oto-cliques in the dataset

Arabic	Chinese	English	Greek	Italian	Norwegian	Polish	Russian	Swedish
	吉他	guitar	κιθάρα	Chitarra	gitar	gitara	гитара	gitarr
ملكة		queen	βασίλισσα	Regina	dronning	królowa	королева	drottning
	三十	thirty	τριάντα	Trenta	tretti	trzydzieści	тридцать	trettio
مأساة		tragedy	τραγωδία	Tragedia	tragedie	tragedia	трагедия	tragedi

The concepts represented by many-language cliques are of interest, as they are lexicalised in a stable way across languages; one could even propose the method as a way of seeking out universals.

The 51 English words featuring in 8- and 9-language cliques are:

bank bed bomb book bread bridge chair channel church climate coffee dog eye father fish forest future government guitar heart horse hospital kitchen knee level library logic marriage milk music office pocket prison problem psychology queen revolution sand snow source sun system tea ten theory thirty trade tragedy university water week

Word class is not a construct in the database, since <lemma, word class> pairs were reduced to lemmas to avoid mismatches due to non-matching word class inventories. Nonetheless it is apparent that these are all nouns, with the possible exceptions of *future* (also an adjective) and *ten, thirty* (depending on whether numbers are seen as a distinct word class to nouns). The two numbers are in the list but other numbers are not.

Institutions are well-represented: we have eight (*bank, church, government, hospital, library, office, prison, university*, or nine if we include *marriage*). The natural world provides six (*climate, forest, sand, snow, sun, water*), edibles and drinkables, four (*bread, coffee, milk, tea*), animals and body-parts, three (*dog, fish, horse; eye, heart, knee*), and people and furniture, two (*queen, father; bed, chair*).

The 211 English words featuring in 7-word cliques but not in 8- or 9-langauge ones are given in Appendix 3. In addition to contributing further members to the groupings mentioned above, they introduce verbs (*believe, have, hope, read, sleep, write*), adverbs (*almost, already*), adjectives (*big, blind, central, clinical, green, industrial, mathematical, national, nervous, new, philosophical, single, theoretical, tragic, typical*), nationalities (*French, Italian*), months (*February, July, June, November*) and days of the week (*Saturday, Sunday, Thursday*); one can't help wondering what happened to *Monday, Tuesday, Wednesday,* and *Friday*. (As can be seen, allocation of words to word classes is problematic, as, for example, *hope* may be a noun as well as a verb; the analysis here is indicative only.)

In Appendix 4 we present the 33 seven-language oto-cliques (that do not share more than three words with either of the tables above), and in Appendix 5, the 49 eight-language cliques (that do not share more than three words with either of the

tables above or the first table in the appendix).[15] Near-duplicates are a complication: if one language has two words for a concept that is otherwise largely stable, the outcome may be two cliques sharing most words.

4.3 Non-sympairs: why are words *not* in sympairs?

The translation pair $<a$ of language A, b of language B$>$, where *a,* in the source list for A, is a non-sympair if a is not given as a translation of b.

We first distinguish two kinds of non-sympair.

- **Non-sympair-non-source (NSNS)** One kind is where b is not in the source list for B. We can divide the non-sympair set (NS) for the directed language pair $<$A, B$>$ into those where the word in B is in the source list for B, and those where it is not. NSNS can be demonstrated by the Swe–Eng $<$*port, doorway*$>$ where *doorway* is absent from the English source list. Likewise Gr–Eng $<$*προϋποθέτω, presuppose*$>$, where *presuppose* is not included in the English source list.
- **Non-sympair-source (NSS)** The other case is where b is in the source list for B. An example of an NSS is Swe–Eng $<$*förlägga, publish*$>$: *publish* is in the English source list but gets the Swedish translation *publicera*. Another is the Greek–English pair $<$*σχεδόν, practically*$>$: *practically* is in the English-source list but gets the Greek translation *πρακτικά*.

Hapaxes are words that only appear once in the whole database, as the translation of one word of one other language only. They will form a subset of the target words in the non-sympair-non-source (NSNS) set. An example of a hapax is English *starve,* which occurs only once in the database, as the translation of Swedish *svälta.* It is not in the English source list, nor has it been provided as a translation into English from any other language. Another is English *deletion,* translation of Greek *διαγραφή* but not occurring otherwise.

Indirect routes (NSS-0, NSS-1, NSS-m; NSNS-0, NSNS-1, NSNS-m): A further question we may ask about non-sympairs is: can we get from a to b (or *vice versa*) via a third language: is there a word z in a third language Z, such that a translates as z (or *vice versa*) and z translates as b (or *vice versa*). There may be zero routes from a to b via another language, or there may be one, or there may be more than one. We shall call them the *0, 1*, m sets. To understand what these "detours" can look like, consider the following example of an NSS_1: we have the Swedish–English non-sympair $<$*egentligen, really*$>$, but then we can get back from *really* to *egentligen* via Greek, with Eng–Greek $<$*really, πραγματικά*$>$ and then Greek–Swe $<$*πραγματικά, egentligen*$>$.

The classification of types of translation pairs is illustrated in Fig. 2.

We investigated the directed-translation-pairs for eight of the seventy-two directed pairs: Arabic–English, Chinese–Russian, English–Greek, Greek–English,

[15] All tables order columns alphabetically by the English spelling of the language, and rows, by the spelling of the English word, or, if there is no English word, by the word in another Latin-alphabet language, taking the remaining four Latin-alphabet languages in alphabetical order: Italian, Norwegian, Polish, Swedish.

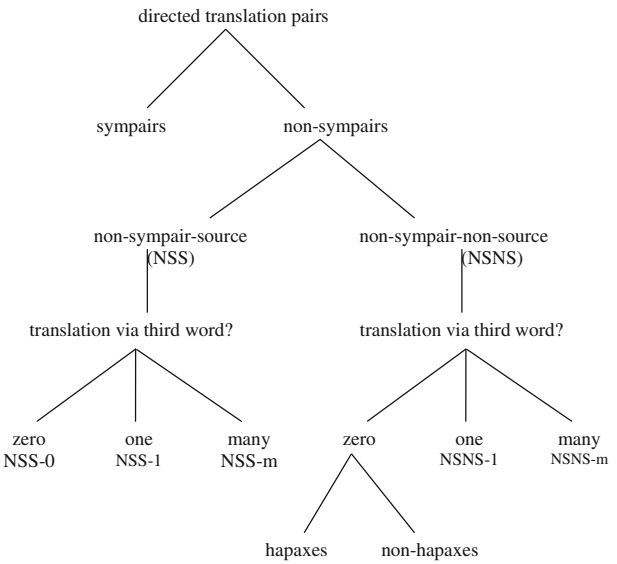

Fig. 2 Types of translation pairs in the KELLY database

Table 7 Analysis of non-sympairs

	Ara–Eng	Chi–Rus	Eng–Gre	Gre–Eng	Nor–Swe	Rus–Chi	Swe–Eng	Swe–Rus
NS	4,692	3,871	5,599	5,519	2,958	5,443	3,120	3,553
NSS	2,918	2,647	2,381	3,339	1,864	2,706	2,095	2,453
NSS-0	628	1,191	701	1,135	683	1,221	633	801
NSS-1	630	807	527	664	531	749	576	712
NSS-m	1,660	649	1,153	1,540	650	736	886	940
NSNS	373	328	1,923	554	81	1,155	214	295
Hapax	1,401	896	1,295	1,626	1,013	1,582	811	805
Other NSNS-0	286	262	594	355	36	303	103	106
NSNS-1	75	60	638	176	28	504	97	149
NSNS-m	12	6	691	23	17	348	14	40

Norwegian–Swedish, Russian–Chinese, Swedish–English and Swedish–Russian. We identified how many translation pairs there were in each category, and give the counts in Table 7.

