# LEGATO: A flexible lexicographic annotation tool

**David Alfter**
Språkbanken
Department of Swedish
University of Gothenburg
Sweden
david.alfter@gu.se

**Therese Lindström Tiedemann**
Department of Finnish, Finno-Ugrian
and Scandinavian Studies
University of Helsinki
Finland
therese.lindstromtiedemann@helsinki.fi

**Elena Volodina**
Språkbanken
Department of Swedish
University of Gothenburg
Sweden
elena.volodina@gu.se

## Abstract

This article reports an ongoing project aimed at analyzing lexical and grammatical competences of Swedish as a Second language (L2). To facilitate lexical analysis, we need access to linguistic information about relevant vocabulary that L2 learners can use and understand. The focus of the current article is on the lexical annotation of the vocabulary scope for a range of lexicographical aspects, such as morphological analysis, valency, types of multi-word units, etc. We perform parts of the analysis automatically, and other parts manually. The rationale behind this is that where there is no possibility to add information automatically, manual effort needs to be added. To facilitate the latter, a tool LEGATO has been designed, implemented and currently put to active testing.

## 1 Introduction

Lexical competence has been acknowledged as one of the most important aspects of language learning (e.g. Singleton, 1995; Milton, 2013; Laufer and Sim, 1985). Some claim that we need to understand 95–98% of the words in a text to manage reading comprehension tasks (cf. Laufer and Ravenhorst-Kalovski, 2010; Nation, 2006; Hsueh-Chao and Nation, 2000). It has also been observed that vocabulary is actively taught at all levels of L2 proficiency courses with a tendency to be dominating at more advanced levels in comparison to other linguistic skills, see for example findings from a course book corpus COCTAILL (Volodina et al., 2014, p.140). Lexical features have also been found to be one of the best predictors in text classification studies (e.g. Pilán and Volodina, 2018; Xia et al., 2016; Vajjala and Meurers, 2012) with important implications to the area of educational NLP. Deciding on which vocabulary to use and include is thus an important part of teaching a foreign language, in designing course materials and tests. In theoretical descriptions of L2 acquisition, lexical knowledge was previously "side-lined" according to Milton, but within academic circles its place has been "significantly revised" and received an increasing amount of interest over recent decades (Milton, 2013).

There are multiple characteristics of vocabulary that are interesting from the point of view of both theoretical analyses, as well as for pedagogical and NLP-based applications. Such characteristics include, among others, vocabulary size & breadth (e.g. Nation and Meara, 2010; Milton, 2013), corpus frequency (Dürlich and François, 2018; François et al., 2016), word family relations (Bauer and Nation, 1993), syllable structure, morphological characteristics, semantic relations, topical domain categorization (Alfter and Volodina, 2018), and many others (e.g. Capel, 2010, 2012).

While frequency information comes from corpora, most linguistic characteristics are non-trivial to acquire by automatic methods and require either manual effort or access to manually prepared resources – lexicons being the most extensive and reliable sources for that. However, dictionaries and lexicons are often proprietary resources (e.g. Sköldberg et al., 2019), which complicates automatic lexicon enrichment. Among freely available lexicons for Swedish, we can name Saldo (Borin et al., 2013), Swesaurus (Borin and Forsberg, 2014), Lexin (Hult et al., 2010) and a few other resources provided through Språkbanken's infrastructure Karp (Borin et al., 2012), although, even there many aspects of vocabulary are not documented, e.g. the transitivity of verbs, the morphological structure of the words (root, prefix, suffix) or the topical domain of the words.

To circumvent the problem of access to the information that may prove crucial in the context of the current project for the three outlined areas of application (theoretical studies, pedagogical studies/applied linguistics and educational NLP), we have initiated semi-automatic annotation of learner-relevant vocabulary interlinking available resources with manual controls of those, and adding missing aspects manually. The work is ongoing, and below we present the reasoning around this annotation process and the main components of the system that facilitate that.

## 2 Second language profiles project

In the current project, *Development of lexical and grammatical competences in immigrant Swedish* funded by Riksbankens Jubileumsfond, the main aim is to provide an extensive description of the lexical and grammatical competence learners of L2 Swedish possess at each CEFR[1] level, and to explore the relation between the receptive and productive scopes. The exploration of the grammatical and lexical aspects of L2 proficiency is performed based on two corpora, COCTAILL (Volodina et al., 2014), a corpus of course books used in teaching L2 Swedish and the SweLL-pilot (Volodina et al., 2016a), a corpus of L2 Swedish essays. The corpora are automatically processed using the SPARV pipeline (Borin et al., 2016), and include, e.g., tokenization, lemmatization, POS-tagging, dependency parsing, and word sense disambiguation.

