SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners

Thomas François[1,3], **Elena Volodina**[2], **Ildikó Pilán**[2], **Anaïs Tack**[3]
[1] Chargé de recherche FNRS
[2] Språkbanken, University of Gothenburg
[3] Cental, IL&C University of Louvain
thomas.francois1@uclouvain.be, elena.volodina@svenska.gu.se, ildiko.pilan@svenska.gu.se

Abstract

The paper introduces SVALex, a lexical resource primarily aimed at learners and teachers of Swedish as a foreign and second language that describes the distribution of 15,681 words and expressions across the Common European Framework of Reference (CEFR). The resource is based on a corpus of coursebook texts, and thus describes receptive vocabulary learners are exposed to during reading activities, as opposed to productive vocabulary they use when speaking or writing. The paper describes the methodology applied to create the list and to estimate the frequency distribution. It also discusses some characteristics of the resulting resource and compares it to other lexical resources for Swedish. An interesting feature of this resource is the possibility to separate the wheat from the chaff, identifying the core vocabulary at each level, i.e. vocabulary shared by several coursebook writers at each level, from peripheral vocabulary which is used by the minority of the coursebook writers.

Keywords: CEFR-graded lexicon, Swedish as a second/foreign language, ICALL

1. Introduction

When developing a second or foreign language (L2) course, setting vocabulary goals for learners remains a challenging task. Second language acquisition (SLA) research has shown that a reader has to know 95-98% of the running words in a text to understand it correctly (Laufer and Ravenhorst-Kalovski, 2010). Such studies are useful to estimate the size of the vocabulary needed to read a text, but they do not provide a method to define lexical curriculum for L2 learners.

One possible way to address this problem consists in creating vocabulary lists in which each word is located on a proficiency scale. Given the level of a target reader, it is therefore possible to have an estimate of the words he/she is supposed to know. With this in mind, we present a lexical resource, SVALex, aimed not only at learners and teachers of L2 Swedish, but also at lexicographers, L2 test and curriculum developers, as well as researchers within Intelligent Computer-Assisted Language Learning (ICALL). This resource distinguishes itself from the existing similar lists (see section 2.) due to its descriptive nature covering the distribution of 15,681 words and expressions across a widely used L2 proficiency scale, the Common European Framework of Reference (CEFR) (Council of Europe, 2001). To make sure that the selected lexical items in SVALex are relevant for learners at different levels of language proficiency, we calculate lexical distributions based on COCTAILL, a corpus of L2 coursebooks graded by teachers in compliance with the CEFR scale (Volodina et al., 2014).

In section 2., we describe previous work aimed at establishing CEFR-graded lists for various languages or adopting an automatic approach to classification of words according to the CEFR. Section 3. details the methodology applied to create SVALex from a CEFR-labeled corpus of L2 Swedish texts. It explains how the texts were processed in order to get POS-disambiguated words and multi-word expressions (MWE). The frequency distribution of linguistic terms was then estimated and normalized before the list was finally manually cleaned. The obtained resource is analyzed in section 4., where we also compare SVALex with other existing resources for Swedish.

2. Previous work

The CEFR contains guidelines for the harmonization of language teaching and assessment across languages and countries, which has become the reference for L2 learning in Europe and beyond. It also defines 6 proficiency levels (from beginner to mastery): A1, A2, B1, B2, C1, C2. Unfortunately, the language of the document describing the requirements of each level and skill is very general and often does not provide a clear definition of how to interpret and assess the target skill (see Figure 1). In a number of countries, there have been made efforts to interpret the CEFR guidelines in the form of reference level descriptors[1]. These books describe the language competences expected from an L2 learner in each of the CEFR levels, including a list of words, syntactic structures, and expressions associated with specific communicative functions or themes.

