11 A practical guide to the Swedish L2 lexical profile

Therese Lindström Tiedemann
University of Helsinki

David Alfter
University of Gothenburg

Elena Volodina
University of Gothenburg

Vocabulary is a fundamental aspect of any language since without words you cannot communicate, nor learn other aspects of a language, such as grammar or pronunciation. The Swedish L2 profile offers many ways in which researchers can explore the vocabulary which learners can produce and are expected to understand at different proficiency levels. It also provides a foundation for innovative ways of teaching Swedish, for instance, through Computer Assisted Language Learning (CALL) and Data Driven Learning (DDL).

In this chapter we show how the lexical part of SweL2P can be used to explore the vocabulary growth of language learners both receptively and productively in a step-by-step overview. Starting from a bird's eye view of vocabulary in course books and learner essays we show how to zoom in on some specific aspects of vocabulary, choosing adjectives as an example. We use SweL2P to show how adjectives occur in course books and how they appear in learners' texts – comparing the lexis in both, but also showing the potential to explore the way learners acquire vocabulary more broadly. Finally, we present how results in SweL2P can be easily compared to other Swedish corpora.

1 *Introduction*

Vocabulary is a fundamental aspect of any language since without words you cannot communicate, nor can you learn other aspects of a language, such as grammar or pronunciation, without words to practise. The Swedish Second Language (L2) profile (SweL2P; Volodina, Alfter & Tiedemann 2024) is an open resource for Swedish as a second language which presents information about the occurrence of lexis, grammar and morphology in Swedish L2 course books and learner essays side by side in relation to the levels of the

HUM INFRA Therese Lindström Tiedemann, David Alfter & Elena Volodina. 2025. A practical guide to the Swedish L2 lexical profile. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 355–403. University of Tartu Library. DOI: 10.58009/aere-perennius0180

Common European Framework of Reference for languages (CEFR; Council of Europe 2020). It offers many ways in which researchers from different disciplines, but also language teachers, assessors and material designers, can explore what learners can produce and what they are expected to understand at different proficiency levels. SweL2P provides a new perspective that can have a substantial impact on language teaching and language learning, as well as on linguistic research.

In this chapter we show how the lexical part of SweL2P can be used to explore the vocabulary growth of language learners both receptively and productively in a parallel fashion which facilitates comparisons, something which according to usage-based perspectives can be important in relation to language learning (cf. Davidson et al. 2008). After a bird's eye view of vocabulary in course books (receptive) and learner essays (productive) we show how to zoom in on specific aspects of vocabulary in two case studies, to explore adjectives further for research purposes, or to support the teaching of Swedish as a second language. We also show how links in the resource lead the user to the empirical data in Korp,[1] Språkbanken Text's corpus management and search platform (Borin et al. 2025). There the user can explore the corpus data more closely and make comparisons to the results in other Swedish corpora, e.g. possible input such as newspapers.

The lexical SweL2P provides new opportunities for empirical research, but also new ways to digitally enhance usage-based language learning and teaching, e.g. through Data driven learning (DDL; see e.g., Gilquin & Granger 2010, Warren 2016, Boulton & Vyatkina 2021), while also showing a promising potential for future use in NLP applications such as Computer assisted language learning (CALL; see e.g., Mohsen et al. 2024) and automatic assessment.

The chapter starts with a section about the research context (Section 2). This is followed by a description of the SweL2P resource including the corpora used to create SweL2P (Section 3). In Section 4 we focus on the lexical profile, after which Section 5 explores some research and teaching scenarios briefly. We then continue with two case studies with guidelines on how to use SweL2P. First, we explore adjectives by frequencies in relation to linguistic research (Section 6), and then we present some ideas on how the subprofiles for adjectival declension, comparison and MWE can be useful in teaching (Section 7). After this we introduce some downstream NLP tasks for which the lexical profile can be used and we include links to downloadable datasets associated with the lexical profile (Section 8), before we end with some final remarks and future prospects (Section 9).