4.3.1 Non-sympair analysis

We then took a sample of 100 non-sympairs for each language pair, for closer examination. The sample was a random sample, structured as follows (Table 8):

Table 8 Structure of sample for non-sympair manual analyses

NSS-0	NSS-1	NSS-m	Hapax	NSNS-0	NSNS-1	NSNS-m	Total
15	15	15	30	5	5	5	100

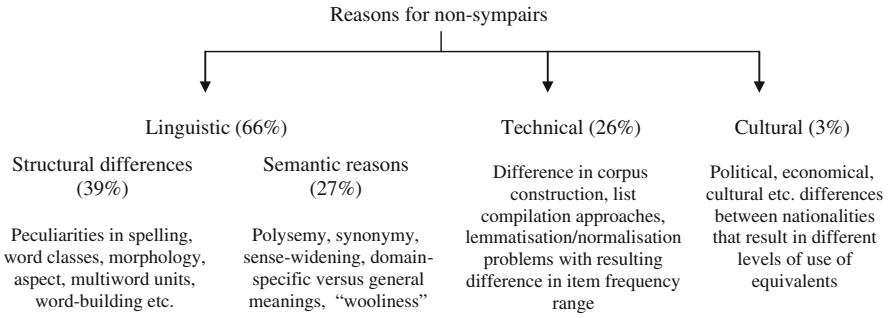

Fig. 3 Summary of reasons for non-symmetric translations

A team member who knew the two languages analyzed them for possible reasons why the directed pair $<a, b>$ was not a sympair: that is, why there was not a translation $<b, a>$ in the database. We identified several common reasons. Figure 3 provides a summary of the most important ones grouped according to their types. The numbers provided in brackets are averages, and indicative only (Fig. 3).

Translation is to an extent subjective in character, depending on the personality, skills and experience of translators. However, certain linguistic characteristics of individual languages make subjective choices made by translators objectively explicable, especially in projects like ours with words taken out of their contexts. The analysis confirmed our intuitions that "bad translation" was only occasionally the reason for non-sympairs, covering between 2 and 10 % of the sample, depending on the language. The most frequent reasons for non-symmetric translations proved to be either technical, i.e. due to differences in compiling the lists and corpora for deriving the lists, or linguistic, i.e. due to differences between the languages. Here we give descriptions and examples of cultural, technical and linguistic reasons.

Cultural

This group covers cultural, political, economical and other nation-specific mismatches: *a* denotes a salient concept in the culture of A-speakers but the concept is not present or is not as salient for B-speakers. Many hapaxes fall here:

- Vocabulary reflecting flora, fauna, or other "natural" phenomena specific for the A culture, e.g. Swe–Eng $<$*gran, fir*$>$: there are not so many fir trees in the UK
- Political reality not represented in B languages, e.g. Swe–Eng $<$*kommun, municipality*$>$; Swe–Rus $<$*republikan, республиканец*$>$ ('republican')
- Presence of geographic names specific to A-languages: Swe–Rus $<$*stockholm, стокгольм*$>$, Swe–Eng: $<$*nordisk, nordic*$>$

- 'Easter': the Swe–Rus language combination has <*påsk, пасха*>. The item *пасха* is not in the top 6,000 items of the Russian frequency list since the religious holidays were suppressed for 70 years under Soviet rule. We wait to see how this might change. <*teolog*, *теолог*> ('theologian') followed a similar pattern
- Swe–Rus <*färja*, *паром*> (Eng *ferry*)—this type of transportation is underused in Russia compared to Sweden
- In the Arabic list, a relatively high frequency of religious terms and phrases were found. Those with relevance to general language learners were kept on the list and a number of irrelevant terms were omitted. An example of a non-symmetric pair from the terms that were retained is the Ara–Eng <مصحف, *holy koran*>. The term *holy Koran* is not in the top 6,000 items of the English frequency list, nor are its equivalents *koran, quran, holy quran.*

Technical

The 'technical' reasons comprise the following types of mismatch:

- **Corpus differences:** *a* is only there because of a skew in the A corpus. An example is the political bias of the Swedish corpus which gives Swe–Eng <*marxist, marxist*> (hapax for English), <*ordförandeskap, chairmanship*>, <*feminist, feminist*>. The corpus for Arabic proved to have a bias towards religious terminology whereas the English gave a high number of medical texts.
- **Frequency** (often arising as a consequence of (a) above): *b* is not frequent enough to get into the source list for B. This is the default for NSNS and does not apply to NSS cases. In principle it may be because the source corpus either displays the relative unimportance of that concept for the speakers of the B language or that the corpus material has a bias towards some other topics and domains thus downgrading the concept to a lower frequency range. However, many cases are simply the result of marginal frequencies. If an item present in language A has a frequency that has given it a position at the bottom of the A list, whereas the item in language B has a frequency that has left it just outside the B source list there is little to be said. A Swe–Rus example is <*korsning, пересечение*> ('crossing'): the Swedish *korsning* has rank 5,725 of 6,000, whereas the Russian *пересечение* just missed the Russian list. Other examples of the "marginal" type are Swe–Rus <*skicklighet, ловкость*> ('skill'), <*smälta, таять*> ('melt'), <*bättra, улучшать*> ('improve'); Swe–Eng <*systematiskt, systematically*>, <*nyfikenhet, curiosity*>. **S**ome of the vocabulary absent in B languages but present in a number of other languages was identified during the post-translation phase and was added to the final monolingual lists for the B language.
- **List compilation differences:** e.g. part of speech taxonomy mismatches. Some language teams decided against having certain word classes in their lists which resulted in hapaxes in the B language, e.g. Swe–Rus <*varenda, каждый*> ('every'), <*själv*, *сам*> (reflexive pronoun). These items, though important and frequent in Russian, were not present on the original Russian monolingual list for translation since pronouns were not included in the list. They found their way into the final B list after the post-translation phase.