Unfortunately, such descriptors are not available for Swedish. Moreover, the reference level descriptors, although they are valuable tools, have raised some concerns among researchers. Alderson (2007) agrees with Little to say that "the methodologies being used [to compile these descriptions] are unclear or suspect". According to Beacco et al. (2011), the authors of the French reference level descriptor, their approach rests on several sources: the expertise of educationalists and linguists, the expertise of decision-makers in L2 teaching programming, and achievements of research in second language acquisition (SLA). However, Hulstijn (2007) argues that these

[1] See the list of concerned languages at `http://www.coe.int/t/dg4/linguistic/dnr_EN.asp?`

references should be based instead on empirical data collected among learners as well as on statistical data obtained through a corpus analysis.

Besides the CEFR descriptors, several attempts have been made to create frequency-based vocabulary lists for Swedish L2, among which are the Kelly list (Kilgarriff et al., 2014), the Base Vocabulary Pool (Forsbom, 2006), SveVoc (Heimann Mühlenbock and Johansson Kokkinakis, 2012) and Swedish Academic Wordlist (Jansson et al., 2012). Even Lexin, a mono- and bilingual lexicon series, has been developed with Swedish language learners in mind (Hult et al., 2010). However, of those resources, only Kelly has attempted to link vocabulary to the CEFR proficiency scale thus indicating which vocabulary should be introduced at which level.

In its original state, the Swedish Kelly list (hereinafter Kelly) provides CEFR labels for 8,425 headwords (Volodina and Kokkinakis, 2012). Kelly is a frequency-based wordlist generated from web corpora, and translated into and compared between nine languages for identification of core vocabulary across these languages (Kilgarriff et al., 2014). However, Kelly has shortcomings, namely that (1) frequency statistics are collected from web texts aimed at L1 speakers of Swedish, which can be misleading since the vocabulary used for L1 speakers may differ from what beginner L2 speakers need to concentrate on; (2) the division into the CEFR levels is based on frequency and L1 text coverage, which needs explicit validation to confirm its relevance for a CEFR-based curriculum; and (3) Kelly lacks some vocabulary useful in the L2 context, such as *table, alphabet, toothpaste* - i.e. vocabulary not appearing in L1 web texts.

Another frequency-based lexical resource for Swedish that is potentially appropriate for L2 learners is the Base Vocabulary Pool (BaseVoc) (Forsbom, 2006), that relies on the assumption that domain-specific or genre-specific words should not be part of the base vocabulary pool. The core of such a pool should rather consist of stylistically neutral and general-purpose words collected from as many domains and genres as possible (in this case in at least three genres/domains). As a result, out of 69,371 entries in the vocabulary based on Stockholm Umeå Corpus (SUC) (Ejerhed et al., 1992), only 8,215 lemmas were retained as forming the base vocabulary pool. Yet, in spite of a proportionally small number of lemmas constituting the base vocabulary pool, they account for 88.2% of the SUC texts (Forsbom, 2006). In the context of second language learning it means that a learner who has acquired the knowledge of these words can read and understand most of the modern Swedish texts. However, as was the case for the coverage approach of Laufer and Ravenhorst-Kalovski (2010), this list only describes a global learning goal and does not contain any indications for appropriateness at different levels of learner proficiency.

A third list aimed at L2 learners of Swedish is the Swedish Academic word list (Carlund et al., 2012). It is a domain-specific list aimed at advanced learners at the university level who need to acquire specific vocabulary used for writing academic papers. The list has been generated from a corpus consisting of academic papers at master and doctoral level, as well as academic articles. The target L2 group for the academic word list should eventually be learners at C1 or C2 level, in other words only a small subset of the learners we are addressing in SVALex.

> Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.
> Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.

Figure 1: Description of Vocabulary range at A2 (Council of Europe, 2001, 112). Subject to interpretations is *sufficient vocabulary, familiar situations and topics, basic communicative needs, simple survival needs.*

Finally, Lexin is a series of lexicons aimed at immigrants, which comprises a mono-lingual version of the lexicon as well as a number of aligned bilingual editions (Hult et al., 2010). It is a comparatively large resource of approximately 30,000 entries, each of them including definitions, examples, and information on grammatical patterns. It is a useful electronic resource, which however has no indication of learner level or frequency ranking.