1 https://spraakbanken.gu.se/korp/

2 *Research context*

Second language acquisition (SLA) can be studied in various ways. Since the late 1980s it has become more and more common to try to study second language acquisition through corpora containing learners' own production. This type of research is nowadays often referred to as Learner corpus research (LCR). When learner corpora were introduced it was already evident that first language (L1) corpora could be of much use in relation to second language teaching, e.g. in designing teaching materials. However, as Granger (2002: 21) points out L1 corpora can only "provide valuable information on the frequency and use of words, phrases and structures" they cannot say anything about "the difficulty they [=the words] present for learners in general or for a specific category of learners". Specialists on vocabulary teaching have been claimed to agree that "both frequency and difficulty have to be taken into account" in language teaching (Granger 2002: 22) and to some extent it has been shown that frequency in corpora correlates to the likelihood that learners will use words (Davidson et al. 2008: 138). Learner corpora can provide empirical evidence of what learners might find difficult by studying what they do not use, and hence might not have learnt, or might be avoiding, as well as whether they use vocabulary items and vocabulary types in a manner somehow different to L1 speakers of the language.

In this chapter we first summarise some of the research done in connection to usage-based approaches and data driven learning (DDL) (Section 2.1) and vocabulary profiling (Section 2.2). In section 2.3 we introduce Contrastive interlanguage analysis (CIA) which has been used quite frequently in learner corpus research to compare L1 and L2, but also to compare the language of learners with different L1 backgrounds. Finally, section 2.4 presents some previous research on Swedish adjectives as background to the case studies in section 6.

2.1 *Usage-based approaches and data driven learning (DDL)*

This type of acquisition research fits particularly well with usage-based theories of language acquisition (e.g. Davidson et al. 2008, Wulff 2020, Römer 2023). In usage-based linguistics a prominent idea is that language learning is affected by language input as well as by learner output and general cognitive abilities. The type of language which we encounter as input and use ourselves is seen as likely to affect what we learn and when we learn things. This makes it highly relevant to be able to make comparisons between learner language and first language usage, but also between different proficiency levels.

In order to test the hypothesis that usage affects learning, it is important to have comparative data for different language usage. This is one of the reasons why we believe it is important to be able to compare course-book usage at different CEFR-levels to learners' production at different levels. In order to do this as well as possible we would need access to information about the course books which the learners who wrote the essays in the learner corpus used and preferably also what else they read and what they listened to, etc. This kind of information is not easy to come by and, instead, the SweL2P project chose to use a course-book corpus (COCTAILL, Volodina, Pilán, Eide et al. 2014) consisting of several course books per CEFR level, aimed at adult learners of Swedish, as a way of ascertaining what learners are expected to be able to understand (receptive knowledge). At the same time, this corpus also gives important insights into the input which learners normally get. To also be able to gauge the productive abilities of learners at different proficiency levels, we used a learner corpus for Swedish as a second language, (SweLL-pilot, Volodina, Pilán, Enström et al. 2016a, Volodina 2024). Both corpora are introduced further in Section 3.1.

2.2 *Vocabulary profiling*

Profiling has been described as the "identification and definition of developmental stages" (Granfeldt & Ågren 2014) and was initially used in relation to first language development and language disorders (Crystal et al. 1976) but it has also become asssociated with second language research (see e.g. Clahsen 1985, Pienemann & Mackey 1993, Granfeldt et al. 2005, Keßler & Liebner 2011, Hawkins & Filipović 2012, Granfeldt & Ågren 2014).

For Swedish important work has also been done on a typologically-based lexical profile based on translation corpora, in the work of Viberg (1990, 1992, 2006b,a, 2013), with an aim to support L2 research. This perspective lists the most common words in a certain part of speech and/or in a particular semantic field by frequency and also studies how much of the vocabulary that is made up of the most frequent words in a particular group under study. Based on the idea that differences between languages can result in learning challenges, typological profiling can be of interest in relation to language acquisition. For instance, previous research has shown that Swedish nuclear verbs are "favored at early stages of second language acquisition" (Viberg 2006b: 127). Viberg defines this type of lexical profile as "an account of the distinctive character of its [=the language's] lexical structure in relation to other languages" (Viberg 2006b: 103).