- **Lemmatisation/normalisation:** this was a particular issue for Arabic. Arabic verbs were lemmatised as the simplest form, the past simple (third person masculine). However, some verbs from other language lists were translated into the present simple. For example, تبيع as the present tense translation of 'to sell', rather than the past simple tense; يعطي ;باع rather than اعطى for 'to give'.

Linguistic

The group of linguistic reasons is split into two distinct subgroups:
The first subgroup covers **structural differences between the languages**:

- Difference in **word-building mechanisms**. Swedish and Norwegian exhibit compounding: they merge two root morphemes into one word. In the other languages such compounds have to be translated with multiword units which are seldom included as headwords in the source lists for B languages, e.g. Swe–Rus <*förfalla, приходить_в_упадок*> ('to fall due, degrade, delapidate'), <*kartlägga, наносить_на_карту*> ('to chart'), <*finansminister, министр_финансов*> ('finance minister'). Half of the Russian hapaxes found in Swedish–Russian pairs fall into this group.
- **Spelling and form variants** have also influenced the translation asymmetry. Many words can be spelled in several ways, all frequent and accepted, e.g. Swedish *utge*, *utgiva* ('issue'), which has then given rise to the non-symmetric translations Swe–Rus <*utge, издавать*>, Rus–Swe <*издавать*, *utgiva*>. Russian words containing "ë" that can also be spelt with "e", as in the case of Swe–Rus <*seg, жёсткий*> (Eng. *tough*) where instead of the spelling variant *жёсткий*, the Russian B-list contained the variant *жесткий*. The same can be observed in English, e.g. *med**iae**val* versus *med**ie**val* where Swe–Eng. <*medeltida, med**iae**val*> is not matched with the Eng–Swe <*med**ie**val*, *medeltida*>. Another case is full versus shortened forms of the same word, e.g. Swedish *bio* versus *biograf* (Eng *cinema*) which gave the non-symmetric pairs Swe–Eng <*bio, cinema*>, Eng–Swe <*cinema*, *biograf*>, or the English *photo/photograph*, which, with Greek, resulted in <*photo, φωτογραφία*> and <*φωτογραφία, photograph*>.
- **Aspectual differences:** In languages where aspectual difference is expressed lexically (as opposed to grammatically) there exist several variants of the same item for different aspects, e.g. Russian *пон**я**ть* and *пон**има**ть*, both translated as *understand* in English. The difference between the items lies in the semantics of the aspects—one having the meaning of a "completed action" (perfective aspect) and the other of an "action in progress" (imperfective aspect). The translators have been asked to use the imperfective aspect only in their translations. However, in some cases the members of the perfective/imperfective aspect pairs have different usage preferences and carry a slightly different denotation, so aspect normalisation was problematic. Thus, the Swe–Rus pair <förstå, понять> does not match the Rus–Swe <понимать, förstå>.
- **Homography across word classes**, where *a1* in language A is translated as *b* in language B, which is back-translated as *a2* in A, and *b* is based on the same lemma/root though representing different headwords, an example of such case is Swe–Eng. <*bo* (verb), *live*>, <*live*, *levande* (adjective)>.

The second subgroup of linguistic reasons covers mismatches that are **semantic** in character and are in principle cases of synonymy or polysemy. However they are often not clear-cut. Translators might give any of several translations so it is not surprising they do not match up.

Consider Swe–Rus <*sakna*, *скучать*>, Rus–Eng <*скучать*, *miss*>, Eng–Swe <*miss*, *sakna*>. Both the original Swedish word and the possible English translations are polysemous. The Swedish one has two frequently used meanings: *to lack* and *to miss (somebody)*; the translator into Russian chose the second one whereas the translator from Russian into Swedish picked the first. Another example is Swe–Rus <*transaktion*, *сделка*>, Rus–Arabic <*сделка*, صفقة>, Arabic–Swe <صفقة, *transaktion*>. It seems that, across several languages, words in this area have both polysemy and near synonyms.

Even in specific domains like computer technology, there are often disambiguation problems for translators due to a number of alternatives, e.g. the Swedish source word *webb* (Eng *web*). In the translation pair Swe–Rus <*webb*, *интернет*> the Swedish *webb* makes an allusion to a spider's web whereas in the Russian term for a spider's web, *паутина* is never used in the internet-related sense[16] and the translator from Russian to Swedish chose not to translate *интернет* as *webb,* but rather as *internet*.

Here Swe–Eng also gives a "never-closing translation circle": Swe–Eng <*webb*, *website*>, Eng–Swe <*website*, *webbplats*>. In the Swe–Eng case the translator opted for the sense narrowing of the source term and neither of the back translations used the sense widening to get back to the source term.

- **Polysemy**, i.e. *b* has more than one meaning and the translation given from B to A is not the meaning of *a*. Examples of this kind are Swe–Rus <*tilltala*, *обращаться*> and Rus–Swe <*обращаться*, *behandla*>, where *обращаться* means both 'address something to someone' and 'treat'; Swe–Rus <*destination*, *назначение*> and Rus–Swe <*назначение*, *förordnande*> where *назначение* means both 'destination' and 'appointment'; Gr–En <*απόδειξη, receipt*>, En–Gr <*receipt, λήψη*> where we have the polysemy of *receipt* between the proof of a purchase, and the event.
- **Synonymy and cognates**: if *a1* translates to *b*, but also has a synonym *a2*, then the back translation might be *a2*, so <*a1, b*> is not a sympair. This often arose because there was both a loan word and a native near-synonym available, as for the Swe–Rus pair <*intervention*, *вмешательство*>. The Swedish source word is a borrowing from English, but the back translation uses the native variant, Rus–Swe <*вмешательство*, *ingripande*> (which also translates to English *intervention*). In another example the translator has chosen a cognate: <Swe–Rus <*lokalt, локально*> (Eng: *locally*). A synonymous Russian word *местно* is a native variant. In general, on very many occasions where there was a native option and a cognate option, a mismatch resulted.
- **Synonyms with different shades of meaning:** for example Swe–Rus <*lyssnare*, *слушатель*> and Rus–Swe <*слушатель*, *åhörare*>, (both, broadly, 'listener')

[16] The only exception is a (relatively) rarely used expression *всемирная паутина* (Eng *worldwide web*).

where *lyssnare* and *åhörare* are synonyms with *åhörare* being a more restricted/infrequent vocabulary item. Synonymy in the B-language can also be a reason for non-sympairs: Swe–Rus <*forskare*, *учёный*> (Eng. *researcher*) versus Rus–Swe <*исследователь*, *forskare*> (Eng. *researcher*).

- **Sense-narrowing:** i.e. an ambiguous source item is provided with a more specific translation (narrowing semantic coverage of the source item), often typical of some domain, e.g. Swe–Eng <*fil*, *lane*>, Eng–Swe <*lane*, *körfil*>. The original Swedish *fil* is polysemous. The translator from Swedish provided two translations, one of them being *lane*. The translator from English into Swedish chose a more specialized term and used compounding, specifying the kind of *lane* by adding *driving* (*kör-*) to the Swedish translation to avoid ambiguity, which resulted in *körfil* (Eng *driving lane*). The Swedish–Norwegian translation went in a different direction: Swe–Nor <*fil*, *rad*> (Eng *row*), Nor–Swe <*rad*, *rad*>. At the same time the Swedish source item *rad* has been symmetrically translated into Norwegian with *rad*.
  Further examples of sense-narrowings are:
  - Swe–Eng <*utspel, gambit*>—*gambit* is particularly in chess whereas *utspel* is more general
  - Swe–Eng <*framkalla, develop*>, Eng–Swe <*develop, utveckla*>. Swedish *framkalla* is polysemous, in one of its meanings relating to photography ('develop a photograph'), while the other is general. The translator to English has selected a narrower term within the domain of photography, which taken independently outside the translation pair can also be interpreted either as a domain-specific item within photography or as a general language word. In the B to A translation the translator chose a more general term.

Finally, **bad translations** (6 % of the analyzed mismatches). Examples include Swe–Eng <*censur, censure* (noun)>, <*censurera, censure* (verb)> where in both cases the English word should have been *censor.* Or the Gr–En example <*παρέα, bunch*>, where *παρέα* should have been translated as "company" or "gang".

Another example needs more explanation: the English noun *surprise* exhibits a systematic polysemy between a 'psychological' and 'external' reading. The Swedish word *förvåning* has only the psychological reading whereas the Russian word *сюрприз* has only the external one so the translation pair <*förvåning, сюрприз*> was not good. The Rus–Swe pair <*сюрприз, överraskning*> was correct, with only the external reading.

4.4 Analysis by language family

One might expect there to be more sympairs where the languages are more closely related. We can test the hypothesis in that Swedish and Norwegian are both North Germanic languages, a branch of the Germanic family, to which English also belongs; Polish and Russian are both Slavic.[17] The percentage of sympairs for these

[17] See eg Ethnologue: http://www.ethnologue.org (Lewis 2009).

Table 9 Sympairs by language family

Scandinavian	36.9 %
Other Germanic (En–Sw, En–No)	39.5, 30.4 %
Slavic (Ru–Pl)	29.2 %
Other (where one of the pair is Arabic or Chinese)	Percentages vary: Ar–Ru 16.5 %, Ar–Sw 7.3 %

is given in Table 9. (Data here is a subset of data in Table 3; we just bring attention to the language families.)

We have used oto-sympair percentages (Table 3) as a metric of lexical similarity to compute a complete-linkage cluster analysis. The resulting tree is given in Fig. 5. Broadly speaking, the clustering corresponds to the genetic relationships between languages. English and Italian are closer than in Fig. 4: perhaps this is because of the mixed lexicon of English, with much that is Romance as well as much that is Germanic. In comparing the two trees we need to bear in mind that the genetic relationships between languages do not take into account later lexical borrowing, in particular the extent to which English words have permeated the vocabularies of various languages.

We can also explore three-language cliques. The sets of three languages for which there are most three-language oto-cliques are:

No–Ru–Sw (535), No–Po–Sw (528), En–No–Sw (503), It–No–Sw (485)
Po–Ru–Sw (473), No–Po–Ru (412), It–Po–Ru (404), En–Po–Ru (397)

The top four triples all include the two closest languages, Norwegian and Swedish. They are joined with first their two geographical and cultural neighbours, Russian and Polish, before their cousin in the language tree, English.

All triples including one of the non-European languages, Arabic and Chinese, scored lower than all-European triples. The lowest score for an all-European triple was 164, for En–Gr–No, whereas the highest for a triple including a non-European language was 99 for Chinese–Polish–Russian. The lowest-scoring triple of all was Arabic–Chinese–Greek with just 22 three-language oto-cliques.

4.5 Are words and their translations of similar frequencies?

It is not clear whether there is any reason to expect words in a sympair to have similar frequencies. Of course our frequencies will come from our corpora, so, if food words are commoner in Italian than Polish, this could be a feature of the corpus —hence uninteresting—or it could be a feature of the language, with Italians talking more about food than Poles—hence interesting—and we will not be well equipped for unpicking the two. But our corpora are broadly comparable in their methods of construction and we can at least begin to explore the question.

First, for all the European languages, for all words in the database, we identified the frequency in the main source corpus, and normalised to frequency per million. We left out Chinese and Arabic because the difficulty in segmentation of the texts into words (for Chinese) and lemmatisation (for Arabic) meant the prospects of

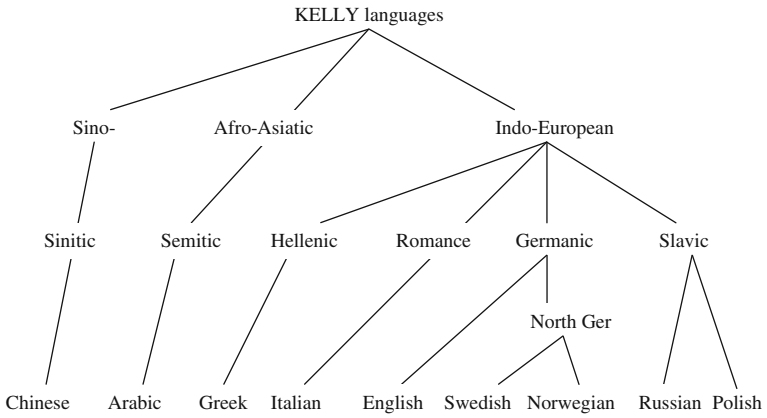

Fig. 4 Genetic relationships between the nine languages in the KELLY project

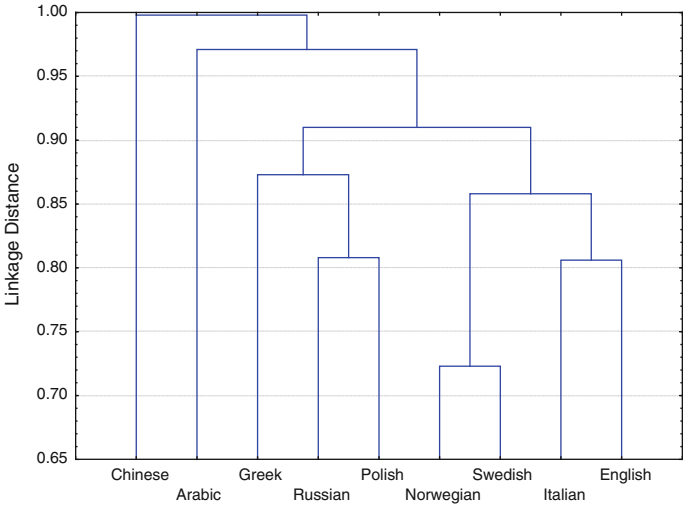

Fig. 5 Cluster analysis of KELLY languages based on sympair distance, one-translation-only

comparing like with like across corpora, without human intervention, was low. Throughout, we normalised to lower-case.

For each oto-sympair[18] for the (undirected) language pairs English–Russian, English–Swedish and Russian–Swedish, we calculated the ratio of the higher normalised frequency to the lower (so the lowest possible value of the ratio, when the normalised frequencies are equal, is 1). In Table 10 we present the numbers of sympairs where this ratio was less than two, between two and four, four and eight, eight and sixteen, and over sixteen.

[18] We excluded the few oto-sympairs containing a multiword from the analysis.

Table 10 Ratios of frequencies for oto-sympairs

Lg pair	No. of oto-sympairs	Ratio <2	2–4	4–8	8–16	>16
Eng–Rus	1,044	634	306	64	14	2
Eng–Swe	1,308	749	401	126	22	10
Swe–Rus	1,292	716	430	119	19	8

The cases of interest are those where ratios are high. For these four language pairs, a member of the group who knew both languages of the pair has looked at all items with a ratio greater than four.

4.5.1 Frequency discrepancy analysis in oto-sympairs

With the help of the KELLY database, we explored why vocabulary that comes in oto-sympairs belongs to different frequency ranges in different languages. Hypothetically, shared vocabulary indicates that it is "basic", i.e. either general in character or coming from core domains. If this is so, why are translation equivalents in different frequency ranges? Could it depend upon the cultural differences, or accidents of corpus composition, or anything else?

During the analysis we once again tried to identify and group reasons. The reasons have proven to be technical, cultural and linguistic, largely as in the non-sympair analysis, with some different indicative numbers, shown in Table 11. Numbers are averages and come from analysis of the three language combinations.

Cultural reasons

This group covers **culturally dependent word choices or usages.** For example, in Eng–Rus <*farm, ферма*> the Russian noun is underused (7 times less frequent), most probably since for a long time a more common term has been *kolkhoz* (collective farm): individual farms are only starting to establish themselves. Another example is Eng–Rus <*queen, королева*> where *queen* is 6 times more frequent in the English corpus, which is not surprising given the political structure of the two countries. It is not difficult to guess how frequency is distributed in Eng–Rus <*soviet, советский*> (1:11). Eng–Rus <*mile, миля*> (6:1) reflects the differences in measurement systems of the two countries. Holiday names also bear witness to cultural differences between countries, e.g. Swe–Rus <*jul, рождество*> (Eng. *Christmas*) where the Swedish item is 4 times more common.

Technical reasons

(a) **Corpus differences:** i.e. *a* is more frequent because the A corpus has many texts in the relevant domain, compared with B. One example, already

Table 11 Analysis of types of frequency discrepancy

Technical reasons	55 % (non-sympair analysis: 26 %)
Linguistic	34 % (non-sympair analysis: 66 %)
Cultural	8 % (non-sympair analysis: 3 %)

mentioned, is medicine in the English corpora where medical terms like *cancer, protein* are used 5 times more often than in the Swedish corpus, and *clinical*, 12 times more often. Another is education (numbers in brackets are the frequency ratio for the pair): <*curriculum, läroplan*> (7:1), <*scientist, vetenskapman*> (6:1), <*university, universitet*> (5:1), <*library, bibliotek*> (5:1), <*classroom, klassrum*> (4:1), <*discipline, disciplin*> (4:1). The Swedish corpus has an inclination towards politics which is shown by the higher frequency of such headwords as <*politician, politiker*> (1:7), <*politics, politik*> (1:6), <*democracy, demokrati*> (1:5), <*islam, islam*> (1:5), <*unemployed, arbetslös*> (1:5).

(b) **Headword selection principles:** To this group belongs the strategy of adding certain learner-relevant vocabulary manually with arbitrary high frequency to secure the item's high rank, e.g. manually added Swedish numerals compared with the Russian list *twenty, thirty, forty*, etc. In the Arabic list, everyday vocabulary items such as food and household objects were underrepresented and so some items were added according to their frequency in other language lists, e.g. دش (shower), لحاف (duvet), زبادي (yogurt) and برتقال (oranges).

Linguistic reasons

The first subgroup is where one word has a broader range of meaning(s) and use(s) than another. For many of these cases there will be several possible translations of the broader word, and it may seem that the proper comparison of frequencies is not with the single narrower term, but with a number of narrower words accounting for the different meanings and uses.

- Eng–Swe <*handsome, snygg*> (1:6): Swedish *snygg* can be used to describe people and objects as well as in an exclamation that means *nice!*
- Eng–Swe <*paper, papper*> (6:1). The other potential translations into Swedish are *tidning (newspaper)/dokument (document)/avhandling (thesis).*
- Rus–Eng <*фамилия, surname*> (8:1) *фамилия* can also be translated as *last name* or *family name*.
- Rus–Swe <*предок, förfäder*> (25:1) where the Russian item is 25 times more frequent. The Russian item has, apart from the provided meaning *ancestor*, a number of other uses, e.g. colloquial *parent.*
- Rus–Swe <*хотя, fastän*>(27:1) (Eng. *though*) *w*here the Russian can potentially also be translated by *även om, dock, även.*
- Rus–Swe <*страсть, passion*> (4:1) *страсть* can be used as an intensifier noun (equal in meaning to *awfully*) as well as a regular noun (*passion*) whereas the Swedish *passion* is used only as a regular noun.

A common special case was cognates, often selected by translators over higher frequency alternatives:

- Rus–Swe <*реальность, realitet*> (6:1) (Eng. *reality*) where *realitet* is a cognate of the Russian, a more frequently used alternative/synonym being the native *verklighet.*

- Eng–Swe <*policy, policy*> (16:1) <*attract, attrahera*> (8:1); <*incident, incident*> (6:1); <*destination, destination*> (5:1)
- Eng–Rus <*innovative, инновационный*> (38:1).

The second subgroup relates to structural differences between languages:

- **Word-building** alternatives may give rise to a number of translation variants for the same source item. In its nature this "reason" is very close to synonymy. For example, the English *reading* can be translated into Swedish as *läsning* or *läsande*; *written* as *skriven* or *skriftlig*; English *surprise* (verb) can be translated into Russian as *удивить* or *удивлять*. In all those cases the translation variants have the same stem plus different affixes, giving a slight semantic difference between the translation variants, often aspectual in character. All of the translation variants can be alternatively used in different contexts and their frequencies should perhaps be summed.
- **Syntactic reasons:** some word classes are more widely used in some languages than others. Nouns are very often more frequent in English than Russian since in English, nouns can be used as noun modifiers whereas in Russian an adjectival phrase is used. For example, the English noun *Sunday* can be used as a pre-modifier *Sunday morning*. Russian allows two variants, one using the noun *Sunday* as a post-modifier (*утро воскресенья*); the other using an equivalent adjective in a pre-modifier position (*воскресное утро*). This may explain the higher frequency of *Sunday* (ratio: 4.5:1).

The interaction of language, meaning and corpus frequency is a topic worthy of much fuller study, for which we hope we provide a launchpad.

