Semi-automatic selection of best corpus examples for Swedish: initial algorithm evaluation

Elena Volodina, Richard Johansson, Sofie Johansson Kokkinakis
elena.volodina@svenska.gu.se
richard.johansson@svenska.gu.se
sofie@svenska.gu.se
Department of Swedish & Språkbanken, University of Gothenburg, Sweden

Abstract

The study presented here describes the results of the initial evaluation of two sorting approaches to automatic ranking of corpus examples for Swedish. Representatives from two potential target user groups have been asked to rate top three hits per approach for sixty search items from the point of view of the needs of their professional target groups, namely second/foreign language (L2) teachers and lexicographers. This evaluation has shown, on the one hand, which of the two approaches to example rating (called in the text below algorithms #1 and #2) performs better in terms of finding better examples for each target user group; and on the other hand, which features evaluators associate with good examples. It has also facilitated statistic analysis of the "good" versus "bad" examples with reference to the measurable features, such as sentence length, word length, lexical frequency profiles, PoS constitution, dependency structure, etc. with a potential to find out new reliable classifiers.

1 Introduction

This evaluation has been carried out as a part of a pre-study partly financed by the Centre for Language Technology (CLT) at the University of Gothenburg.

In this study we have evaluated two different approaches, namely algorithm #1 and #2, to the selection of examples. Both algorithms perform in such a way that, given a number of corpus hits for a search item, examples are sorted withdrawing or awarding points for presence or absence of formalized linguistic features, so called constraints. This brings to the top examples that correspond best to the constraints.

Using a specifically designed user interface and database, we performed the first evaluation. This step has provided us with a body of linguistic evidence for further refinement and tuning of the algorithm in *general* terms for Swedish.

Our hypothesis is, though, that users of different target groups would value presence (or absence) of different linguistic features; and that the same set of parameters cannot satisfy all potential target groups. Moreover, even within different target groups, the definition of a "good example" would change depending upon the practical aim at hand, e.g. examples for learners of different levels will need to take into account different language characteristics.

Thus, during the second iteration planned for near future our intention is to implement a user interface for working with different configurations of extended set of parameters according to the results of the first evaluation. We intend to evaluate parameter configurations again, this time concentrating on whether requirements set on examples differ between different target groups, and different tasks at hand. As a result we hope to suggest optimal parameter configurations for each individual target group, and eventually for different practical tasks at hand.

2 Background

Selection of authentic examples that can appropriately demonstrate vocabulary items of interest is a vital question for lexicographers and L2 teachers. At present it is often unknown for instance, on what principles dictionary examples are selected or where examples for illustrating new vocabulary for L2 learners come from. One way of providing examples is to make them up – they are then as typical as the person that comes up with them thinks they should be, but they lack authenticity. Another way is to use some source of authentic texts, e.g. a linguistic corpus, and select examples using concordance software. The only constraint set on the corpus hits is then the occurrence of the target word in the text span (as opposed to sentence) which makes the number of hits often innumerable. In this case examples are authentic, but the selection process can be very tedious and the quality of "candidate" examples can be very different. One more option is to pre-select sentences automatically using a number of constraints downgrading inappropriate samples. The user is then offered top candidate samples he or she can choose from. The resulting list of ranked candidate sentences can be used for further manual or automatic selection (or editing) of top high-quality sentences, reducing the costs and time spent on manual pre-selection of those. The candidate examples can be used: for dictionary entries; to illustrate language features for students of linguistics; to exemplify vocabulary for language learners; to create test items for L2 learners; to accompany electronic texts (e.g. via clicking on the unknown word the user can see another example of the usage of this word), and eventually for a number of other tasks.

The ranking algorithm can eventually be used to test web texts for appropriateness for inclusion into a corpus. The target user groups are therefore lexicographers, L2 teachers, teachers of linguistics, test item creators, designers of electronic course materials and corpus linguists.

The question arising in this connection is whether we can comprehensively describe and model "good examples". This question has been addressed in different studies (Kilgariff et.al. 2008, Husák 2008, Kosem et.al. 2011, Segler 2007, etc.), though up to date never for Swedish as a target language. Our starting point is that parameters of good examples are language dependent and need to be tested for each language separately.

Algorithms for ranking corpus hits for Swedish have been designed with two practical applications in mind: *Swedish FrameNet* (SweFN, Friberg Heppin and Toporowska Gronostaj 2012) and *Lärka* (Volodina and Borin 2012).

SweFN is a lexical resource under development based on frame semantics, put forward by Charles J. Fillmore. The central idea is that word meanings are described in relation to semantic frames which are schematic representations of the conceptual structures of the language. Work on each frame consists in identifying relevant lexical items and providing authentic corpus (sentence-long) examples for each frame-related meaning. At the moment the work on finding examples involves a tedious look through several hundreds of examples in search of one that is good enough for the task. An algorithm that would be able to sort inappropriate examples away can considerably accelerate work on each frame.

Lärka (Eng. Lark) is an ICALL platform for deploying different language learning activities, at the moment consisting of an exercise generator for linguists and language learners. The language learner part contains a preliminary version of multiple-choice exercise items for vocabulary training. Training context for exercises is at the moment limited to sentences due to copyright restrictions set on most of the corpora available through the Swedish Language Bank. We need, therefore, a reliable automatic approach to selection of appropriate example sentences for language learners, which means, that we need to take into account learner proficiency levels and relevance for different types of vocabulary aspects.