2.3 *Contrastive interlanguage analysis (CIA and CIA²)*

Granger has shown how different corpora can be used to compare the production from different groups, e.g. learners with different first languages or learners and L1 speakers or expert users. This is a method which she has called Contrastive interlanguage analysis (CIA, Granger 1996, 2015), and in its revised form CIA². CIA and CIA² build on the ideas from contrastive analysis with the help of corpora. In CIA there is an emphasis on comparing like with like, e.g. only foreign language learners with other foreign language learners, not second language learners with foreign language learners, only texts of the same genre etc (e.g. Granger 1996: 44). To compare interlanguage with first language production, special L1 corpora were compiled in relation to the International Corpus of Learner English (Granger 1996: 44), something which we have not had a chance to do in the SweL2P project. Lately, it has also become common to compare advanced learners to "proficient" student writing "regardless of native speaker status" (Granger 2015: 12), as well as to make comparisons with large L1 reference corpora such as Corpus of Contemporary American English (COCA) (cf. Granger 2015).

Granger (2015) revised the method due to the criticism of CIA. The criticisms primarily questioned native speaker language as a target for learners. In the revised form, CIA², Granger emphasises variation both in the learner language and in the native/expert language. However, it is worth to bear in mind that Granger (2009) noted that "all the studies that compare learners of different proficiency levels are in fact based on an underlying L1 norm" (Granger 2015: 13–14). Tenfjord et al. (2006: 101–2) has also pointed out that comparison to a target language can be part of an objective method to study various phenomena in the learner language.

The data which we have gathered in SweL2P can be easily used to make comparative studies, e.g., between different proficiency levels, between learner language and course-book language, according to CIA². Furthermore, links to Korp (see Section 6.2) fascilitate comparison of the SweL2P data to corpora from different text types and varieties.

2.4 *Learning Swedish adjectives*

One aspect of language that clearly develops from early on in second language acquisition is the use of adjectives. Swedish adjectives are also both lexically and grammatically complex which is another reason why they can be interesting to observe over the different proficiency levels. Their numbers increase, they change semantically (cf. e.g. Axelsson 1994), as well as

Table 1: The most frequent adjectives in Löhndorf's (2021) study of L1 Swedish.

Grade 3	Grade 5	Grade 9	Grade 11/12
arg 'angry'	*bra* 'good'	*bra* 'good'	*stor* 'big'
rädd 'afraid'	*stor* 'big'	*rädd* 'afraid'	*bra* 'good'
glad 'happy'	*liten* 'small/little'	*hel* 'whole'	*många* 'many'
osäker 'insecure'	*rolig* 'amusing'	*stor* 'big'	*olika* 'different'
stor 'big'	*kul* 'fun'	*trött* 'tired'	*annan* 'other'
liten 'small/little'	*viktig* 'important'	*liten* 'small/little'	*ny* 'new'
jätterädd 'very afraid'	*fin* 'nice'	*ny* 'new'	*kränkt* 'insulted'
vanlig 'common/usual'	*gammal* 'old'	*mörk* 'dark'	*hel* 'whole'
nästa 'next'	*ny* 'new'	*fler* 'more'	*olika* 'different'
olika 'different'	*andra* 'other'	*många* 'many'	*fler* 'more'

morphologically. Morphologically, adjectives are interesting to study both in relation to word formation and to inflection.

The fact that Swedish adjectives are inflected both inside the noun phrase (Sw. *attribut*, for modifiers inside the noun phrase) and as predicatives (Sw. *predikativ*) on the clausal level, can prove challenging to learners. In addition, some adjectives are more common in a premodifying position inside the noun phrase whereas others are more common as predicatives.

In Allén's *Frequency Dictionary of Present-Day Swedish* (Sw. Nusvensk frekvensordbok) (Allén 1970, 1971) adjectives make up 6% of all tokens and 10% of all lemmas. Adjectives have also been shown to be much less common than nouns in the vocabulary of young L1 speakers (Löhndorf 2021: 79). Löhndorf also shows that children learning Swedish as a first language first tend to use adjectives to modify concrete nouns, only later increasing their use with more abstract meanings. She lists the ten most common adjectives in her materials from grades 3, 5, 9, 11/12 (see Table 1) which can be of interest for comparisons with L2 Swedish.

Young first language (L1) speakers have been shown to use adjectives mainly as predicatives at first (Löhndorf 2021: 82). Later they start to use adjectives more in premodifying position in noun phrases, and extended noun phrases are used more by proficient L1 speakers. In the L1 writing of upper-secondary students with good marks, adjectival pre-modifiers (Sw. attribut) make up a substantial part of all adjectives (cf. Hultman & Westman 1977). The amount of simple vs extended noun phrases have also been shown to correlate to whether texts represent typical spoken language or written language (cf. Collberg 2021) and to the information density in the text, lower in speech, higher in written texts especially in more formal genres.

