Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1326–1331
Marseille, 20-25 June 2022

NyLLex: A Novel Resource of Swedish Words Annotated with Reading Proficiency Level

Daniel Holmer, Evelina Rennes
Department of Computer and Information Science
Linköping University, Linköping, Sweden
{daniel.holmer, evelina.rennes}@liu.se

Abstract

What makes a text easy to read or not, depends on a variety of factors. One of the most prominent is, however, if the text contains easy, and avoids difficult, words. Deciding if a word is easy or difficult is not a trivial task, since it depends on characteristics of the word in itself as well as the reader, but it can be facilitated by the help of a corpus annotated with word frequencies and reading proficiency levels. In this paper, we present NyLLex, a novel lexical resource derived from books published by Sweden's largest publisher for easy language texts. NyLLex consists of 6,668 entries, with frequency counts distributed over six reading proficiency levels. We show that NyLLex, with its novel source material aimed at individuals of different reading proficiency levels, can serve as a complement to already existing resources for Swedish.

Keywords: lexicon, easy language, reading proficiency, text complexity

1. Introduction

Being able to decode and comprehend text is an important skill for participating in the modern society. However, there are many groups of people that, for various reasons, struggle with reading, and could be helped by more accessible texts.

In order to help writers create more accessible texts, easy language guidelines have been developed. For Swedish, such guidelines include the easy-to-read guidelines proposed by Swedish Agency for Accessible Media (MTM, 2021), whereas commonly referred international initiatives include the IFLA guidelines (Misako Nomura and Tronbacke, 2010), the Inclusion Europe guidelines (Inclusion Europe, 2020), and the Plain Language guidelines (PLAIN, 2011). Some of these initiatives are more focused on specific target audiences, and some adopt a broader approach, aiming to create more accessible texts for all.

Although the various sets of guidelines differ, depending on, for instance, intended target audience or type of text, one advise that is common for all easy language writing is to use *simple words*. This is, however, expressed in slightly different ways. For example, the MTM guidelines advise to use *"simple and preferably short words. If a more difficult word is needed, it should be explained in the text where it occurs"*. The IFLA guidelines, on the other hand, say that one should *"avoid difficult words but use language that is adult and dignified. Unusual words should be explained through context clues"*.

Thus, it is clear that easy language texts should use *simple* words, and avoid *difficult* words. But how can we know, then, if a word is simple or not? A common way of determining this is to look at the relative frequency of a word in a corpus. This gives an indication of how common a word is, and thus, might say something about how simple it is, as more common, everyday words presumably are more familiar to the reader, thus, simpler. However, more frequent words might also be more polysemous, which conversely indicates higher complexity (Alfter, 2021), and the distribution of frequencies is highly dependent on the reference corpus used since factors as the genre and text type (i.e. texts of spoken or written words, targeting audiences of different demographics, etc.) will be reflected in the word frequency distribution (Wengelin, 2015). Another aspect of word complexity is whether the word domain is known to the individual reader or not. A low-frequency word, such as the term *gambit* in a chess domain, could be perfectly readable to a reader with knowledge in chess. Other features that could indicate word complexity include word length or age of acquisition.

Identifying complex words is not an easy task. One way of gaining further knowledge about word complexity is to compile lexical resources of texts targeting poor readers. Such resources can provide an important source of information about the receptive vocabulary of the target audience, which by extension could be used, for instance, in automatic text simplification applications, or for measuring text complexity. For instance, the classical Dale-Chall readability formula (Chall and Dale, 1995) utilises a list of simple words in its calculation, where the ratio of difficult words (i.e. words not occurring in the list) is used for estimating the readability of a text.

Nypon och Vilja förlag is Sweden's largest publishing company for texts in easy language, targeting both youths and adults, and the published books are graded according to reading proficiency level. In this article, we aim to compile a lexical resource from this material, resulting in a graded vocabulary of easy Swedish, that could be used for lexical simplification applications as well as text complexity assessment.

For Swedish, some similar resources exist, for instance SWEVOC (Heimann Mühlenbock and Johansson Kokkinakis, 2012), a Swedish base vocabulary consisting of about 7,600 Swedish lemmas, divided into sub-categories. SVALEX (François et al., 2016) is a lexicon of approximately 16,000 words originating from the COCTAILL corpus (Volodina et al., 2014), a corpus of reading comprehension texts collected from coursebooks targeting L2 learners of Swedish. Whereas SVALEX contains the receptive vocabulary of L2 learners, the SWELLEX (Volodina et al., 2016b) lexicon is focusing on productive vocabulary. SWELLEX is extracted from the SweLL corpus (Volodina et al., 2016a), containing essays written by second language learners of Swedish. Both SVALEX and SWELLEX are annotated with the CEFR level classification indicating reading proficiency level.