In this study we have evaluated two different approaches to the selection of examples.

In the first algorithm, each example is scored independently of all other examples using a manually defined set of heuristic rules, each of which has an associated weight:

–sentence length: sentences shorter that 10 words or longer than 15 words have 5 points withdrawn for each item not in the range;

–rare words: two relevance points are subtracted for each infrequent word, defined as words above the frequency threshold of 200 based on a frequency list over word forms in the Swedish Wikipedia Corpus;

–keyword position: five points are withdrawn if the keyword item appears after the tenth position in the sentence;

–finite verb: sentences without finite verbs get 100 points withdrawn.

This is in principle similar to the well-known GDEX algorithm often used in lexicography (Kilgariff et.al. 2008, Husák 2008).

In the second algorithm (Borin et.al. 2012), we additionally took into account the intuition that in order to get a good overview of the usages of a word, e.g. to represent different senses of a lexical item in SweFN context, the examples should not only be typical but also different.

This notion of difference is formalized as a similarity metric. The joint optimization of the sum of goodness score according to the heuristic rules and the dissimilarity scores is a computationally intractable problem in general, but can be approximately solved using diversification methods developed in the information retrieval community (Minack et al. 2011). We used a similarity measure based on the Euclidean distance between feature vectors; these vectors represented words in the context of the search terms, as well as a number of syntactic features derived from dependency trees.

The critical question for the present study is whether the two approaches target the parameters that ensure acceptable example ranking; which of the two approaches performs better; what other parameters might be necessary to consider to improve algorithm performance as predictors of good examples. The goal of the study is, in other words, to evaluate the two above-mentioned algorithms; and as a side effect – to identify other potential parameters for Swedish that need to be considered.

3 Related research

Of all the research aimed at selecting authentic examples, the main bulk of studies have been dealing with text readability as opposed to sentence readability. Text readability measures have been explored in a number of studies (Flesh 1948; Björnsson 1968; Huckin 1983; Cedergren 1992; Fulcher 1997; Collins-Thompson and Callan 2004; Mühlenbock and Johansson Kokkinakis 2009; Volodina 2010, etc.); some of them describe CALL and ICALL applications that make use of the measures for automatic selection of texts of appropriate language learner proficiency levels (REAP[1], Read-X[2], Ott & Meurers 2010).

Even though larger contexts, like text, are usually preferred in language learning setting, sentence, nevertheless, cannot be neglected in this discussion. It is a popular linguistic unit when it comes to demonstrating use of vocabulary items for students, e.g. to provide an extra example to usage of an item. In our case it is a necessary limitation imposed by copyright restrictions set on many corpora. Therefore the issue of sentence readability needs to be addressed separately.

When it comes to the source of examples, there have been lively discussions about their nature – should they be authentic, invented or should there be a compromise between the two in the form of simplified corpus examples. Authentic examples, though of course praised by many, are criticized for being rather long and containing too many infrequent words; and that “authenticity” as it is plays greater role for native speakers than for language learners or lexicon users. On the other hand, it is time-consuming to invent examples. Automatic selection of examples from authentic corpora speeds up the process, but it is known to be controversial since the notion of ”good examples” is subjective and often conflicts with the notion of ”authentic examples”. However, it is argued that with semi-automatic approaches using so-called “curation”, i.e. applying human proofreading and editing where necessary, authentic materials can acquire the necessary precision, accuracy and appropriateness (Hubbard, 2012).

Good examples change their characteristics depending upon who is defining them. Most of research within automatic example rating has been done within the domain of lexicography (Kilgariff et.al. 2008, Husák 2008, Kosem et.al. 2011, Didakowski et.al. 2012); only a few studies exploring characteristics of good sentence-long examples within L2 learning (Segler 2007) or aimed at people with special needs (Heimann Mühlenbock 2012).

Regardless of the target group, it has been proven that sentence length is one of the most reliable predictors of sentence readability. Other classifiers vary within different projects and for

[1] http://reap.cs.cmu.edu/

[2] https://sites.google.com/site/elenimi2/read-xpublications

different languages. For example, linguistic features such as sentence length, word frequencies, pronouns, main clauses have been found useful as main predictors of sentence readability for English; punctuation and proper names being used as additional indicators of how well-formed and easy-to-understand a sentence is (Kilgariff et.al. 2008; Husak 2008). The Slovenian team (Kosem et.al. 2011) tested different configurations of linguistic classifiers and compared them in several iterations, having naturalness, typicality and intelligibility as primary criteria for human evaluators. Even sentences showing potential to be turned into good dictionary examples have been considered as good ones. The classifiers that have shown the best predicting ability for Slovene have turned out to be: preferred sentence length, relative keyword position, penalty for keyword repetition, optimal word length.

Different approaches treat linguistic constraints differently. For instance, unlike the English and Slovenian GDEX approaches described above, where all the features are non-obligatory, i.e. none needs to be necessarily met, an approach adopted by the German team (Didakowski et.al. 2012) applies harsher selection. They define a set of parameters with some of them being "hard", i.e. examples are not considered at all if the constraint is not met.