Table 2: The most frequent adjectives in Axelsson's 1994 study of spoken L2 Swedish.

Adjectives	Proportion of all adjectives (%)
bra 'good'	10
svår 'difficult'	7
stor 'large'	4
bättre 'better'	3
svensk 'Swedish'	3
liten 'small'	3
lång 'long'	2
olika 'different'	3
inte bra 'not good'	2
ny 'new'	1
lätt 'easy'	2
halv 'half'	2
gammal 'old'	2
	45

Furthermore, according to Löhndorf (2021: 88), Ravid & Levie (2010) have shown that girls have higher adjective density, at least in Hebrew.

Axelsson (1994) investigated the order of acquisition of adjectives in Swedish as a second language in relation to semantic fields and the frequency of occurrence, but she also studied adjectival agreement. Her definition of adjectives relies primarily on "semantic meaning and syntactic function" (Axelsson 1994: 112) unlike the Swedish Academy Grammar (SAG, Teleman et al. 1999) which tends to put morphology first in word class classification. But like SAG, Axelsson (1994: 113) classifies adjectives used as adverbs (or rather adverbials) as adjectives. In addition, participles which are used as adjectives and certain pronouns are also classed as adjectives to be in line with Allén (1970, 1971), e.g. *olika* 'different', *själv* 'self'.

For her study, Axelsson collected learner data in the form of interviews, rather than written data as in SweL2P. The most common adjectives in her spoken data are listed in Table 2 and any adjectives that were less common were used by less than 20 individual learners (Axelsson 1994: 117). She emphasises that the adjectives used were affected by the topics discussed (Axelsson 1994: 118), but we can see that there is an interesting overlap with the data from L1 speakers in grade 3, 5, 9 and 11/12 in Löhndorf's study presented in Table 1.

If we compare these adjectives to the ones which Goddard & Karlsson (2008) have suggested should be included in a Swedish version of the Natural

semantic metalanguage (NSM) semantic primes, based on the work on prime words done by Anna Wierzbicka (e.g. Wierzbicka 1996) and Charles Goddard (e.g. Goddard 2012) only three are the same. The semantic primes include the adjectives *bra* 'good', *dålig* 'bad', *stor* 'big', *liten* 'small', *nära* 'near', *avlägsen* 'far'. In her master dissertation, Arle (2018) found that the concrete adjectives *dålig* 'bad', *stor* 'big', *liten* 'little' were more common in easy language versions of the novels she studied, than in the original texts, hence showing some similarity between easy language novels and learner language.

3 *The Swedish L2 profile*

The Swedish Lexical Profile is part of a bigger Swedish L2 Profile (SweL2P, Volodina, Alfter & Tiedemann 2024), which is described in an online hands-on manual by Volodina and Lindström Tiedemann.[2] The SweL2P was created as part of the research project *Development of lexical and grammatical competences in immigrant Swedish* and released in 2023. Below, we introduce the parent resource, SweL2P, and its user interface briefly (for more information, please see e.g. Volodina, Alfter & Tiedemann 2024) and then we focus on the lexical profile with all of its subparts in Section 4.[3]

3.1 *Swedish L2 profile – An overview*

The Swedish Second Language Profile (SweL2P) features the following profiles:

- a lexical profile, organised into subprofiles by words, multiword expressions, adjectival declensions and adjectival and adverbial structure;
- a grammatical profile, including noun patterns and verb patterns (Lindström Tiedemann et al., in preparation);
- a morphological profile, organised into word families and morpheme families (Volodina et al. 2021, Volodina, Ali Mohammed et al. 2022, Volodina, Mohammed et al. 2024)[4] (cf. Sköldberg et al. 2019).

SweL2P is empirical in nature since it is based on data from two corpora: COCTAILL – a corpus consisting of course books used for teaching Swedish

2 spraakbanken.github.io/L2_profiles/SwedishL2profiles, last updated in April 2023.