## 3 LEGATO tool

LEGATO[2] - **LE**xico**G**raphic **A**nnotation **TO**ol - is a web-based graphical user interface that allows for manual annotation of different lexicographic levels, e.g. morphological structure (root, affix etc), topic, transitivity, type of verb (e.g. auxiliary, motion verb), etc. The interface shows a lemgram for a given word sense, the part of speech and the CEFR level, as well as the Saldo sense and the primary and secondary sense descriptors used in Saldo (Borin et al., 2013), and up to three example sentences taken from the COCTAILL corpus. If there are fewer than three sentences available at the target CEFR level, the maximum number of sentences found is shown. It also features search, filter and skip functionalities as well as external links to other information sources such as Karp (Ahlberg et al., 2016); SAOL, SO & SAOB via svenska.se (Malmgren, 2014; Petzell, 2017); and the Swedish Academy's Grammar (SAG, the main grammar of the Swedish language) (Teleman et al., 1999). Figure 1 shows the user interface for the annotation of *nominal type* category.

### 3.1 Data for lexicographic annotation

For lexical analysis, we generate word lists (SenSVALex and SenSweLLex) based on senses from the two linguistically annotated corpora, both lists being successors of the lemgram-based ones from the same corpora (François et al., 2016; Volodina et al., 2016b). The lists contain accompanying frequency information per CEFR level according to the level assigned to the texts/essays where they first appear. In practical terms, the task of preparing a resource for lexical studies involves:
1. labeling all items for their "target" level of proficiency – that is, the level at which the item is expected to be understood (receptive list) or actively used (productive list). The CEFR level of each item is approximated as the first level at which the item appears, i.e. the level would be B2 for entry X if it was first observed at level B2 (cf. Gala et al., 2013, 2014; Alfter and Volodina, 2018).
2. interlinking items with other resources for enrichment, e.g. adding information on adjective declension
3. manually controlling the previous step for a subset of items to estimate the quality
4. setting up an annotation environment for adding missing information.

While (1) above has been partially addressed by Alfter et al. (2016) and Alfter and Volodina (2018), steps (2–4) are described shortly in the sections below.

### 3.2 Automatic enrichment

An overview of linguistic aspects annotated using LEGATO is provided in Table 1. All aspects are kept as close as possible to the terminology and the description of Swedish grammar in SAG (Teleman et al., 1999). A subset of those aspects, marked as *A* or *A-M* in Table 1 (column "Mode") are annotated automatically using a range of available resources mentioned in the column "Resources for auto-enrichment". Other aspects are added manually (*M*) following guidelines[3] explaining choices

[1] CEFR = Common European Framework of Reference (Council of Europe, 2001)

[2] `https://spraakbanken.gu.se/larkalabb/legato`; user "test" for testing purposes

[3] `https://urlzs.com/PZoRm`

| Aspect | Explanation / choices | Mode | Resources for auto-enrichment |
|---|---|---|---|
| 1 Adj/adv structure | comparisons: periphr.: *(mer/mest) entusiastisk*; morph.: *vacker-vackrare-vackrast*; irreg.: *god-bra-bäst* | A-M[2] | Saldo-Morphology |
| 2 Adj declension | decl. 1 & 2, irregular, indeclinable | A-M | Saldo-Morphology |
| 3 Morphology 1 | word analysis for morphemes:<br>*oändlig*: prefix:*o-*; root:*-änd-*; suffix:*-lig* | M[3] | |
| 4 Morphology 2 | word-building: root, compound, derivation, suppletion, lexicalized, MWE[1] | M | |
| 5 MWE type | taxonomy under development | M | |
| 6 Nom declension | decl. 1-6, extra | A[4] | Saldo-Morphology |
| 7 Nom gender | common, neuter, both, N/A | A | Saldo-Morphology |
| 8 Nom type | abstract–concrete, (un)countable, (non)collective, (in)animate, proper name, unit of measurement | M | |
| 9 Register | neutral, formal, informal, sensitive | M | |
| 10 Synonyms | free input, same word class | A-M | Swesaurus |
| 11 Topics/domains | general + 40 CEFR-related topics[5] | A-M | Lexin, COCTAILL |
| 12 Transitivity | (in-, di-)transitive, N/A | A-M | SAOL (under negotiation) |
| 13 Verb category | lexical, modal, auxiliary, copula, reciprocal, deponent | M | |
| 14 Verb conjugation | conjugations 1-4, irregular, N/A | A | Saldo-Morphology |
| 15 Verb action type | motion, state, punctual, process[6] | M | |

Table 1: Linguistic aspects added to SenSVALex and SenSweLLex items
[1]MWE = Multi-Word Entity; [2]Manual based on automatically enriched input; [3]Manual; [4]Automatic; [5]Topics come from the CEFR document (Council of Europe, 2001), COCTAILL corpus (Volodina et al., 2014), and some other resources; [6]Incl. limited and unlimited process verbs

and argumentation based on SAG and other work on the Swedish language and linguistic description in general.