4 Method

Starting from the previous practical and theoretical findings, we designed our evaluation set-up:

Given the two existing algorithms for Swedish, we needed to evaluate their prediction performance on authentic examples and compare them with human judgment. To do that, we selected 60 test items (keywords) from the Swedish Kelly-list, an L2 learner frequency list of modern Swedish (Volodina & Johansson Kokkinakis 2012), taking ten items from each learner proficiency level as defined by Common European Framework of References, CEFR (Council of Europe 2001). Only lexical word classes have been considered, i.e. nouns, verbs, adjectives, adverbs. The number of selected items per word class reflects part-of-speech distribution per CEFR level in the Kelly-list. By having items from a learner-oriented list we tried to address both lexicographers, linguists and L2 teachers as potential user groups.

The 60 items have been sent to the algorithms that made corpus searches in Korp (Borin et.al. 2012a) and ranked the hits. Three top results per algorithm and keyword have been saved in a specially designed database. We kept all the annotations coming from corpora for later statistic analysis of linguistic parameters.

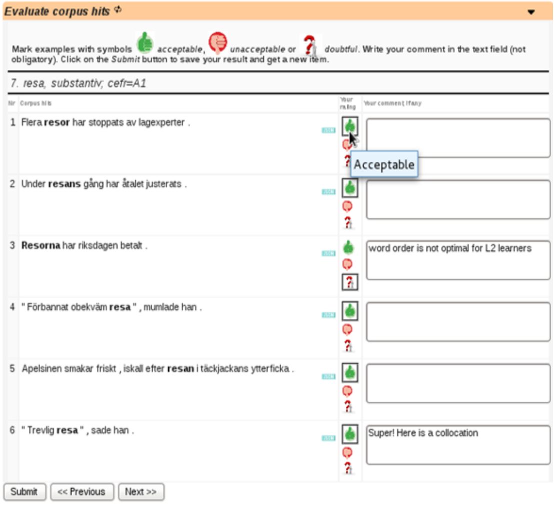
Search for examples was made in several corpora: *SUC* (Stockholm Umeå Corpus), which is often used as the «gold standard» of POS annotation since it has been manually proofread; it amounts to 1,2 mln tokens (Källgren et.al., 2006); *Talbanken*, which is a manually constructed treebank from the 1970s, that is considered to be the «gold standard» of syntactic annotation; the professional prose part used in this project contains 86,000 words (Teleman 1974; Einarsson 1976; Nivre et. al. 2006); and *LäsBarT*, a collection of easy-to-read texts from the 2000s amounting to 1 mln. words (Heimann Mühlenbock 2012).

We initially planned to use only the 3 above-mentioned corpora since they can boast reliability in PoS and syntactic annotations. However, the number of hits for some of the keywords on the list (for CEFR levels B2-C2) proved to be not extensive enough. Therefore to ensure variability of hits per keyword, we added some other corpora, namely; 1) four corpora of fiction prose: *Bonniersromaner I and II* from 1976-1981, *Nordstedtsromaner* from 1999 and *SUC romaner* from 1990s, totaling at about 18 mln words; 2) *PAROLE*, a corpus of mixed texts (novels, newspapers, journals and web text) to balance down the amount of novels (about 24,5 mnl words).

Once the database was populated with corpus examples, the user interface was set up (figure 1) with an option for "voting" for appropriateness of examples: *acceptable* ("thumbs up"), *unacceptable* ("thumbs down"), *doubtful* ("question mark").

We provided a possibility to leave a comment about each example, but it wasn't obligatory. The user was given an opportunity to go back to the previous answers and change them. To avoid any bias in their answers, users were not given information about which of the two algorithms has suggested this or that example sentence. The JSON[3] button, however, (placed in the same cell as examples) reveals all corpus- and user-related information about each example.

[3] JavaScript Object Notation



Evaluate corpus hits

Mark examples with symbols 👍 *acceptable*, 👎 *unacceptable* or ? *doubtful*. Write your comment in the text field (not obligatory). Click on the *Submit* button to save your result and get a new item.

7. resa, substantiv; cefr=A1

Nr	Corpus hits	Your rating	Your comment, if any
1	Flera **resor** har stoppats av lagexperter .	[✓ 👍] 👎 ?	
2	Under **resans** gång har åtalet justerats .	[✓ 👍] 👎 ?	
3	**Resorna** har riksdagen betalt .	👍 👎 [✓ ?]	word order is not optimal for L2 learners
4	" Förbannat obekväm **resa** " , mumlade han .	[✓ 👍] 👎 ?	
5	Apelsinen smakar friskt , iskall efter **resan** i täckjackans ytterficka .	[✓ 👍] 👎 ?	
6	" Trevlig **resa** " , sade han .	[✓ 👍] 👎 ?	Super! Here is a collocation

Figure 1. User interface for evaluation

All users had to evaluate the same population of example sentences. In the result set we had each particular sentence associated with five human votes and optional comments. In addition, sentences contained linked information about which of the algorithms has suggested them, whereas user votes had information about the user target group.

We have asked 5 people to perform evaluation. They come from two different professional backgrounds, some of them working across several subjects, namely: one lexicographer, one lexicographer/computational linguist, and three L2 teachers/computational linguists. Three of them have Swedish as their mother tongue; two others are non-native proficient users of Swedish; all of the participants have doctoral degrees; two of them are men, three are women.