In this article, we aimed to 1) create an easy language lexical resource which complements the available resources for Swedish with regards to source material and target audience, and 2) validate the resource by examining its unique aspects and overlaps with similar resources.

2. Method

This section presents the procedure of deriving the lexical resource out of the source material, and describes the techniques used for filtering, preprocessing, and frequency estimation.

2.1. Material

The source material consists of 247 books from *Nypon och Vilja förlag* where each book is classified according to the readability gradation presented in Table 1. As a validation of the reading proficiency levels, we calculated LIX (Björnsson, 1968), a commonly used metric to indicate readability in Swedish texts. LIX is given by Equation 1, where $n(w)$ denotes the number of words and $n(s)$ the number of sentences.

$$\text{LIX} = \frac{n(w)}{n(s)} + \left(\frac{n(words\ >\ 6\ chars)}{n(w)} \times 100\right) \quad (1)$$

As seen in Figure 1, the LIX value does indeed increase in conjunction with the reading proficiency levels.

In the source material, the books vary in genre, and include both fiction and non-fiction. Some of them belong to the same series of books, for example; the *Gå till ...* [Go to...] series, describing various everyday tasks (such as going to the library or the dentist), aimed mainly at L2 learners; and the *Ett liv* [A life] series of easy-to-read biographies adapted for different levels of education. Furthermore, there are also easy-to-read versions of classic novels like *Kallocain* by Karin Boye and *1984* by George Orwell, amongst others. All the books were provided in digital format as PDF files.

Level	Description
Level 1	Each page contains very little text. Simple words and sentences. Many illustrations that support the story.
Level 2	Everyday language. The text is divided into short paragraphs with short line lengths. The content depicts relatable situations and focuses on sequences of events. In the books targeting a younger audience, there are illustrations for each spread.
Level 3	Well-known words and expressions. The story is chronologically presented and the connection between cause and effect is clear. There is a sequence of events and descriptions of characters and environment. A number of illustrations. The graphical form includes larger font size, line spacing and margins.
Level 4	Chapter books with few or no illustrations. Slightly more difficult names, words and expressions and longer sentences. The graphical form is spacious with large font size.
Level 5	Adopts an easy-to-read focus regarding language, content and graphical form, but presents a larger challenge to the reader.
Level 6	Books produced with special care regarding language, content and graphical form. Books at this level are supposed to be a gateway to traditional books.

Table 1: The Nypon reading proficiency levels, interpreted and loosely translated by the authors. The complete descriptions of each level can be found at `https://www.nyponochviljaforlag.se/om-oss/om-lattlast/lattlastnivaer/lattlastnivaer-nypon/`

2.2. Preprocessing

Each book was parsed from PDF to plain text with a Python implementation of Tika Parser[1]. Since the format of the books varied widely, and no catch-all rules could be applied to exclude all non-textual data in the books, the parsing was followed by manual corrections. For example, in some cases we removed parts of the front matter (excluding any prologues) and back matter (excluding any epilogues). In other cases, mainly in the lower reading proficiency levels, we manually corrected sentences that had no space separation. This problem most probably stemmed from the presence of a mix of illustrations and text in varying horizontal alignments on the same page, where the parser had a hard time retaining the text formatting.

[1] `https://github.com/chrismattmann/tika-python`

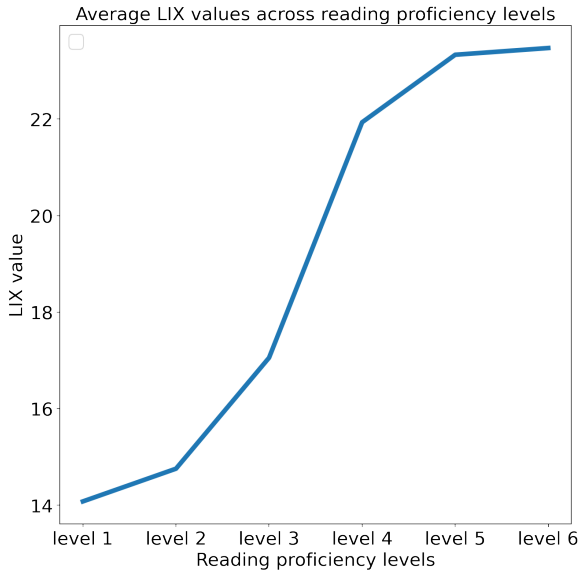

Figure 1: Average LIX values of the books in the different reading proficiency levels .

	Books	Tokens	
		Before	After
Level 1	57	23301	22942
Level 2	47	83990	82998
Level 3	60	212000	208723
Level 4	64	362289	352595
Level 5	17	110476	106966
Level 6	2	22573	22007
Total	**247**	**814629**	**796231**

Table 2: Number of books and tokens in the Nypon dataset. Before and after filtering.

The plain texts were subsequently tokenised, part-of-speech tagged, and lemmatised in the EFSELAB pipeline[2]. The EFSELAB pipeline has shown promising results compared to other taggers (Östling, 2018) and has successfully been implemented in similar projects (see for instance Megyesi et al. (2016)). Each word was kept in its lemma form together with its part-of-speech tag, meaning that a word that appears with two different part-of-speech tags has two separate entries in the resource. We excluded all entries that were tagged as punctuation (the SUC[3] tags MAD, PAD, and MID) as well as all personal names (listed as the SUC tag PM). Entries tagged as an ordinal number (SUC tag RO) in the form similar to *2:a* [2nd], *3:e* [3rd] were manually merged to their base lemmas; *andra* [second] and *tredje* [third]. A description of the SUC tags can be found in the SUC manual (Gustafson-Capková and Hartmann, 2006). In some cases, page numbers were present in the texts even after the initial pre-processing step. To ensure the fairness of the resource, all digit form cardinal numbers were excluded.

In line with similar resources, NYLLEX includes multi-word expressions (MWEs). These were identified by matching *n*-grams in the source texts to MWE entries in the SALDO lexicon (Borin et al., 2013) .

2.3. Filtering

The vocabulary of the full Nypon dataset consists of a total of $16,841$ entries spread across the six readability levels. $10,167$ (61.7%) entries were only found in one of the readability levels. Of these level-specific entries, $6,781$ were unique and only appeared once in the whole dataset, while 3386 could be found multiple times in the level. Many of the level-specific entries were a consequence of errors during the reading of the PDFs, even after the manual corrections of the raw texts produced by the PDF reader. There was also a substantial number of entries that represented a correct word, but where most of the usage stems from very specific contexts[4]. In order to get a more representative vocabulary, with fewer outliers and highly specific entries, we applied a filter to remove these kinds of entries. We tried filters that only kept entries that had a specified raw frequency count of N, but found that while this approach was fairly effective for pruning rare words, the entries that were faulty readings from the PDFs were not easily caught with a relatively low N. As mentioned before, the faulty readings mainly stemmed from graphical elements in the books. Since these elements tended to reoccur several times in the same book, the same kind of error was often repeated, resulting in multiple instances of the same faulty entry. The issue then became to find an N high enough to remove even repeated faulty entries, but low enough to not remove too many correct entries. However, since the exploration of the best value of N would require a manual evaluation of the entire resource for a wide range of Ns, we used the approach applied in Forsbom (2006), where a filter based on genre[5] contribution was applied. Since the goal was to keep as many lemmas as possible, we also set the threshold as low as possible, and filtered out entries that were not present in at least two of the six reading proficiency levels. This approach was found to strike a good balance in the pruning of both rare and faulty entries.

As seen in Table 2, even though the filter removed 62% of the unique entries, our resource still covers 98% of the total number of tokens present in the dataset.

[2] https://github.com/robertostling/efselab

[3] https://spraakbanken.gu.se/resurser/suc2

	Total entries	Entry overlap	New entries
NYLLEX	6,668	-	-
SVALEX	15,686	4,544 (68.15%)	2,124 (31.85%)
SWELLEX	6,967	2,733 (40.99%)	3,935 (59.01%)
SWEVOC	7,408	3,505 (52.53%)	3,163 (47.47%)

Table 3: Comparison of the final resource to other resources.

2.4. Manual editing

Similarly to SVALEX, the manual editing of NYLLEX was carried out in a circular fashion that allowed for the frequency estimations to be recalculated when a correction was made to an entry. We also took advantage of the fact that SVALEX had undergone this process of manual editing before. As seen in Table 3, the total overlap of the NYLLEX and SVALEX is 4,544 entries, all of which could be seen as correct. Initially that left us with 2,214 entries present in NYLLEX, but not in SVALEX, all of which became the subject of our manual editing. In total, 83 of the 2214 entries were flagged as erroneous, and subsequently manually assigned to their correct lemma and word forms. The remaining 2,130 no-matches with SVALEX were thus correct, but novel, NYLLEX entries.