To augment SenSVALex & SenSweLLex, we use different resources. Besides the information already present in these lists (word senses, Saldo descriptors, automatically derived CEFR level, part-of-speech), we use Saldo / Saldo morphology (Borin et al., 2013), Swesaurus (Borin and Forsberg, 2014), Lexin (Hult et al., 2010) and potentially SAOL (Malmgren, 2014) to enrich the lists.

Saldo morphology is used to add nominal gender, nominal declension and verbal conjugation. Adjectival declension and adjectival (and adverbial) structure are derived from the comparative and superlative forms given in Saldo morphology and checked manually. Synonyms are added using Swesaurus. Other named resources are planned for enriching topics and transitivity patterns. The remaining categories are left to be manually annotated.

### 3.3 Tool functionality

LEGATO offers a range of useful functionalities. It allows moving forward as well as backwards through the list; to search through the list of word senses to be annotated and to filter by certain criteria; to skip words you are uncertain about. Items that are skipped are added to a dedicated ‘skip list’ which makes it is easy to come back to these items. It also keeps track of your progress, allowing the annotator to close the interface, come back at a later time and continue where they left. Finally, it includes (automatically generated) links to different external resources such as Saldo (through Karp), Wiktionary, svenska.se, Lexin, synonymer.se, Korp and SAG.

For user friendliness, we keep guidelines, issue-reporting and lookup/reference materials linked to the front page of the tool. It is possible to leave comments, start issues/discussion threads, as well as see an overview of all completed tasks and tasks that are remaining.

### 3.4 Piloting the tool

To test LEGATO’s functionality as well as to control that the automatic linking of items is sufficiently reliable, we carried out an experiment with 100 SenSVALex items, divided equally between nouns, verbs, adjectives and adverbs. The selected words represent all the CEFR levels available in the COCTAILL corpus, various morphological paradigms and other types of linguistically relevant patterns as shown in Table 1.

In order to test the tool, two of the authors volunteered as annotators. After gathering data from the intial test phase, we calculated inter-annotator agreement (IAA) between the automatic analysis and annotator one (IAA 1), as well as the inter-annotator agreement between annotator one and annotator two (IAA 2). Table 2 shows Cohen’s $\kappa$[4] for the various categories. For IAA 1, only categories where annotator one had completed all tasks, and where automatic enrichment was used, were taken into account. For IAA 2, only categories where both of the annotators had completed all tasks were taken into account. This explains why some of the values are missing in the Table.

As can be gathered from Table 2, categories with closed answers, e.g. only one possible answer value, lead to higher agreement (nominal declension, nominal gender, verbal conjugation), while categories that allow multiple answers or free-text input show less agreement (nominal type, adjectival adverbial structure, morphology 1). For example, for nominal type, if one annotator selects “abstract, countable, inanimate" and another annotator select “concrete, countable, inanimate", this would be counted as disagreement. In order to address such problems, one would have to calculate partial agreement. One notable exception is adjectival declension, which only allows one value, but has low agreement between the automatic analysis and annotator one. This discrepancy could stem from the fact that all forms in Saldo morphology are automatically expanded, according to regular morphology, thus potentially producing forms that are incorrect.

| Category | IAA 1 | IAA 2 |
|---|---|---|
| nominal declension (6) | 0.85 | 0.80 |
| nominal gender (7) | 0.82 | 0.73 |
| nominal type (5) | | 0.20 |
| verbal conjugation (14) | 0.82 | 0.94 |
| adjectival declension (2) | 0.49 | |
| adjectival adverbial structure (1) | 0.39 | |
| morphology 1 (3) | | 0.48 |
| Overall $\kappa$ | 0.73 | 0.60 |

Table 2: Inter-annotator agreement. Numbers in brackets (Column 1) refer to the numbering of categories in Table 1

As a result of the IAA calculations, a subset of categories has been deemed reliable enough to be added automatically (categories 6, 7, 14 in Table 1), and another subset will be offered in a semi-automatic way, where a manual control check will be performed (categories 1, 2, 10, 11, 12 in Table 1).