In selection of evaluators the most important factor was that they all are actively involved in the development of the two resources that the algorithms have been developed for – SweFN and Lärka. They are well-trained and qualified to make judgments about example appropriateness and therefore their answers are relevant in terms of requirements set on the example selection.

The users have been instructed to look at every example and assign it a vote (“acceptable”, “unacceptable”, “doubtful”) following the same judgment they would use selecting examples when working with one of the two projects.

This way we collected information about how often human graders agreed with algorithm judgments and could make conclusions about appropriateness of different rating approaches to example selection. Moreover, the optional comments provided us with insights about the linguistic features that we need to take into account in the future versions of algorithms.

A word about bias and limits of this research: we would like to note that four of five participants are computational linguists which supposedly has influenced the type of comments they provided. We presume that their answers are more reasonable in terms of what technology can perform. This might also have influenced their ratings in favor of the algorithms. Users without technical background tend to set higher requirements on technology. We have been aware of that and in fact very interested in their responses since they could help us pinpoint technically reasonable classifiers and predictors which we overlooked from the start.

5 Results and discussion

5.1 Quantitative data

We have looked into how the approach represented by algorithm #1 has performed compared to the approach in algorithm #2 – first in general and then for each target user group, for individual parts-of-speech and finally for learner proficiency levels.

As shown in table 1, algorithm #1 has “won” over algorithm #2 by 6,3% (56,6% to 50,3%). Reasons could be different, one of them being that #2 presents top examples with dispersion built in, i.e. it presents versatility of a lexical item demonstrating it in a group of examples with different realization of meanings and in various syntactic patterns; and thus should be evaluated as a group of examples, rather than individual examples in isolation.

	acc	*unacc*	*doubtful*	*total*
alg# 1	509 56,6%	177 19,7%	213 23,7%	899 100%
alg #2	453 50,3%	242 27%	204 22,7%	899 100%
Total (#1+#2)	962 53,5%	419 23,3%	417 23,1%	1798 100%

Table 1. Distribution of acceptances between the two algorithms.

Algorithm #2 has also suggested more examples that, evaluated individually, were more often found *unacceptable* than the ones suggested by algorithm

#1 (27% to 19,7%). The number of *doubtful* examples, however, is almost equal between the two algorithms (23,7% to 22,7%).

Distribution of acceptances between the two user groups looks as illustrated in table 2.

user groups	*acc*	*unacc*	*doubtful*	*total*
lexico-graphers	458 63,6%	144 20%	118 16,4%	720 100%
alg #1	238 66,1%	67 18,6%	55 15,3%	360 100%
alg #2	220 61,1%	77 21,4%	63 17,5%	360 100%
L2 teachers	504 46,7%	275 25,5%	299 27,7%	1078 100%
alg #1	271 50,2%	110 20,4%	158 29,3%	539 100%
alg #2	233 43,2%	165 30,6%	141 26,1%	539 100%

Table 2. Distribution of votes per user group

Table 2 indicates that lexicographers slightly favored algorithm #1 compared to algorithm #2 (66,1% to 61,1%, *acceptable* examples); the *unacceptable* votes do not either have a clear tendency to distinguish algorithm #1 as a better one (18,6% to 21,4%). Numbers for L2 teachers, however, show an obvious tendency to favor algorithm #1: 50,2% to 43,2% of votes given to *acceptable* examples versus 20,4% to 30,6% to *unacceptable* ones. Here, too, individually well-formed examples from #1 seem to play a more important role for L2 teachers than versatility of a lexical item presented in a group of examples which seems to be important for lexicographers.

An interesting tendency has been shown in ratings viewed from the point of learner proficiency levels.

CEFR levels	*acc*	*unacc*	*doubtful*	*total*
A1	153 51,3%	81 27,2%	64 21,5%	298 100%
alg #1	73 49%	41 27,5%	35 23,5%	149 100%
alg #2	80 53,7%	40 26,8%	29 19,5%	149 100%
A2	146 48,7%	62 20,7%	92 30,7%	300 100%
alg #1	86 57,3%	18 12%	46 30,7%	150 100%
alg #2	60 40%	44 29,3%	46 30,7%	150 100%
B1	143 47,7%	94 31,3%	63 21%	300 100%
alg #1	84 56%	34 22,7%	32 21,3%	150 100%
alg #2	59 39,3%	60 40%	31 20,7%	150 100%
B2	175 58,3%	56 18,7%	69 23%	300 100%
alg #1	91 60,7%	25 16,7%	34 22,7%	150 100%
alg #2	84 56%	31 20,7%	35 23,3%	150 100%
C1	161 53,7%	63 21%	76 25,3%	300 100%
alg #1	83 55,3%	29 19,3%	38 25,3%	150 100%
alg #2	78 52%	34 22,7%	38 25,3%	150 100%
C2	184 61,3%	63 21%	53 17,7%	300 100%
alg #1	92 61,3%	30 20%	28 18,7%	150 100%
alg #2	92 61,3%	33 22%	25 16,7%	150 100%

Table 3. Distribution of votes per learner proficiency level (CEFR-based)

In table 3 we can see a clear tendency of algorithm #1 performing better than #2 for items coming from intermediate proficiency levels B1 and B2, both in terms of higher acceptance and lower rejection rates. This tendency is less clear at levels A2 and C1. Performance per algorithm is strikingly equal for items at levels A1 and C2. Hypothetically, this might indicate that the lower the learner level is, the stricter constraints might need to be applied to example well-formedness to make them appropriate. At intermediate levels (B1, B2) “normally” well-formed examples are much more easily accepted; and the requirement for examples to be well-formed decreases by level C2 (= “proficient language user”); so that both algorithms are performing equally well.