As regards other languages, a number of attempts to create CEFR-based lists have been carried out. The English Vocabulary Profile project (Capel, 2010; Capel, 2012), or EVP, is one of the inspiring examples where efforts are made to interpret the CEFR document using a corpus-informed approach based on learner production (*Cambridge Learner Corpus*). This list was obtained through a threefold process. First, the initial set of words considered were those among the top 6000 words or senses in English that had been manually assigned a frequency tag for the *Cambridge Advanced Learner's Dictionary*. These words were then further scrutinized using the *Cambridge Learner Corpus* in order to detect the most used words worldwide and to assign a CEFR level to the selected words or senses. This level was finally checked against the *Cambridge English Lexicon*.

The EVP project aims at describing which words are **actually known** by learners at different levels rather than providing a list of terms that learners should be **exposed to** (Capel, 2010; Capel, 2012). The methodology applied has the great advantage to be able to assign different difficulty levels to the different senses of a word. However, it is also a long and costly process to repeat for other languages. Moreover, Alderson (2007) stressed that relying almost entirely on the *Cambridge Learner Corpus*, a collection of performances on Cambridge examinations, may be an issue for generalisation.

A more time-effective approach has been proposed by François et al. (2014), in which the authors automatically extracted a list of lexical items from a corpus of texts linked to the CEFR scale. As a result, French L2 learners and teachers have FLELex, a freely available list of 17,871 French words and multi-word expressions (MWE) associated with their frequency distribution across the six CEFR levels. We reused part of FLELex methodology to generate

SVALex, another list in the "CEFRLex" family.

3. Methodology

One of the reasons why the FLELex methodology has not been applied more widely previously is the need of a corpus aimed at L2 learners that is (1) large enough to allow a robust estimation of the frequency distribution of words, and (2) in which every text has been located on the CEFR scale. To our knowledge, such corpora are only reported for Swedish by Volodina et al. (2014) and for French by François (2014). Similar resources exist for a few other languages, but they deviate in one way or another from the requirements above. For example, a corpus of reading exam materials for Portuguese (Branco et al., 2014) are linked to the CEFR levels, but the size of the resource is rather modest. A corpus with English coursebooks is relatively large, but the texts are linked to a different scale of proficiency (Heilman et al., 2007).

3.1. Source corpus

The corpus COCTAILL (Volodina et al., 2014) contains digitised coursebooks used for teaching CEFR-based L2 Swedish, where each level is represented on average by four coursebooks, except for C2, for which no coursebook was available, most likely because it is a near-native proficiency level. Coursebooks have been included into the corpus based on teachers' judgements, which has been the most important criteria in corpus compilation. This means that only coursebooks that have been confirmed as appropriate for teaching CEFR-based courses at an announced level are included into COCTAILL. Each coursebook is manually structured into lessons, and within lessons into texts, exercises, lists and language examples. Apart from that, rich pedagogical and textual annotation has been added to the corpus, which now allows to filter material for texts by their topics and/or genres, and the rest of the material according to target skills and competences (e.g. vocabulary, speaking, reading, listening etc.), formats (brainstorming, gaps, matching, etc.) and units (single words, phrases, dictionary entries etc.). Linguistic annotation in the form of POS-tags, syntactic relations and lemmatization has been automatically added to the corpus using the Korp pipeline (Borin et al., 2012). To ensure comparability to FLELex (François et al., 2014), only a subset of COCTAILL containing texts aimed at reading comprehension has been used for the generation of SVALex.

3.2. Estimation of lexical frequencies

The entries in SVALex consist of lemmas[2], their parts of speech (POS) and frequencies across 5 of the 6 CEFR levels. Apart from single words, our list is populated with multi-word expressions (MWE), which were extracted by the Korp-pipeline using a pilot feature based on a knowledge-based approach. If the analyzed token is a potential constituent in an MWE that is listed in the SALDO lexicon (Borin et al., 2013), the sentence is checked for the presence of the remaining constituents and their order, and a number of POS-specific rules are applied to test linguistic behaviour constraints.