3 This section is based on Volodina, Alfter & Tiedemann (2024), which has been invited in an extended form as a chapter to this handbook

4 Additional information in the form of guidelines for the morphological analysis are available https://docs.google.com/document/d/1G5PEfeDEKg4dAZaupj6FmUUWBGiegiqagzXgTA3cDSY/edit?usp=sharing

to L2 students (Volodina, Pilán, Eide et al. 2014); and SweLL-pilot – a corpus of learner essays written by L2 learners of Swedish (Volodina, Pilán, Enström et al. 2016a, Volodina 2024). Texts in both corpora have been annotated with CEFR (Council of Europe 2020) levels by experts, starting from A1 (beginner) to C1 (advanced). There are also a few essays from the C2 level but no course books. COCTAILL has been used to get an approximation of the vocabulary and grammar L2 learners meet when reading, and therefore what they are expected to understand *receptively*. SweLL-pilot has been used to get an approximation of the vocabulary and grammar L2 learners are able to produce actively when writing, and therefore represents learners' *productive* writing abilities.

Within a usage-based approach one normally does not believe in clear boundaries between lexis and grammar and even though it may look as though we have clearly divided the SweL2P into lexis, grammar and morphology, this is not quite the case. The lexical profile naturally overlaps with the morphological profile which focuses primarily on word formation (e.g. derivation, compounding), whereas inflectional aspects are only rarely included since we have based the morphological profile on the lemma and base form. The only reason there are some inflected forms is that multi-word expressions (MWEs) may contain lexicalised inflected forms (e.g. *på allas läppar* 'on everyone's lips') and because a lemma-final morpheme which changes in other inflected forms in the morphological paradigm has been seen as an inflectional morpheme, (e.g. *flicka, flickor* 'girl, girls', where the *-a* in the singular form has been annotated as an inflectional ending).[5]

To further illustrate the fuzzy borders between lexis and grammar we also see that there are grammatical aspects under the lexical profile since there we can search by part of speech (POS), gender, nominal declension, and verbal conjugation.

3.2 *Graphical user interface*

The graphical user interface for browsing the SweL2P[6] has some features that are shared by all of its modules and subprofiles, such as a filter for CEFR levels, a possibility to enter a search item/word (except in the grammatical profile) and an option to see frequencies and samples from receptive and productive data for one's search. Some other features are specific for a specific (sub)profile.

5 For further information about the morphological annotation principles please see https://docs.google.com/document/d/1G5PEfeDEKg4dAZaupj6FmUUWBGiegiqagzXgTA3cDSY/edit?usp=sharing.

6 https://spraakbanken.gu.se/larkalabb/svlp

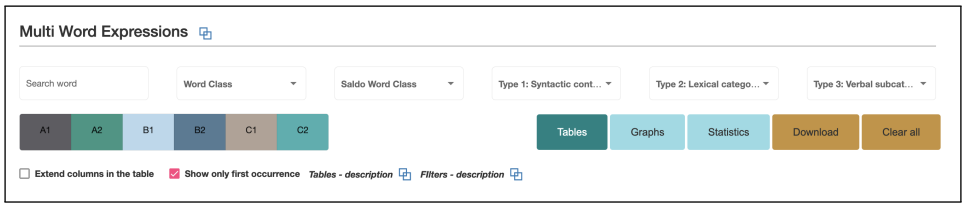


Figure 1: Filters for Multiword Expressions (Lexical profile, SweL2P)

CEFR level	Word	Lemgram	Sense	Word Class	Saldo Word Class	Receptive	Productive
A1	jeans	jeans..nn.1	jeans..1	Noun (NN)	Noun (nn)	0.77 (2)	0.00 (0)
A1	kläder	kläder..nn.1	kläder..1	Noun (NN)	Noun (nn)	3.09 (8)	7.32 (3)
A1	shorts	shorts..nn.1	shorts..1	Noun (NN)	Noun (nn)	1.16 (3)	0.00 (0)
A2	makaroner	makaroner..nn.1	makaroner..1	Noun (NN)	Noun (nn)	0.44 (3)	0.00 (0)

Figure 2: Table view for Sen*Lex (Lexical profile, SweL2P), filtered for 'always plural' nouns

Filters appear at the top of the page, providing a set of filters for each subprofile, e.g. MWE *Types 1-3* in Figure 1. The resource can be explored using several views: Table, Graphical and Statistical.