Viewed as a whole, the total number of *acceptable* examples (from both algorithms) is nearly equal to the sum of *unacceptable* and *doubtful* examples: 53,5% versus 46,4%. It means that algorithms suggest 54% of examples that users accept as good ones. This leaves us with the task of improving the rating strategies to offer a higher rate of *acceptable* examples.

5.2 Qualitative data

Analysis of the comments left by the evaluators reveals a range of positive and negative arguments, with critical ones prevailing, which can be summarized as follows: structural, lexical, related to annotation and heterogeneous comments.

1. A large group of comments mention structural features of the sentence, among them:

- *Use of ellipsis*. Elliptic sentences can be found both among the approved examples and among the discarded ones, e.g. `Dämpar inflationen. [Decreases iflation]` or `Sannolikt, sade van Delden. [Most likely, said van Delden].` In both cases, ellipsis has been criticized, e.g. one of the comments says: «elliptical construction; it can function as a possible usage example; but it is not a typical use of this word». A possible approach to this problem could be to check each example for finite verb and subject, and especially check for completeness the clause where the keyword is used.
- *Use of passive*. A recurrent criticism has been aimed at sentences containing passive, even in cases where examples have been marked as *acceptable* ones, e.g. `Midsommarens ritualer genomgicks. [The rituals of Midsummer was explained.]` The evaluator has written: «passive should rather be avoided». Other comments criticized use of passive in combination with other complicating parameters, e.g. *«compounds; plus domain-specific vocabulary; plus passive»* for the rejected example: `Uppgången dämpades i avvaktan på fredagens sysselsättningssiffror för maj. [The increase was dimmed awaiting employment figures for May on Friday.]`
- *Limited context*. Some of the examples have been rejected on the basis of being difficult to understand in the provided context. One of the comments says: «*too short to be illustrative*» for the rejected example `Påstår Gunnar alltså. [States Gunnar that is.]`
- *Non-typical word order*. Example of the rejected *sentence:* `Efter semifinalförlusten känns därför behovet av en föryngring akut. [After the loss in the semi finals the need for a rejuvenation seemed acute.]`
- *Use of anaphora/pronouns*: An example of such is `I åratal hade det sparats till den. [One had been saving up for it for years.]` Use of anaphoric expressions inhibit understanding, therefore presumingly it would be reasonable to avoid sentences where both subject and object are expressed by pronouns.
- *Not appropriate for the learner level*. This type of comment has often been provided for sentences containing a combination of complicated factors, such as unusual (non-frequent) vocabulary, compound words; structurally difficult sentences with inverted word order or long phrase structure, e.g. `Som kandidat till utrikesministerposten utpekades EU-parlamentarikerm Elisabeth Guigou. [Elisabeth Guigou singled out as candidate for the job as minister of foreign affairs.]`

Structural parameters seem to have influenced a lot of decisions against the acceptance of suggested examples. Technically viewed, several of the listed parameters can be easily incorporated into the future algorithms, e.g. restriction on elliptic sentences, on use of passive and pronouns; others, e.g. non-typical word order, might require some brainstorming in terms of which structures to classify towards typical versus non-typical word order; and more importantly in which contexts to classify them as unusual (e.g. for language learners at beginner levels). Limited context is another such parameter. It seems that sentences of the same length can sometimes be sufficiently informative, and at other times highly unrevealing of the word meaning.

A more radical way of treating syntactic characteristics could be a *discriminative approach to target word classes*, e.g. building specific parameter configurations for each word class, e.g. for verbs – check semantic and syntactic valency in a GLDB (The Göteborg Lexical DataBase) (Järborg 1989, http://www.ilc.cnr.it/EAGLES96/rep2/node19.html) and look for identified patterns; for adjectives – look for typical patterns, e.g. keyword in pre-modifier, post-modifier or in attributive positions, etc. Checking statistic results for most frequent structural and lexical patterns for the keyword, so-called word pictures, mutual information, Z-score and other measures of degree of collocality between the keyword and its neighboring words is another possible approach.

2. Second group of comments focuses on lexical features of example sentences:

Stricter word frequency filtering. In many cases examples have been rejected because of the difficult word choice, i.e. containing *domain-specific* or *advanced* vocabulary. One example of such sentence is `Avskrapade smulor av yxorna blandades i bly vid hagelstöpning för att uppnå bättre träffsäkerhet. [Scraped crumbs from the axes were mixed with led at the hail steeping in order to achieve better accuracy.]` A more detailed recommendation has been provided about frequency range of finite verbs (on more than one occasion), for example: «a finite verb should be more frequent one» as a comment for sentence `Tyvärr snuddar också "Studio Ett" en smula vid den sortens generalisering. [Unfortunately, the radio program "Studio Ett" also touches upon that sort of generalization.]`