2.5. Frequency estimation

Each vocabulary item of the resource is associated with a per-level frequency estimation. For the calculation of frequencies across each of the six readability levels, we followed the procedure described in SVALEX (François et al., 2016) using dispersed frequency. For calculating frequencies on the dataset as a whole, we used adjusted frequency, which has previously been used for ranking the words of the Swedish Base Vocabulary Pool (SBVP) (Forsbom, 2006).

3. Final Resource Description

The final resource totals $6,668$ items (of which 443 are MWEs) distributed over six reading proficiency levels as defined by *Nypon och Vilja förlag*. Each item is not restricted to a single level, and since the filter method we applied (each entry must be present in at least two levels) removed all hapaxes, the rarest items in our resource are the items present only once in two levels respectively. Our resource includes 771 of these rare items, of which 50.2% are nouns (e.g. *kärlekshistora* [love story] and *ledtråd* [clue]). Their distributions across different reading proficiency levels are listed in the column *Rare entries* in Table 4. Conversely, 922 entries are present in all of the six levels. It comes as no surprise that the most frequent entries in this category are highly used words like *jag* [I], *vara* [be], and *och* [and].

As the well-established LIX metric indicates (see Figure 1), the complexity of the books increases in conjunction with the reading proficiency level. Another (simple, but fairly effective) metric for text complexity is the average word length of the text (see for example Falkenjack et al. (2013)), where longer average word lengths indicate a more complex text. This increase is, however, not clearly visible in the average entry lengths in NYLLEX. Whereas level 1 is observed to be slightly lower, the rest of the levels follow no obvious pattern. The same trend can be seen with the number of MWEs present across different levels. While Volodina et al. (2016b) found that the number of MWEs increased steadily with a higher CEFR level, the clearest pattern is that level 1 is the odd one out with a lower proportion of MWEs compared to the other levels. On the other hand, levels 2 through 6 can be seen to have a fairly similar MWE ratio. Why this happens is an interesting topic for further study, and we plan to investigate additional text complexity metrics on the same dataset.

We also compared NYLLEX with the aforementioned similar resources for Swedish; SVALEX, SWELLEX, and SWEVOC and the results of this comparison are presented in Table 3. It should be noted that NYLLEX had a varying degree of overlap with the other resources. SVALEX, which is the most extensive of the other resources, was found to have the highest overlap, with $4,544$ identical entries. For the two comparatively smaller resources, SWELLEX and SWEVOC, the percent of overlaps was also lower. For SWEVOC, $3,505$ entries overlapped, while $3,169$ entries were novel to NYLLEX. For SWELLEX, $2,733$ entries overlapped, while $3,941$ entries were novel. The resource size probably accounts for most of the difference in overlap percentages between SVALEX and SWEVOC and SWELLEX. Additionally, the fact that SWELLEX is a productive vocabulary sets it apart from the other resources, and can possibly explain some of the proportionally large overlap difference between SWELLEX and SWEVOC, even though they are similar in size.

Since SVALEX was the most similar resource to NYLLEX in terms of overlapping entries, and has similar structure with words annotated at different reading proficiency levels (CEFR), we did a more extensive comparison of the overlap of our resource and SVALEX. This comparison is presented in Table 5. Level 1–6 correspond to the six Nypon reading proficiency levels, and A1–C1 correspond to the CEFR levels. Each row displays the proportion of Level 1–6 entries also present in each of the CEFR levels of SVALEX. Even though level 6 has slightly more entries represented than level 1, it only consists of two books (see Table 2) compared to the 57 books of level

[4]For example; *kafferast* [coffee break], *debattera* [to debate], and *kakdeg* [cookie dough]

[5]In this work each reading proficiency level is seen as a genre

	Entries	MWEs	Avg. entry length	Rare entries
Level 1	1,876	72 (3.8%)	5.26	45 (2.4%)
Level 2	3,347	206 (6.2%)	5.53	151 (4.5%)
Level 3	5,145	315 (6.2%)	5.95	413 (8.1%)
Level 4	6,147	382 (6.3%)	6.14	556 (9.2%)
Level 5	4,386	250 (5.7%)	5.89	304 (6.9%)
Level 6	2,087	108 (5.6%)	5.61	73 (3.8%)
Total	**6668**	**443 (6.6%)**	**6.2**	**771 (11.5%)**

Table 4: Overview of the final resource.