The Table view (Figure 2) lists all items which fit the filter(s) chosen. Columns contain descriptive information, among others, a clickable word (e.g. *jeans* in Figure 2) that opens a link to an entry with this item at `https://svenska.se/`, as well as a manual morphological analysis from the project and an earlier analysis from the Swedish Academy Glossary (SAOL, *Svenska Akademiens ordlista*) and the Contemporary Dictionary of the Swedish Academy (SO, *Svensk ordbok utgiven av Svenska Akademien*). In addition, this view also presents receptive and productive frequencies (both relative and absolute) based on the corpora mentioned above. If you click on the frequencies it opens a corpus search tool, Korp (Borin et al. 2025), containing the actual occurrences in the (Swedish) corpora we use. The receptive corpus (COCTAILL) is open to all users whereas the productive corpus (SweLL pilot) is restricted for ethical reasons.

The range of the columns depends on the profile – notably, for the grammatical profile, "pattern" is listed instead of "word" and here relative frequencies are currently given in terms of occurrences per 100 sentences rather than in relation to tokens.[7]

7 For more information e.g. about the choice of relative frequency, please see Lindström Tiedemann et al. (in preparation).

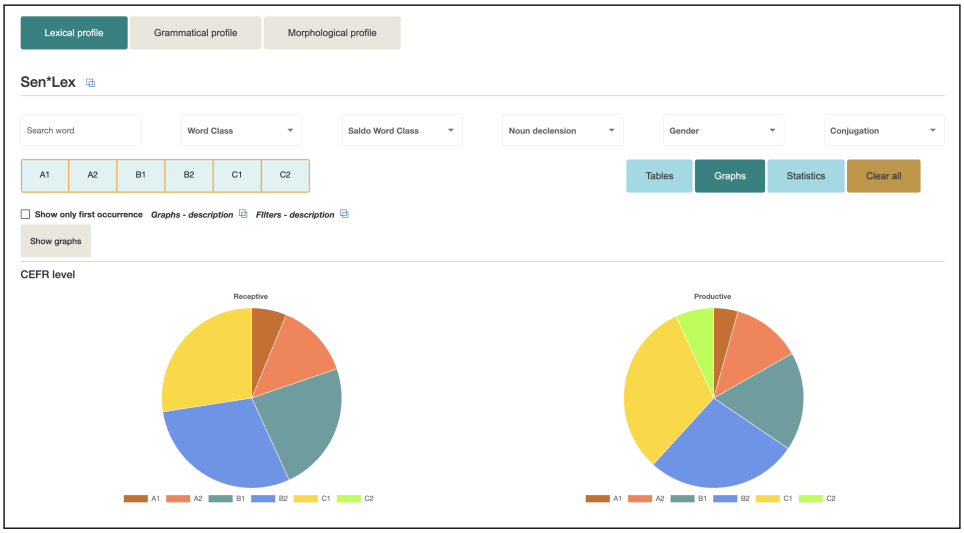


Figure 3: Graph view over Sen*Lex across CEFR levels (Lexical profile, SweL2P)

Graphical and Statistical views summarise the statistics and distribution of various features for the current selection in the two sources, receptive and productive. For example, Figure 3 shows a graph view of Sen*Lex (i.e. the whole word list in the lexical profile, see further Section 4.1) across CEFR levels, including all occurrences. The pie charts show you the proportion of lemmas which are represented at each CEFR-level.

In Figure 4, instead of graphs and distributions, we see counts in terms of types (cf. lemmas), tokens and type-token ratios (TTRs) per filter category so that we can study a statistical breakdown of each selection contrasting receptive and productive competences. In Figure 4 the number of types and tokens in Sen*Lex, in the lexical profile, are listed per CEFR-level (A1–B2 is visible in the figure) and across word classes (showing adjectives and adverbs in the figure).

The entire dataset or filtered data selection can be downloaded from the website as an excel file or as a csv file by clicking on the button to the right of the Statistics tab, "Download Statistics".

4 *The Swedish lexical L2 profile – resource description*

The SweL2P resource allows the user to choose different ways to approach the lexis. Under the heading Lexical profile, the following subprofiles can be accessed: adjectival declension, adjectival and adverbial structure, multi word expressions and Sen*Lex (cf. