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

Daniel Holmer, Evelina Rennes

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.

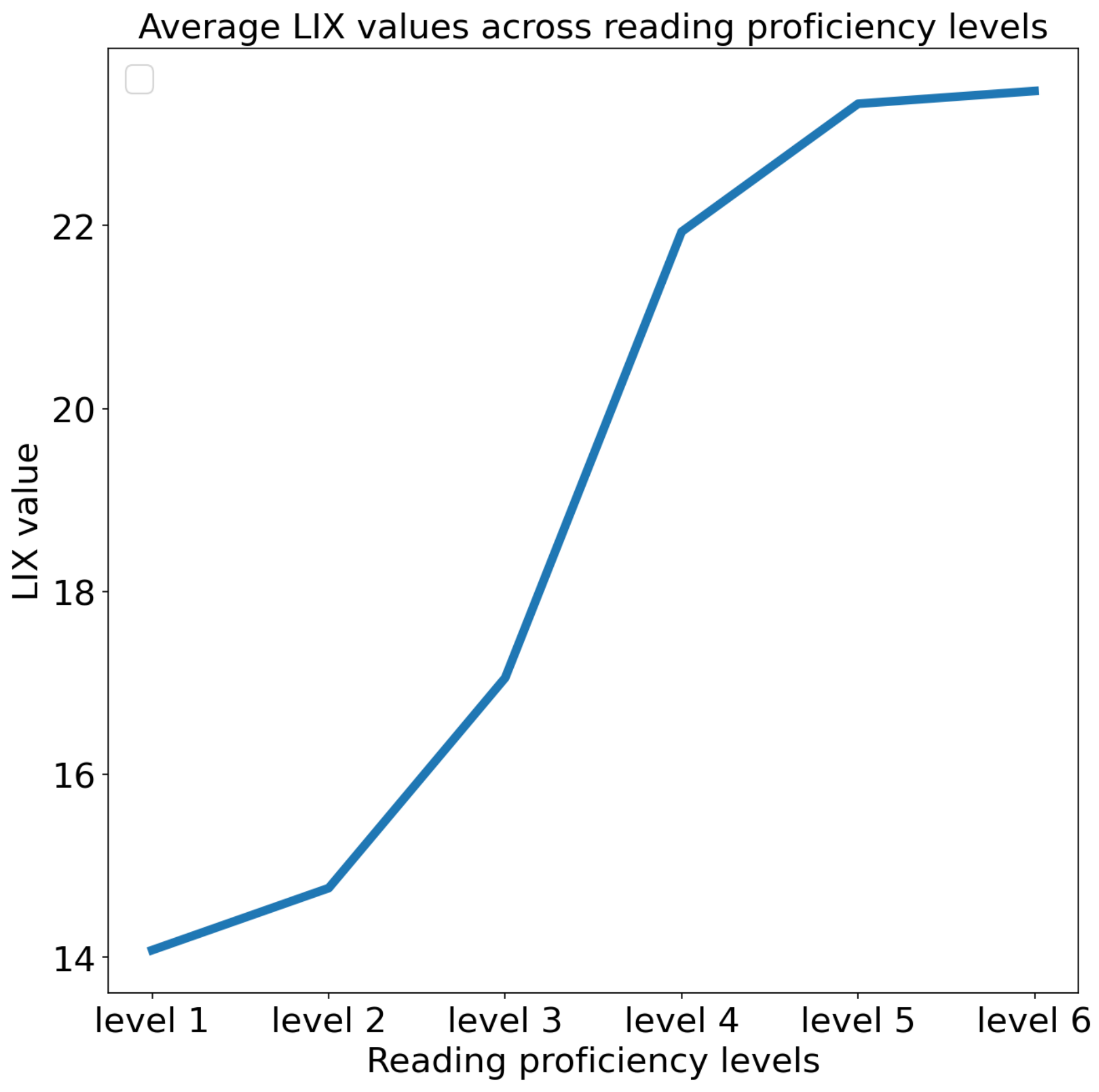
Material

The material consists of 247 books from *Nypon och Vilja förlag.* The books are annotated by experts at different reading proficiency levels.