1. It is therefore hard to draw any general conclusions about level 6. For the rest of the levels, although the degree of overlap between the resources is quite modest, there seems to be a slight trend toward a higher Nypon reading proficiency level also having a higher degree of overlap with a higher CEFR level.

We believe that this small, but nonetheless clearly established, overlap further strengthens the validity of the Nypon reading proficiency levels since it indicates a general alignment with the more widely used CEFR reading proficiency scale. Furthermore, as shown in Table 3, NYLLEX introduces a significant number of new words compared to similar resources. NYLLEX can therefore serve as a complementary resource and improve the performance of applications relying on Swedish readability-graded lexical resources.

	A1	**A2**	**B1**	**B2**	**C1**
Level 1	20.5%	21.2%	15.5%	13.1%	13.0%
Level 2	16.2%	21.9%	19.3%	17.4%	16.7%
Level 3	13.3%	21.9%	22.6%	21.4%	20.0%
Level 4	12.0%	20.8%	23.8%	22.8%	21.4%
Level 5	14.0%	20.9%	22.0%	20.5%	20.0%
Level 6	16.2%	18.5%	15.8%	13.8%	14.2%

Table 5: Overlap with CEFR levels in SVALEX.

4. Conclusion

We presented NYLLEX, a new lexical resource of words annotated with six reading proficiency levels.

When compared to similar resources, we saw that NYLLEX did not overlap completely with any of the resources. Thus, we believe that NYLLEX could work as a complementary lexical resource, which could be used for further work in, for instance, applications for text complexity assessment or lexical simplification. NYLLEX will be made freely available.

5. Acknowledgements

Thank you, *Nypon och Vilja förlag* for generously sharing the material making this resource possible.

6. Bibliographical References

Alfter, D. (2021). *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, Department of Swedish, University of Gothenburg, Gothenburg, Sweden.

Björnsson, C. H. (1968). *Läsbarhet*. Liber, Stockholm.

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.

Chall, J. S. and Dale, E. (1995). *Readability revisited: The new Dale–Chall readability formula*. Brookline Books, Cambride, MA.

Falkenjack, J., Heimann Mühlenbock, K., and Jönsson, A. (2013). Features Indicating Readability in Swedish Text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway*, number 085 in NEALT Proceedings Series 16, pages 27–40. Linköping University Electronic Press.

Forsbom, E. (2006). A Swedish Base Vocabulary Pool. In *Swedish Language Technology conference, Gothenburg*.

François, T., Volodina, E., Pilán, I., and Tack, A. (2016). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Gustafson-Capková, S. and Hartmann, B. (2006). Manual of the Stockholm Umeå Corpus version 2.0. `https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf`.

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*, pages 28–34, Lund, 10. Linköping University Electronic Press.

Inclusion Europe. (2020). Easy-to-read checklist. checklist to make sure your document is easy to read. `https://www.inclusion-europe.eu/wp-content/uploads/2020/06/Easy-to-read-checklist-Inclusion-Europe.pdf`. Accessed: 2021-09-30.

Megyesi, B., Näsman, J., and Palmér, A. (2016). The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Proceedings of the Tenth*

International Conference on Language Resources and Evaluation (LREC'16), pages 3192–3199.

Misako Nomura, G. S. N. and Tronbacke, B. (2010). *Guidelines for Easy-to-Read materials*. International Federation of Library Associations and Institutions.

MTM. (2021). Att skriva lättläst. `https://www.mtm.se/var-verksamhet/lattlast/att-skriva-lattlast/`. Accessed: 2021-10-05.

PLAIN. (2011). Federal Plain Language Guidelines, revision 1. `https://www.plainlanguage.gov/media/FederalPLGuidelines.pdf`. Accessed: 2021-10-18.

Volodina, E., Pilán, I., Eide, S. R., and Heidarsson, 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*, pages 128–144.

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G., and Sandell, M. (2016a). SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Volodina, E., Pilán, I., Llozhi, L., Degryse, B., and François, T. (2016b). SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 76–84.

Wengelin, Å. (2015). Mot en evidensbaserad språkvård? en kritisk granskning av några svenska klarspråksråd i ljuset av forskning om läsbarhet och språkbearbetning. *Sakprosa*, 7(2).

Östling, R. (2018). Part of speech tagging: Shallow or deep learning? *North European Journal of Language Technology*, 5:1–15.