When estimating the word distribution across the 5 CEFR-levels considered, we did not rely on raw frequency by level (*RFL*) since, as noted by Francis and Kucera (1982), lower frequency words tend to be context-specific, appearing in a small number of texts, but sometimes with an unusually high frequency within those texts. To reduce the impact of this issue, we have applied a dispersion index (D) to the *RFL* using the formula described in Carroll et al. (1971):

$$D_{w,K} = [\log(\sum p_i) - \frac{\sum p_i \log(p_i)}{\sum p_i}] / \log(I) \quad (1)$$

For a corpus with K levels of difficulty (in our case, $K = 5$), the D of a given word w for the level K requires to use p_i, the probability that a word appears in the textbook i, and I, which is the number of textbooks at the level k. When $p_i = 0$, $p_i \log(p_i)$ is also considered 0. After Ds are computed, we can combine the *RFL* with the D values to obtain the normalized frequency per million for a given word w, referred to as U. The formula is as follows (Carroll et al., 1971):

$$U = \frac{1,000,000}{N_k}(RFL * D + (1 - D) * f_{min}) \quad (2)$$

in which N_k is the total number of tokens for level k and f_{min} represents $1/N$ times the sum of the products f_i and s_i, where f_i is the frequency of a word in textbook i and s_i corresponds to the number of tokens in the textbook.

3.3. Manual editing

Whereas FLElex was cleaned once for all after its generation, a number of manual adjustments to SVALex were performed in a circular fashion, i.e. alternating steps of manual editing and regeneration of the list for further checking. Such process allowed us to recover the frequency counts of problematic forms and assign them to the correct forms (which was not done for FLELex). As regards manual editing, we first checked every word form for which a lemma could not have been identified during the automatic linguistic annotation of COCTAILL, amounting to about 3,500 items in total. Some reasons why the automatic processing was problematic in these cases include: compounding, proper names, use of other languages (e.g. English), inconsistent spelling and incorrect optical-character recognition (OCR) of the texts from the corpus. Each of these items have been looked up in the lexical-semantic resource SALDO (Borin et al., 2013) and, if a corresponding entry was found, then the lemma and its POS were manually corrected. Besides that, all participles have been manually converted to either verbs or adjectives to adjust to an updated version of the annotation pipeline currently under development (Adesam et al., 2015).

As an additional check, the candidate list was matched with 3 other resources: Base Vocabulary Pool, Kelly and Lexin to identify SVALex items not present in any of the resources. This way, a number of problematic cases, such as MWE written without a space, have been identified and

[2]The rationale of using the lemma instead of inflected forms is that words having numerous inflected forms (such as verbs) would then have their probability mass split between their inflected forms in comparison with invariable words.

Resource	# items	# overlap	# missing
SVALex	15,681	N/A	N/A
Swedish Kelly	8,425	5,757	9,924
Base Vocabulary	8,220	4,964	10,717
Lexin	30,684	9,039	6,642

Table 1: Size of the resources in number of entries. For each resource other than SVALex, the number of overlapping items with SVALEX and missing SVALex items is given.

corrected. While Kelly and BaseVoc are shorter lists than SVALex and thus cannot be expected to contain all the items, Lexin is a more extensive resource and provides a good point of reference (Table 1).

Surprisingly, a total of 6,189 of correct SVALex items did not have any match in any of the three resources (called here *no-hit-items*). Inspection of those correct entries has shown that:

- ∼80% of no-hit-items are compounds consisting of several stems. Swedish is known for its rich compounding, and many of the used word stems are represented in the other resources as independent entries, though not in their combination: e.g. *klokttro* [cleverly believe], *femhundra* [five hundred], *byxben* [trousers legs], *astråkig* [super boring]
- ∼15% are multi-word expressions, e.g. *bre ut sig* [to spread around], *från och med nu* [starting from now], *betala kalaset* [take consequences/pay the price]
- ∼1% are abbreviations: e.g. *eKr* [AC, After Christianity], *odyl* [and the like], *kvm* [m2, square meter]
- ∼1% are modern or colloquial words: e.g. *app, gin, luska* [to nose about], *fniss* [(a) giggle]
- ∼1% are alternative (e.g. colloquial, modern or old-fashioned) spelling variants: *förrn* [before] instead of *förrän*, *likasom* [as well as] instead of *liksom*. Appearance of old-fashioned expressions in SVALex can be explained by use of poems, lyrics and historical documents in the coursebooks, whereas colloquial words and expressions appear due to extensive use of dialogues.

4. Description of the resource

SVALex contains 15,681 items that Swedish L2 learners are *exposed to* during their courses. Of these, 10% are MWE. The distribution of vocabulary that is expected to be recognized by learners at each of the five CEFR levels is shown in Table 2.

The vocabulary is partially overlapping between the levels (see column *# items*), which means that, for example, C1 learners do not learn 7,564 new items during the C1 course, but the total vocabulary used in the C1 coursebooks contains 7,564 different unique lemma-POS combinations, part of which have been used at previous levels. Column 3 shows the number of new items that have not been used in the texts at lower levels. As expected, the number of new items at higher levels (B1-C1) is greater.

The strength of our approach is that it allows us to find an objective **core** vocabulary at each level versus **peripheral**, good-to-know items. Table 4 shows, for example, the number of vocabulary items shared by number of coursebooks per level, where columns *4 of 4 CB* and *3 of 4 CB* reflect - hypothetically - the number of the **core, need-to-know** items, whereas column *1 of 4 CB* is the vocabulary used in one coursebook only (which reflects subjective author bias), and potentially qualify for **peripheral, good-to-know** vocabulary for L2 learners. By identifying shared versus peripheral vocabulary, we are taking the first step away from subjective lexical selection (typical, for instance, of individual groups of coursebook writers) towards a more objective principled way of wordlist compilation.

In a second step, we compared SVALex to similar resources, namely the EVP and FLELex. The comparison between SVALex and the English Vocabulary Profile (see column *EVP*) shows that SVALex contains more new items per each level than EVP suggests, most probably because SVALex covers receptive lexical knowledge needed for reading comprehension, whereas EVP makes a case for productive vocabulary knowledge used actively in writing. It is worth mentioning that no extra filtering has been applied yet to SVALex, which means that items appearing in one text only, the so-called document hapaxes (8,363 items in total, see columns 5 and 6), are kept in the list.

FLELex is a sister resource of SVALex intended for French learners and was obtained using the same methodology. The number of new items per level, total items per level and number of new MWEs per level is described in Table 3. The main difference between FLELex and SVALex concerns the rhythm of introduction of new items per level. FLELex includes as much as four times more A1 items than SVALex and has more items at the levels A2 and B1. For higher levels, the tendency is reversed and more previously unseen terms appears in SVALex, probably because they correspond to words already included in lower levels in FLELex. It is likely that this pattern, that is repeated for the case of the MWEs, is due to difference in the corpus size. FLELex includes more textbooks per level and has seen more data than SVALex. As a result, lower levels include more peripheral words, i.e. words encountered in only one or two textbooks, in FLELex.

This finding reveal that the corpus size used to train such graded lexicon as SVALex influences the level at which words appear for the first time. This was confirmed by Tack et al. (2016), who used the first level of appearance in FLElex to predict L2 learners' lexical knowledge and found that this criterion was too optimistic, i.e. tends to consider words as known too easily.

The list is available for download on the SVALex platform [3]. The primary use of the list is planned in automatic exercise generation and readability analysis of learner materials. Moreover, for Swedish L2 learners, we have also developed a web interface that allows to query SVALex in

[3]The address of the platform is `http://cental.uclouvain.be/svalex/`.

Level	# items	# new items	# MWE	# doc.hapax	Doc.hapax examples	# EVP
A1	1,157	1,157	92	99	*postnummer* "zip code"	601
A2	3,327	2,432	300	635	*jurist* "lawyer"	925
B1	6,554	4,332	617	1,868	*öga mot öga* "face to face"	1,429
B2	8,728	4,553	880	3,051	*snigelfart* "snail speed"	1,711
C1	7,564	3,160	783	2,709	*inom synhåll* "within eyesight"	N/A

Table 2: The distribution of SVALex entries per CEFR level, including the number of items, new items, multi-words expressions, and number of document hapaxes per level. We also provide the number of new items for English Vocabulary Profile (EVP) for comparison (Capel, 2010).

	Number of shared items in				
Level	4 of 4 CB	3 of 4 CB	2 of 4 CB	1 of 4 CB	Total
A1	115	225	373	775	1,157
A2	306	628	1,206	2,215	3,327
B1	844	1,424	2,510	4,267	6,554
B2	704	1,442	2,067	6,860	8,728
C1	N/A	N/A	1,597	6,248	7,564

Table 4: Shared vocabulary per number of coursebooks (CB) and level. C1, with 2 coursebooks, does not contain any information in the first two columns.

Level	# items	# new items	# new MWE
A1	4,976	4,976	465
A2	6,995	3,516	458
B1	10,780	4,970	604
B2	7,349	1,653	222
C1	8,348	2,122	227
C2	7,433	634	61

Table 3: The distribution of FLELex entries per CEFR level, including the number of items, new items, and multi-words expressions.

a user-friendly manner. As shown in Figure 2, a user can visualize the distribution of a given word across the 6 levels of the CEFR (C2 being always empty) or compare the distributions of two words.

5. Perspectives and conclusions

We described a new lexical resource for L2 Swedish, SVALex, based on knowledge extracted from a L2 corpus related to reading comprehension tasks. We argued that such a resource can be useful to distinguish between core and peripheral vocabulary. The next step of the development is creating a more education-oriented version of SVALex, focused on the core vocabulary and linking every entry to a single CEFR level (at which it should be first introduced), rather than to a frequency distribution. Various methods could be investigated to reach this goal. It is possible to rely on the distribution of words across coursebooks of one level to favour those encountered in various documents. Another possibility would be to tune the training corpus size in order to limit the amount of peripheral terms at lower levels. We hope that such more educationally-oriented resource could help to answer some questions related to L2 vocabulary learning such as "How many words per level should learners know?" or "Which words at which levels?"

Other perspectives include adding lexical information to the resource. To this aim, available information from other free lexical resources for Swedish, such as Lexin, Saldo, Swesaurus (Borin and Forsberg, 2014), etc. could be linked to SVALex items, enriching them with definitions, synonyms, English translations, domain mark-up, valency information, selected corpus examples demonstrating different senses of the words, compound analysis, etc. Furthermore, we are considering making SVALex available in linked open data format using the *lemon* model (McCrae et al., 2011). Some of the aforementioned resources are already available in such a format which would facilitate adopting a similar structure for SVALex.

Finally, as regards the SVALex web platform, we plan to offer more diverse and task-related access to the list. For instance, any individual user could set a target CEFR level, then insert a text in which all higher level words could be highlighted. A further perspective would be to get some feedback from users about the words that are known by them and use this information to define a personalized model of their lexical knowledge.

Adesam, Y., Bouma, G., and Johansson, R. (2015). Defining the Eukalyptus forest – the Koala treebank of Swedish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 1–9.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, 91(4):659–663.

Beacco, J.-C., Blin, B., Houles, E., Lepage, S., and Riba, P. (2011). *Niveau B1 pour le*