5 Summary and outlook

In this paper we have presented the KELLY project. We have described its work on developing word lists, monolingual and bilingual, for language learning, using corpus methods, for nine languages and 72 language pairs. We have presented the method and discussed the many complications encountered. We have loaded the data into an online database and made it accessible for anyone to explore: we presented our own first explorations of it.

The vocabulary has been selected firstly using objective, statistical criteria, namely the monolingual frequency lists initially generated for each language; secondly, by translating all lists into all eight other languages and investigating the network of translations in all directions to identify omissions and anomalies; thirdly, using any other corpora and wordlists that were available; and fourthly, the scrutiny of linguists. In this way we have developed resources for second language learners and for linguistic research.

We have produced the online KELLY database of searchable corpora for nine languages. We have identified key concepts for describing and exploring the database, including *sympair, oto-sympair, n-language clique* and different categories

of *non-sympair*. With this armoury, we have found sets of words and concepts which tend to get straightforward and direct translations, and presented them in the text and appendices. Nouns dominated the lists. Institutions and the natural world were well-represented. The number of sympairs for a language pair reflects both the language family tree, and the cultural and geographical proximity between the countries where the languages are spoken.

We investigated a sample of cases where translations were not symmetrical, and found a number of recurring patterns involving differences in the corpora, list construction methods, culture, and linguistics. The linguistic differences included differences of syntax, morphology and word-formation between languages, as well as synonymy (particularly that involving cognates), homonymy and polysemy. We also examined cases where translations were symmetrical, but frequencies for a word and its translation were very different. The reasons, again, were corpus design, culture, synonymy, polysemy, homonymy.

We invite researchers to evaluate the word lists against others, and their validity in the classroom. We believe the KELLY lists could become key resources, perhaps official vocabularies, for language teaching for those KELLY languages where currently available resources are poor. We shall be making the case for adoption of KELLY lists (or, in all likelihood, their successors) to the language-teaching institutions of several KELLY countries. And we invite others to explore the database to unpick further the tangled threads of meaning, translation and frequency that we have encountered.

Appendix 1

See Table 12.

Table 12 Other resources used (corpora and wordlists)

Language	Other corpora and word lists used
English	BNC, BNC-spoken
Greek	Official list from the Center for the Greek Language
	Hellenic National Corpus (HNC): 50 million words written, various text types and genres
	HNC frequency lists
Italian	Italian PAROLE corpus: 250,000 words, newspapers and periodical
	Corpus Stammerjohann: 100,000 words spontaneous speech
	Corpus per il Confronto Diacronico LABLITA: 1,000,000 words of speech, Florence area
Norwegian	NoTa Corpus of Oslo Speech and Nordic Dialect Corpus
Polish	National Corpus of Polish and the Top 200 frequency list of the PWN corpus
Swedish	EU project Parole corpus with 24 million tokens from newspaper texts, novels, periodical and web-texts from 1976 to 1990s

Appendix 2

See Table 13.

Table 13 Guidelines for inclusion of word types in KELLY lists

Word type	Policy	Comments
Variants	Spelling variants should be amalgamated, so that e.g. organize and organise are counted as one word for frequency calculations. Each language team will have to have a style guide for preferred forms for the list itself. For English, British and US spelling variants such as color/colour will also be amalgamated	
	Lexical variants*, e.g. cash machine/ATM should be treated as separate items	
Inflected forms	These are not shown unless an inflected form has a meaning that is not inherent in the base form, e.g. better in the sense of 'to get better'	Although learners may want to look up inflections, esp. irregular ones, for the purposes of frequency they should be treated together with the base form
Derivational inflected forms e.g. quickly, happiness	To be treated as words in their own right, i.e. as separate lemgrams	
Affixes, including productive affixes	No, an affix will only appear if it forms a word that is common enough in itself to merit inclusion	
Abbreviations	Yes, including abbreviations that are written only, but *only* if they meet the normal criteria of what we are including, so *not* abbreviations for proper nouns and encyclopaedic items. The most common abbreviations will probably be forms of address, weights and measures, Latin abbrevs, and the few cases where an abbreviation is the normal way to refer to an item, e.g. DVD	NB The inclusion of abbreviations will mean searching on the non-alphabet character [.]
Multiword units	Yes for the teams who decided to add them at this stage, no for those who didn't	
Hyphenated compounds	Yes, as long as they can be found automatically	

Table 13 continued

Word type	Policy	Comments
Phrasal verbs	No for English, as they count as multi-words—yes for languages where they have a one word lemma	
Phrases, idioms, proverbs, quotations	No	
Subject-specific vocabulary	Only if it makes it by the normal frequency criteria (it may do, for instance for some computing terms)	NB When it comes to adding CEF levels, we may need to consider grammar vocabulary as a special case because of its usefulness to language learners
Dialect words	No	
Items marked by register, e.g. very formal, slang, offensive	Normal frequency rules apply: if they come in the top 5,000 then yes	NB We agreed that an 'offensive' attribute should be added to the database so that while the frequency lists themselves can be purely frequency based, offensive items can be weeded out if necessary
Geographic terms	*Country name/related adjective/ name of people/language* For these: give your own, then any others that appear in your frequency list in the normal way	
	Oceans/continents/important areas/mountain ranges These should be included on a frequency basis, but privilege items which are *not* from your own area. So for the English list, an item such as 'Mediterranean' would be more important than 'Lake District'. This suggestion is to avoid over-representation of these items—every list is likely to include many from one's own region	
	Cities Your own capital city, plus any really major cities in your country which have a different name in translation. Then any cities from other countries which fulfil the normal frequency criteria *and* have a different name in your language from the original	
	We will *not* cover individual rivers, mountains, deserts etc	

Table 13 continued

Word type	Policy	Comments
Famous places and buildings	Only if they have metonymy, e.g. Hollywood. Likely to be very rare	
Stars, planets, galaxies, etc	No	
Imaginary, biblical or mythological people or place names	No	
Personal names	No	
Famous people and places, and other encyclopaedic information such as names of wars, treaties, names of ancient peoples, names of organizations, etc	No	
Adjectives derived from famous people	Only if they are in the top 5,000	
Festivals and ceremonies	If they are in the top 5,000	
Trademarks	If they appear in the top 5,000 and are the name of an item, but not company names	
Beliefs and religions, and associated nouns and adjectives	If they are in the top 5,000	
Currencies	Include your own currency and any others in the top 5,000	

* Otherwise referred to as synonyms

Appendix 3: English words that featured in 7-language cliques

This list is included as a more readable, duplicate free, but English-only list of items appearing to have a high degree of language-neutrality

afternoon age aggressive air almost already angel apple balcony beer believe big bird blind blood body bus category catholic central chaos cheese christian city clinical club comment constitution contact corruption country court cry culture daughter democracy description diagnosis dialogue dictionary difficulty digital direction director discipline distance document dollar door doubt eighty engineer example experiment family february festival fifteen fifth fifty filter finger five flag flower four french fresh friend garden glass god green guarantee have height hero history hope hundred ice industrial

industry italian july june key kilometre knife lake liberal life light literature litre loan long mathematical mathematics meat mechanism member metal method million minister minute month mother museum myth national nervous new nightmare nine ninth nose november page pain park parliament pay period personality philosophical philosophy planet poem poet police population president price product production professor quality question radio rain read religion restaurant revenge river role root salt saturday scandal school screen sea series seventy shirt simple six sixty sky sleep soldier son stability strategy sugar sunday surprise sweet sword symbol tail talent technology temperature temple text theatre theoretical third three thursday ticket time tobacco tooth tournament tower tradition tragic travel twelve twenty two typical understanding video virus vote war weather white window winter woman word wound write year.