The experiment with the 100 items has also helped us set up and refine guidelines for more extensive annotation by project assistants, as well as improve the functionality of the tool.

### 3.5 Technical details

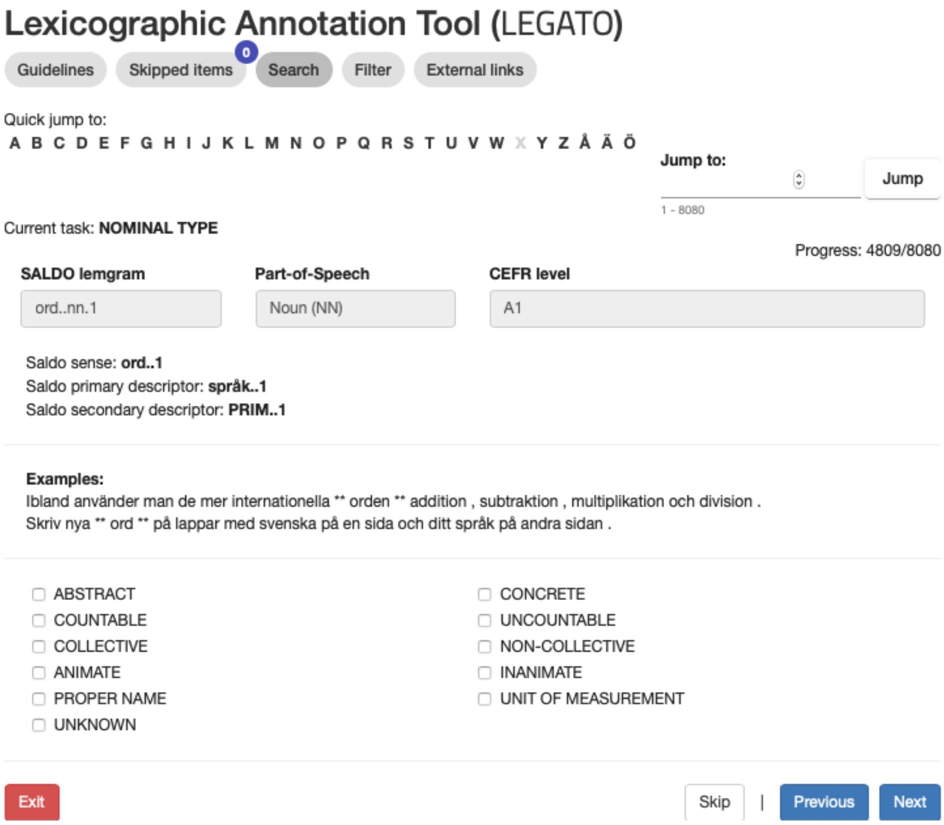
LEGATO is a module integrated with the Lärka-Labb[5] platform. Like its parent platform, the LEGATO front-end is written in TypeScript and HTML using the Angular (previously called *Angular 2*) framework[6]. The back-end is written in Python 2. Data is stored in MySQL format.

Data preparation (i.e. automatic enrichment, see Section 3.2) is done outside of the LEGATO platform using a set of dedicated scripts. In a multi-

[4] While values between 0.40 and 0.60 are generally considered borderline, values of 0.75 and above are seen as good to excellent.

[5] `https://spraakbanken.gu.se/larkalabb`

[6] `https://angular.io`



Current task: **NOMINAL TYPE**

Progress: 4809/8080

| **SALDO lemgram** | **Part-of-Speech** | **CEFR level** |
|---|---|---|
| ord..nn.1 | Noun (NN) | A1 |

Saldo sense: **ord..1**
Saldo primary descriptor: **språk..1**
Saldo secondary descriptor: **PRIM..1**

**Examples:**
Ibland använder man de mer internationella ** orden ** addition , subtraktion , multiplikation och division .
Skriv nya ** ord ** på lappar med svenska på en sida och ditt språk på andra sidan .

☐ ABSTRACT
☐ COUNTABLE
☐ COLLECTIVE
☐ ANIMATE
☐ PROPER NAME
☐ UNKNOWN
☐ CONCRETE
☐ UNCOUNTABLE
☐ NON-COLLECTIVE
☐ INANIMATE
☐ UNIT OF MEASUREMENT

Figure 1: LEGATO graphical user interface

step process, these scripts (1) create the sense-based word list, (2) add Saldo primary and secondary descriptors, (3) add further information such as synonyms and nominal gender by linking lexical resources based on lemgram, sense and part-of-speech tuples and (4) add example sentences. The resulting data is played into the databases on the server side to reduce the number of API calls and reduce runtime. As some of these scripts have a rather long runtime (the average time per entry for example selection is 0.66 seconds on an Intel Core i5-5200U processor, resulting in about 3 hours total for the whole list), they are not distributed as an integrated part of LEGATO and we do not consider advisable to integrate them into the LEGATO platform. However, the code for running interlinking can be made available for reuse.

## 4 Concluding remarks

We are currently exploring a possibility of using Lexin (Hult et al., 2010) and COCTAILL (Volodina et al., 2014) to automatically derive topical domains for vocabulary items. Furthermore, fruitful negotiations are ongoing on a potential access to parts of the SAOL database (Malmgren, 2014) for semi-automatic support of annotation of transitivity patterns.

A full-scale annotation of the two lists is planned for the near future, with the results (i.e. a full resource) expected by the end of 2019. Once the resources are richly annotated, we expect to perform both quantitative and qualitative analysis of L2 lexical competence. The LEGATO tool will have a thorough testing during that time and we hope this will lead to further improvements of the tool.

Since Legato is a module in a highly intricate and interlinked system Lärka, we do not deem it reasonable to release the code for this module only. However, in the future, we would like to make the platform available to other users by allowing them to upload their own data and define what they want to annotate.

## 5 Acknowledgements

This work has been supported by a grant from the Swedish Riksbankens Jubileumsfond (Development of lexical and grammatical competences in immigrant Swedish, project P17-0716:1).

## References

Malin Ahlberg, Lars Borin, Markus Forsberg, Olof Olsson, Anne Schumacher, and Jonatan Uppström. 2016. Språkbanken's open lexical infrastructure. *SLTC 2016*.

David Alfter, Yuri Bizzoni, Anders Agebjórn, Elena Volodina, and Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 1–7. Linköping University Electronic Press.

David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*.

Laurie Bauer and Paul Nation. 1993. Word families. *International journal of Lexicography*, 6(4):253–279.

Lars Borin and Markus Forsberg. 2014. Swesaurus; or, The Frankenstein approach to Wordnet construction. In *Proceedings of the Seventh Global Wordnet Conference*, pages 215–223.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012. The open lexical infrastructure of Språkbanken. In *LREC*, pages 3598–3602.

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

Annette Capel. 2012. Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, 3.

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

Luise Dürlich and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *LREC*.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN 2014*, pages 91–102.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Marcella Hu Hsueh-Chao and Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a foreign language*, 13(1):403–30.

Ann-Kristin Hult, Sven-Göran Malmgren, and Emma Sköldberg. 2010. Lexin-a report from a recycling lexicographic project in the North. In *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010)*.

Batia Laufer and Geke C Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language*, 22(1):15–30.

Batia Laufer and Donald D Sim. 1985. Measuring and explaining the reading threshold needed for english for academic purposes texts. *Foreign language annals*, 18(5):405–411.

Sven-Göran Malmgren. 2014. Svenska akademiens ordlista genom 140 år: mot fjortonde upplagan. *LexicoNordica*, (21).

James Milton. 2013. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In Camilla Bardel, Lindqvist Christina, and Batia Laufer, editors, *L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis*, pages 57–78. EuroSLA monograph series 2.

I Nation. 2006. How large a vocabulary is needed for reading and listening? *Canadian modern language review*, 63(1):59–82.

Paul Nation and Paul Meara. 2010. Vocabulary. *An introduction to applied linguistics*, pages 34–52.

Erik M Petzell. 2017. Svenska akademiens ordbok på nätet. *LexicoNordica*, (24).

Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58.

David Singleton. 1995. Introduction: A critical look at the critical period hypothesis in second language acquisition research. *The age factor in second language acquisition*, pages 1–29.

Emma Sköldberg, Louise Holmer, Elena Volodina, and Ildikó Pilán. 2019. State-of-the-art on monolingual lexicography for Sweden. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 7(1):13–24.

Ulf Teleman, Staffan Hellberg, and Erik Andersson. 1999. *Svenska akademiens grammatik*. Svenska akademien.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173. Association for Computational Linguistics.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 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 at SLTC 2014, Uppsala University*, 107. Linköping University Electronic Press.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. Swell on the rise: Swedish learner language corpus for European reference level studies. *LREC 2016*.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 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 at SLTC, Umeå, 16th November 2016*, 130, pages 76–84. Linköping University Electronic Press.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.