The scale spans from level 1 to level 6, where level 1 contain the easiest texts and level 6 the most complex texts. The genres of the books vary, and they are a mixture of both fiction and non-fiction.

Level	# books
Level 1	57
Level 2	47
Level 3	60
Level 4	64
Level 5	17
Level 6	2
Total	**247**

Proficiency levels & LIX



Nypon reading proficiency levels

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.
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.

Preprocessing & filtering

- Tokenized, lemmatized, and POS-tagged.
- Identified multiword-expressions by the use of the lexicon SALDO.
- Removed entries not present in at least two levels.

	Tokens (before filtering)	Tokens (after filtering)
Level 1	23,301	22,942
Level 2	83,990	82,998
Level 3	212,000	208,723
Level 4	362,289	352,595
Level 5	110,476	106,966
Level 6	22,573	22,007
Total	**814,629**	**796,231**

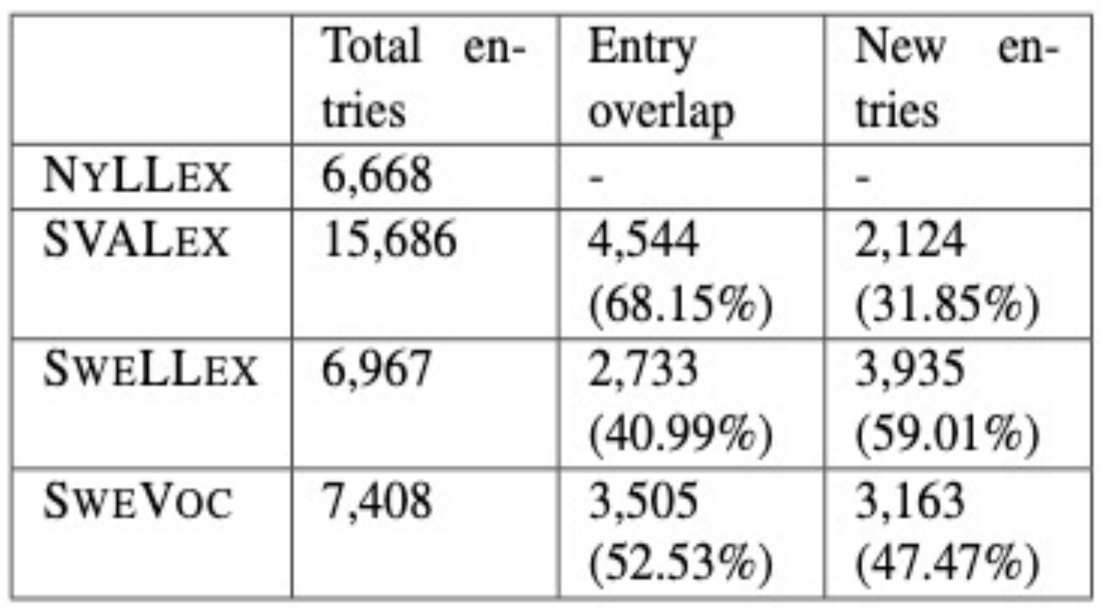
Final resource description

- 6,668 entries
- Annotated with **dispersed frequencies** across the different proficiency levels, and **adjusted frequency** for entries in the whole resource

	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	**6,668**	**443 (6.6%)**	**6.2**	**771 (11.5%)**

Overlap with other resources

We compared NyLLeX with similar Swedish resources, SVALex[1], SweLLex[2], and SweVoc[3].



	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%)

Conclusions

- NyLLex is a new lexical resource annotated with six reading proficiency levels.
- NyLLeX complements similar resources.
- NyLLeX can be used for text complexity assessment or lexical simplification.

Acknowledgements

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

References

1. Francois, 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)*
2. Volodina, E., Pilán, I., Llozhi, L., Degryse, B., and Francois, T. (2016). 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.
3. Heimann Mühlenbock, K. and Johansson Kokkinakis, S. (2012). SweVoc – A Swedish vocabulary re- source for CALL. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*