Appendix 4

See Table 14.

Appendix 5

See Table 15.

Table 14 33 7-language-oto-Cliques

	苹果	apple	μήλο	Mela		jabłko	яблоко	äpple
		apple	μήλο	Mela	eple	jabłko	яблоко	äpple
جبن		cheese	τυρί		ost	ser	сыр	ost
		cheese	τυρί	Formaggio	ost	ser	сыр	ost
فساد		corruption	διαφθορά	corruzione	korrupsjon		коррупция	korruption
		corruption	διαφθορά	corruzione	korrupsjon	korupcja	коррупция	korruption
		february	φεβρουάριος	Febbraio	februar	luty	февраль	februari
		fifteen	δεκαπέντε	Quindici	femten	piętnaście	пятнадцать	femton
		fifty	πενήντα	Cinquanta	femti	pięćdziesiąt	пятьдесят	femtio
	马	horse	άλογο	Cavallo	hest	koń		häst
		july	ιούλιος	Luglio	juli	lipiec	июль	juli
		june	ιούνιος	Giugno	juni	czerwiec	июнь	juni
ركبة		knee	γόνατο	Ginocchio	kne	kolano	колено	
بحيرة	湖	lake	λίμνη	Lago		jezioro	озеро	
		litre	λίτρο	Litro	liter	litr	литр	liter
مليون		million	εκατομμύριο	Milione	million	milion	миллион	
متحف		museum	μουσείο	Museo	museum	muzeum		museum
		museum	μουσείο	Museo	museum	muzeum	музей	museum
		nightmare	εφιάλτης	Incubo	mareritt	koszmar	кошмар	mardröm
أنف		nose	μύτη	Naso	nese	nos	нос	
جيب		pocket	τσέπη	Tasca	lomme	kieszeń		ficka
		pocket	τσέπη	Tasca	lomme	kieszeń	карман	ficka
	沙	sand		Sabbia	sand	piasek	песок	sand
		saturday	σάββατο	Sabato	lørdag	sobota	суббота	lördag
		seventy	εβδομήντα	Settanta	sytti	siedemdziesiąt	семьдесят	sjuttio
شاي	茶	tea	τσάι	Tè		herbata		te

Table 14 continued

شاي		tea	τσάι	Tè	te	herbata		te
		tooth	δόντι	Dente	tann	ząb	зуб	tand
		twelve	δώδεκα	Dodici	tolv	dwanaście	двенадцать	tolv
	二十	twenty	είκοσι	Venti		dwadzieścia	двадцать	tjugo
فيروس	病毒	virus	ιός		virus	wirus		virus
فيروس	病毒	virus		Virus	virus	wirus		virus
ذئب	狼		λύκος	Lupo	ulv	wilk	волк	
	狼		λύκος	Lupo	ulv	wilk	волк	varg

Note that some cliques are largely overlapping. According to our definitions, they are not duplicates since at least one item is different in each case

Table 15 49 8-language-cliques

	银行	bank	τράπεζα	banca	bank	bank	банк	bank
	床	bed	κρεβάτι	letto	seng	łóżko	кровать	sang
قنبلة	炸弹	bomb	βόμβα	bomba		bomba	бомба	bomb
قنبلة		bomb	βόμβα	bomba	bombe	bomba	бомба	bomb
	书	book	βιβλίο	libro	bok	książka	книга	bok
خبز	面包	bread	ψωμί		brød	chleb	хлеб	bröd
	面包	bread	ψωμί	pane	brød	chleb	хлеб	bröd
جسر	桥	bridge	γέφυρα	ponte	bro		мост	bro
	椅子	chair	καρέκλα	sedia	stol	krzesło	стул	stol
قناة		channel	κανάλι	canale	kanal	kanał	канал	kanal
كنيسة	教堂	church	εκκλησία	chiesa	kirke	kościół		kyrka
مناخ		climate	κλίμα	clima	klima	klimat	климат	klimat
قهوة	咖啡	coffee	καφές		kaffe	kawa	кофе	kaffe
قهوة	咖啡	coffee		caffè	kaffe	kawa	кофе	kaffe
كلب	狗	dog		cane	hund	pies	собака	hund
عين		eye	μάτι	occhio	øye	oko	глаз	öga
أب	父亲	father	πατέρας	padre	far	ojciec	отец	
	鱼	fish	ψάρι	pesce	fisk	ryba	рыба	fisk
غابة	森林	forest	δάσος		skog	las	лес	skog
مستقبل	未来	future	μέλλον	futuro		przyszłość	будущее	framtid
حكومة	政府	government	κυβέρνηση	governo		rząd	правительство	regering
	心脏	heart	καρδιά	cuore	hjerte	serce	сердце	hjärta
مطبخ	厨房	kitchen	κουζίνα	cucina	kjøkken	kuchnia	кухня	
مطبخ	厨房	kitchen	κουζίνα	cucina	kjøkken		кухня	kök
مستوى		level	επίπεδο	livello	nivå	poziom	уровень	nivå
منطق	逻辑	logic	λογική	logica	logikk		логика	logic

Table 15 continued

	逻辑	logic	λογική	logica	logikk	logika	логика	logic
زواج	婚姻	marriage		matrimonio	ekteskap	małżeństwo	брак	äktenskap
	牛奶	milk	γάλα	latte	melk	mleko	молоко	mjölk
مكتب		office	γραφείο	ufficio	kontor	biuro	офис	kontor
سجن	监狱	prison	φυλακή	prigione	fengsel	więzienie		fängelse
سجن	监狱	prison	φυλακή		fengsel	więzienie	тюрьма	fängelse
مشكلة		problem	πρόβλημα	problema	problem	problem	проблема	problem
	心理学	psychology	ψυχολογία	psicologia	psykologi	psychologia	психология	psykologi
ثورة	革命	revolution	επανάσταση		revolusjon	rewolucja	революция	revolution
	革命	revolution	επανάσταση	rivoluzione	revolusjon	rewolucja	революция	revolution
	雪	snow	χιόνι	neve	snø	śnieg	снег	snö
مصدر		source	πηγή	fonte	kilde	źródło	источник	källa
	系统	system	σύστημα	sistema	system	system	система	system
	十	ten	δέκα	dieci	ti	dziesięć	десять	tio
تجارة		trade	εμπόριο	commercio	handel	handel	торговля	handel
جامعة	大学	university	πανεπιστήμιο		universitet	uniwersytet	университет	universitet
ماء	水	water	νερό		vann	woda	вода	vatten
اسبوع		week	εβδομάδα	settimana	uke	tydzień	неделя	vecka
	周	week	εβδομάδα	settimana	uke	tydzień	неделя	vecka
مدينة	城市		πόλη	città	by	miasto	город	stad
عشرة	十		δέκα	dieci	ti	dziesięć	десять	tio
مطر	雨		βροχή	pioggia	regn	deszcz	дождь	regn
هاتف	电话		τηλέφωνο	telefono	telefon	telefon	телефон	telefon

References

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. Oxford: Wiley.

Allen, S. (1972). *Tiotusen i topp* [Top ten thousand]. Sweden: Almqvist & Wiksell.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation, 43*(3), 209–226. http://clic.cimec.unitn.it/marco/publications/wacky-lrej.pdf.

Borin, L., Forsberg, M., Friberg Heppin, K., Johansson, R., & Kjellandsson, A. (2012). Search result diversification methods to assist lexicographers. In *Proceedings of the 6th Linguistic Annotation Workshop*.

Bortolini, U., Tagliavini, G., & Zampolli, A. (1972). *Lessico di frequenza della lingua italiana contemporanea*. Milano: Garzanti.

Buckwalter, T., & Parkinson, D. (2011). *A frequency dictionary of Arabic: Core vocabulary for learners*. London: Routledge.

Capel, A. (2010). A1-B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal, 1*(1), 1–11.

Charalabopoulou, F., & Gavrilidou, M. (2011). Creating frequency-based vocabulary lists for L2 learners. In *Proceedings of the 10th international conference on Greek Linguistics, Komotini, Greece* (in print).

Council of Europe. (2001). The common European framework of reference for languages. Available at http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf. Last Accessed March 2, 2010.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37*(1), 65–70.

De Mauro, T. (1997). *Guida all'uso delle parole*. Roma: Editori Riuniti.

De Mauro, T., Mancini, M., Vedovelli, M., & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato*. Milano: EstasLibri.

Efstathiadis, S., Antonopoulou, N., Manavi, D., & Vogiatzidou, S. (2001). *Certificate of attainment in Greek*. Salonica: Ministry of Education Center for the Greek Language.

Forsbom, E. (2006). Deriving a base vocabulary pool from the Stockholm Umeå Corpus.

Gellerstam, M. (1978). Välja sina ord. Reports from Språkdata 9.

Heimann Mühlenbock, K. (2012). *I see what you mean—Assessing readability for specific target groups* (PhD Thesis). Gothenburg: Department of Swedish, University of Gothenburg.

Huld, M. (1986). Proto- and post-Indo-European designations for 'sun'. *Zeitschrift für vergleichende Sprachforschung, 99*(2), 194–202.

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: Cambridge University Press.

Josselson, H. (1953). *The Russian word count and frequency analysis of grammatical categories of standard literary Russian*. Detroit: Wayne University Press.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*, 91–113. http://www.kilgarriff.co.uk/Publications/1997-K-CHum-believe.pdf.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of International Conference on Corpus Linguistics*. Liverpool. http://www.kilgarriff.co.uk/Publications/2009-K-CLLiverpool-SimpleMaths.doc.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX international congress* (pp. 425–432). Barcelona: Universitat Pompeu Fabra.

Kokkinakis, D., & Johansson Kokkinakis, S. (1997). *A Robust and modularized lemmatizer/tagger for Swedish based on large lexical resources*, Inst. f. svenska språket, Göteborgs Universitet.

Kosem, I., Husák, M., & McCarthy, D. (2011). GDEX for Slovene. In I. Kosem, & K. Kosem (Eds.), *Proceedings of eLex 2011 on electronic lexicography in the 21st century: New applications for new users*, Bled, 10–12 November 2011 (pp. 151–159). Ljubljana: Trojina, Institute for Applied Slovene Studies.

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.

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (Sixteenth edn.). Dallas, TX: SIL International. Online version: http://www.ethnologue.com/.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

McCarten, J. (2007). *Teaching vocabulary—Lessons from the corpus—Lessons for the classroom*. Cambridge: Cambridge University Press.

Mikros, G. (2007). Corpora in Modern Greek language research and teaching: An overview of the research project. In *Workshop of Pythagoras: Strengthening Research Groups in the National and Kapodistrian University of Athens* (p. 48).

Mondria, J.-A., & Mondria-de Vries, S. (1994). Efficiently memorizing words with the help of word cards and 'hand computer': Theory and applications. *System, 22*(1), 47–57.

Nakata, T. (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, P. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

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

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly, 11*, 151–167.

Rundell, M. (Ed.). (2007). *Macmillan English dictionary for advanced learners*. London: Macmillan.

Sawalha, M., & Atwell, E. (2010). Fine-grain morphological analyzer and part-of-speech tagger for Arabic text. In *Proceedings of the LREC 2010*, 17–23 May 2010. Valleta, Malta.

Schmitt, N., & Schmitt, D. (1995). Vocabulary notebooks: Theoretical underpinnings and practical suggestions. *ELT Journal, 49*(2), 133–143.

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics, 11*(4), 435–462.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., & Divjak, D. (2008). Designing and evaluating a Russian tagset. In *Proceedings of the sixth language resources and evaluation conference, LREC 2008.* Marrakech.

Sharoff, S., Umanskaya, E., & Wilson, J. (2013). *A frequency dictionary of Russian: Core vocabulary for learners*. London: Routledge.

Shteinfeld, E. A. (1963). *Chastotnyj slovarj sovremennogo russkogo literaturnogo jazyka* [Frequency dictionary of modern Russian literary language]. Tallin.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's handbook of 30,000 words*. New York: Bureau of Publications, Teacher's College, Columbia University.

Turnbull, J., et al. (2010). *Oxford advanced learners' dictionary*. Oxford: Oxford University Press.

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.

Volodina, E. (2010). *Corpora in language classroom: Reusing Stockholm Umeå Corpus in a vocabulary exercise generator*. Saarbrücken: LAP Lambert Academic Publishing.

Volodina, E., & Johansson Kokkinakis, S. (2012). *Swedish KELLY: Technical report*. GU-ISS-2012-01. The Swedish Language Bank, University of Gothenburg.

Waring, R. (2004). *In defence of learning words in word pairs: But only when doing it the 'right' way!*. Available at http://www1.harenet.ne.jp/~waring/vocab/principles/systematic_learning.htm. Last Accessed September 25, 2011.

West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. London: Routledge.

Zasorina, L. N. (Ed.). (1977). *Chastotnyj slovarj russkogo jazyka* [Frequency Dictionary of Russian]. Moscow: Russkij Jazyk.