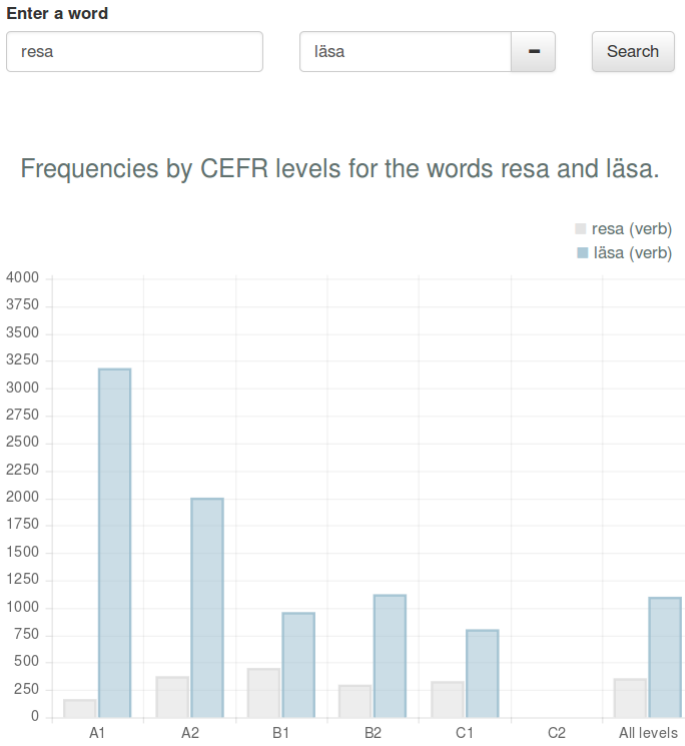

Figure 2: Screen capture of SVALex website, showing the distributions of ”resa” (*to travel*) and ”läsa” (*to read*).

français:(apprenant/utilisateur indépendant): niveau seuil. Didier, Lonrai.

Borin, L. and Forsberg, M. (2014). Swesaurus; or, The Frankenstein Approach to Wordnet Construction. In *Proceedings of the Seventh Global WordNet Conference (GWC 2014)*.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp - the corpus infrastructure of Språkbanken. In *Proceedings of LREC*, pages 474–478.

Borin, L., Forsberg, M., and Lönngren, L. (2013). SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Branco, A., Rodrigues, J., Costa, F., Silva, J., and Vaz, R. (2014). Rolling out text categorization for language learning assessment supported by language technology. In *Computational Processing of the Portuguese Language*, pages 256–261. Springer.

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

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3:1–14.

Carlund, C., Jansson, H., Johansson Kokkinakis, S., and Ribeck, J. (2012). An academic word list for Swedish - a support for language learners in higher education. In *Proceedings of the Swedish Language Technology Conference 2016*.

Carroll, J., Davies, P., and Richman, B. (1971). *The American Heritage word frequency book*. Houghton Mifflin Boston.

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

Ejerhed, E., Källgren, G., O., W., and Åström, M. (1992). *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project*. Publications from the Department of General Linguistics, University of Umeå, no. 33. 1992, Umeå, Sweden.

Forsbom, E. (2006). A Swedish Base Vocabulary Pool. In *Proceedings of the 2006 Swedish Language Technology Conference*.

Francis, W. and Kucera, H. (1982). Frequency analysis of English usage.

François, T., Gala, N., Watrin, P., and Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of International conference on Language Resources and Evaluation (LREC 2014)*, pages 91–102.

François, T. (2014). An analysis of a French as a foreign language corpus for readability assessment. *NEALT Proceedings Series Vol. 22*, pages 13–32.

Heilman, M. J., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, pages 460–467.

Heimann Mühlenbock, K. and Johansson Kokkinakis, S. (2012). SweVoc - A Swedish vocabulary resource for CALL. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*.

Hulstijn, J. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language Proficiency1. *The Modern Language Journal*, 91(4):663–667.

Hult, A.-K., S.-G., M., and E., S. (2010). Lexin - a report from a recycling lexicographic project in the North. In *Proceedings of the XIV Euralex International Congress*.

Jansson, H., Kokkinakis, S. J., Ribeck, J., and Sköldberg, E. (2012). A Swedish Academic Word List: Methods and Data. In *Euralex 2012 Proceedings*.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

Laufer, B. and Ravenhorst-Kalovski, G. (2010). Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a foreign language*, 22(1):15–30.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The semantic web: research and applications*, pages 245–259. Springer.

Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016). Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Volodina, E. and Kokkinakis, J. (2012). Introducing Swedish Kelly-list, a new free e-resource for Swedish. In *LREC 2012 Proceedings*.

Volodina, E., Pilán, I., Rødven Eide, S., and H., H. (2014). You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning.*, volume NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107.