- *Use of proper names*: general recommendation provided by the evaluators is that proper names should be avoided. Examples criticized for (unusual) proper names can, however, be viewed as good potential examples if some human editing is applied. e.g. `Improjekteatern ger 'Ritualer', en improviserad föreställning i Observatorielunden kl.19 . [The "Improjektteatern" gives 'Rituals', an improvised show in the "Observatorielunden" at 19 o'clock.]`
- *Use of acronyms and abbreviations in* example sentences has been criticized on several occasions
- *Use of compounds*: a repetitive criticism. Swedish is famous for its compounding as a productive word-building pattern. Words can therefore become very long and difficult to interpret, e.g. `Och fredagens relativa marknadslugn kan avläsas i kursdiagrammet för lågräntan [And the calm market on Friday could be read in the stock chart of low interest rate.]`
- *Semantic definition through antonyms/synonyms:* marked as a positive feature in examples like `Sammanbrottet är roligare än bygget`, and `Flanera blir till promenera [The collapse is more fun than the construction]` and `[Strolling becomes walking]`.
- *Keyword repetition:* avoid sentences where target item is used more than once since examples becomes non-explanatory.

Lexical features have proven to be crucial, especially for L2 teachers. Most of the listed parameters would be trivial to implement. When it comes to compounds, available methods for identification of compounds, e.g. via Saldo morphology, need to be checked and tested for reliability. To impose a stricter word frequency filtering we need to consider the type of vocabulary, and therefore underlying word lists, relevant for different purposes and target groups.

3. Third group of comments directs critics at annotation.

Problems with *errors in PoS annotation* result from the fact that we have been using corpora that were not manually proofread, and therefore certain percent of annotation errors can be expected.

However, some of the frustration has been caused by the fact that keywords have been more or less systematically provided as a different part of speech than the one specified, e.g. participles where verbs have been targets; adverbs instead of adjectives; and proper names for nouns. This depends upon *search strategies used in Korp* (Borin et.al. 2012) web service that we are using for primary example selection.

4. The last group of comments is heterogeneous and takes up more general aspects of sentences, such as typicality, metaphoric use, etc.:

- *Prototypical*. Approval of the typical meaning and typical context for the target item, e.g. for the approved sentence `Tidigare verk, "Brödrosten" och "Warszawapakt" var två kortoperor. [The previous works "Brödrosten" and "Warszawapakt"were two short operas.]`
- *Not demonstrative of structural or semantic patterns of the target word*, e.g. for the approved sentence `Ordet "möjligen"`

skrämde mig. [The word "möjligen" scared me.]

- *Metaphoric use*; e.g. for the approved sentence Ljuskänglorna dansar i mörkret. [The light cones were dancing in the dark.]
- *Strange* (as a variant: *not clear*, etc.), for example for the rejected sentence Den skällde skräck och lydnad. [It was barking of fear and obedience.]
- *Abstract use*, e.g. for the example marked as doubtful: Avståndet från 'ätbart' till 'jätteäckligt' är mikroskopisk. [The distance from 'ätbart' to 'jätteäcklingt' is microscopical.]
- *Innovative modern use*, e.g. for the approved sentence Öken, tycker Peter om banan i Lierop. [Desert, Peter thought of the course in Lierop.]

Categories like "strange", "metaphoric", "abstract" are difficult to account for automatically. Hypothetically, strange and abstract examples will be reduced among the top results, once we have improved structural and lexical filtering. Techniques derived from word sense discrimination (Purandare and Pedersen, 2004) can also help us reduce such examples among the top results.

5.3 Statistic data over linguistic features

The rich annotation accompanying each sentence token has become an important source of statistical analysis of *acceptable* versus *unacceptable* examples. Below we are looking into whether *acceptable* examples (for both algorithms) share any common features and how these contrast with the *unacceptable* examples.

Linguistic feature	*acc*	*unacc*
Sentence length, range (tokens)	3-27	3-27
Sentence length, average (tokens)	8	9
Sentence length, mean (tokens)	7	7
Word length, range (characters)	1-23	1-23
Word length, average	5	5

Table 4. Surface features in *acceptable* versus *unacceptable* examples

Values for surface features, such as sentence length and word length presented in table 4 do not seem to be discriminating for example acceptability. The optimal sentence length of 7 tokens suggests that sentences do not have complex phrase structure and do not tend to contain subordinate clauses.

As far as the presence of different word classes is concerned, a summary of indications of the examples is found in table 5.

Linguistic feature, % of sentences	*acc*	*unacc*
Absence of nouns	9%	1
Presence of proper names	29%	29%
Presence of pronouns	27%	64%
Presence of adverbs	36%	44%
Presence of numerals	10%	8%
Presence of conjunctions	12%	20%
Presence of subjunctions	2%	3%

Table 5. Presence of selected word classes in *acceptable* versus *unacceptable* examples

Only 9% of *acceptable* examples contain no nouns at all, which means either use of proper names/pronouns or imperative/elliptical sentences. 73% of the *acceptable* examples do not contain any pronouns at all which presumably depends on the fact that pronouns often make anaphoric references which may be difficult to interpret in one-sentence context. The latter fact might have become the reason for rejection of some examples: we can see that 64% of rejected examples contain pronouns.

In 64% cases of *acceptable* examples, they do not contain any adverbs. This might indicate the fact that sentence structure without adverbials is easier to interpret and is therefore to prefer.

Function words indicating more complex sentence structure, like conjunctions and subjunctions, tend to be absent in the *acceptable* examples, e.g. only in 12% of *acceptable* examples conjunctions are used (versus 20% in *unacceptable*); and only in 2% of accepted sentences subjunctions are used.

Some numbers have been obtained for clause level, such as presence of subjects, finite verbs, subordinate clauses, complex phrase structures (table 6).

Linguistic feature	*acc, nr per sentence, in % of sentences*	*unacc., nr per sentence, in % of sentences*
Subject (S)	0 S: 7,2%	0 S: 11%

Linguistic feature	*acc, nr per sentence, in % of sentences*	*unacc., nr per sentence, in % of sentences*
	1 S: 86% 2 S: 5,8% 3 S: 0,7% 4 S: 0,1%	1 S: 80% 2 S: 7,5% 3 S: 0,5% 4 S: 1,4%
Finite verb(FV)	1 FV: 91,2% 2 FV: 8,1% 3 FV: 0,5% 4 FV: 0,1%	1 FV: 87% 2 FV: 11,4% 3 FV: 0,7% 4 FV: 1%
Subordinate clause (SC)	0 SC: 96% 1 SC: 4%	0 SC: 93% 1 SC: 6% 2 SC: 1%
S-passive (SP)	0 SP: 96% 1 SP: 4%	0 SP: 95% 1 SP: 5%
Complex phrases (CP)	0 CP: 11,2% 1 CP: 55% 2 CP: 29% 3 CP: 4.4%	0 CP: 8,8% 1 CP: 60% 2 CP: 28,5% 3 CP: 1,9%

Table 6. Statistics on the clause level

Though showing only a slight difference between the groups of *acceptable* versus *unacceptable* examples, the group of *unacceptable* examples contains more sentences without subjects (11% vs 7,2%); more examples with multiple subjects (9,4% vs 6,6%); they more often contain several finite verbs (13,1%) compared to the group of acceptable examples (8,7%).

Finally, we calculated the lexical frequency profile for each sentence in the evaluation set, see table 8.

LFP, % of sentence tokens	acc, range	unacc, range	acc, average	unacc, average
Voc, CEFR A1	20-100	20-91	60	58
Voc, CEFR A2	0-50	0-40	6	6
Voc, CEFR B1	0-40	0-33	5,3	5,4
Voc, CEFR B2	0-29	0-33	3,8	3,8
Voc, CEFR C1	0-40	0-40	3,2	3,15
Voc, CEFR C2	0-33	0-33	3	2
Voc, C2+	0-75	0-75	18,8	21,4

Table 8. Lexical frequency information

Lexical frequency information has been collected per lemma using Kelly word list; punctuation has been counted towards A1 items assuming that all language users are familiar with it. Words calculated towards C2+ are the ones not appearing among A1-C2 words in the Kelly list, and are thus assumed to be rare and presumingly more difficult to understand.

Looking at the numbers we have received,we can see that lexical complexity of the *unacceptable* sentences only a few percent higher than of the *acceptable* ones: A1 words, i.e. most frequent ones ones (60% vs 58%); and C2+ vocabulary, i.e. less frequent words (18,8% vs 21,4%). We would need to investigate these numbers further to arrive at any relevant measures for sentence lexical complexity measures.

Therefore, we can summarize that lexical frequency statistics and statistics on clause and phrase levels collected for each example do not straightforwardly explain why *unacceptable* examples have not been approved. It can be said, however, that though numbers concerning vocabulary frequency, phrase structure and clause structure differ only slightly between the groups of *acceptable* and *unacceptable* examples, the tendency for difficulty is more consistent in the group of *unacceptable* examples. Taken in isolation, each parameter differs only slightly between the two groups; however in combination these parameters intensify the "complexity" effect making it unattractive for the the end-users.

6 Concluding remarks

We have presented a series of user evaluations of two automatic algorithms for the selection of illustrative examples from corpora. The first algorithm scored the examples independently of each other based on a few manually defined heuristics, while the second one additionally tried to use a distance function to ensure that the selected set was diverse. Contrary to our intuitions, the simpler algorithm with independent scoring consistently outperformed the complex algorithm taking selection diversity into account. There are several possible reasons for this result: our diversity scoring metric may be too simple, and we may need to make use of techniques derived from word sense discrimination (Purandare and Pedersen, 2004); diversification may be of more interest if the target word is highly polysemous, which we did not take into account when selecting lexical items for our evaluations; we selected fairly small output sets, while diversification may be more necessary for large sets.

In addition to the evaluation of the two algorithms, the user study has given us valuable feedback that can lead to the extension and improvement of the heuristic scoring rules. Several new criteria have been proposed by the users: voice and valency features for verbs, word order, the presence or absence of proper names or acronyms, and the strength of collocation with contextual words. The addition of new scoring rules would make the evaluation function more complex and sensitive, but would also allow us to fine-tune it for particular user groups, such as lexicographers or foreign language teachers.

The algorithms in their final improved form promise to be a useful instrument in applications designed for computer-assisted language learning, for teaching of linguistics, and in lexicographic and linguistic projects. We have plans for embedding the web service for example ranking into Korp[4], Karp[5] and Lärka[6] – all of them applications developed and maintained at the Swedish Language Bank.

References

Carl Hugo Björnsson. 1968. Läsbarhet. Liber Stockholm.

Lars Borin, Markus Forsberg, & Johan Roxendal. 2012a. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA. 474–478.

Lars Borin, Markus Forsberg, Karin Friberg Heppin, Richard Johansson, Annika Kjellandsson. 2012b. Search Result Diversification Methods to Assist Lexicographers. P*roceedings of the 6th Linguistic Annotation Workshop*.

Magnus Cedergren. 1992. Kvantitativa läsbarhetsanalyser som metod för datorstödd granskning. <http://iplab.nada.kth.se/pub_all.jsp> (Retrieved 2007-02-08) Stockholm: Inst.för Numerisk analys och datalogi, Kungl. Tekniska högskolan, NADA.

Kevyn Collins-Thompson and James P. Callan. 2004. A Language Modelling Approach to Predicting Reading Difficulty. *Proceedings of the HLT/NAACL Annual Conference*. Boston, MA, USA.

Council of Europe 2001. *The Common European Framework of Reference for Languages*. Cambridge University Press.

Jörg Didakowski, Lothar Lemnitzer & Alexander Geyken. 2012. Automatic example sentence extraction for a contemporary German dictionary. *Proceedings of EuraLex 2012.*

Jan Einarsson. 1976. Talbanken: Talbankens skriftspråkskonkordans/Talbankens talspråkskonkordans. Lund University.

Rudolf Flesch. 1948 A new readability yardstick. *Journal of Applied Psychology*, Vol. 32, pp. 221–233.

Karin Friberg Heppin, Maria Toporowska Gronostaj. 2012. The Rocky Road towards a Swedish FrameNet – Creating SweFN. *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC-2012); Istanbul, Turkey*. p. 256-261

Glenn Fulcher. 1997. Text Difficulty and Accessibility: Reading Formulae and Expert Judgement. *System* vol.25, 497-513.

Jerker Järborg. 1989. *Betydelseanalys och betydelsebeskrivning i lexikalisk databas*. Göteborg: Inst. f. sv. Spr., Göteborgs universitet.

Katarina Heimann Mühlenbock. 2013. *I see what you mean – Assessing readability for specific target groups*. PhD Thesis, Gothenburg University.

Philip Hubbard. 2012. Curation for systematization of authentic content for autonomous learning.. *EuroCALL 2012 Proceedings*, Gothenburg.

Thomas N. Huckin. 1983. A Cognitive Approach to Readability. In: Paul V. Anderson, R. John Brockmann and Carolyn R. Miller, Editors, *New Essays in Technical and Scientific Communication: Research, Theory, Practice*, Baywood, Farmington, NY, pp. 71–90.

Milos Husák. 2008. *Automatic Retrieval of Good Dictionary Examples*. Bachelor Thesis, Brno. Retrieved on 2010-09-22 from http://is.muni.cz/th/172590/fi_b/bachelor_thesis.pdf

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. *Proc EURALEX*, Barcelona, Spain.

Sofie Johansson Kokkinakis and Elena Volodina. 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project KELLY – issues on reliability, validity and coverage.

[4] http://spraakbanken.gu.se/korp/

[5] http://spraakbanken.gu.se/karp/

[6] http://spraakbanken.gu.se/larka/

Proceedings of eLex 2011, Slovenia.

Iztok Kosem, Milos Husák and McCarthy Diana. 2011. GDEX for Slovene. *Proceedings of eLex 2011*, Slovenia, pp.151-159.

Gunnel Källgren, Sofia Gustafson-Capková and Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Department of Linguistics, Stockholm University.

Enrico Minack, Wolf Siberski, and Wolfgang Nejdl. 2011. Incremental diversification for very large sets: a streaming-based approach. In P*roceedings of the 34th International ACM SIGIR Conference on Research and Development of Information Retrieval*, SIGIR'11, pp. 585--594. New York, United States.

Katarina Mühlenbock and Sofie Johansson Kokkinakis. 2009. LIX 68 revisited - An extended readability measure. *Proceedings of Corpus Linguistics 2009*.

Joakim Nivre, Jens Nilsson & Johan Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006) Genoa*: ELRA. 1392-1395.

Niels Ott and Detmar Meurers. 2010. Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education*. Vol 3, No 1-2. Special issue on "Computer-aided language analysis, teaching and learning: approaches, perspectives and applications" edited by George Weir and Shin'ichiro Ishikawa, 2010.

Amruta Purandare and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning* (CoNLL), pp. 41—48. Boston, United States.

Thomas M. Segler. 2007. *Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German*. Doctoral Thesis. University of Edinburgh. Retrieved on 2010-09-22 from http://www.era.lib.ed.ac.uk/bitstream/1842/1750/3/Segler TM thesis 2007.pdf

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Lund.

Elena Volodina. 2010. Corpora in Language Classroom: Reusing Stockholm Umeå Corpus in a Vocabulary Exercise Generator. *LAP Lambert Academic Publishing*, Colne, Germany.

Elena Volodina and Lars Borin. 2012. Developing an Open-Source Web-Based Exercise Generator for Swedish . *EuroCALL 2012 Proceedings*, Gothenburg.

Elena Volodina & Sofie Johansson Kokkinakis. 2012. Introducing Swedish Kelly-list, a new lexical e-resource for Swedish. *LREC 2012*, Turkey.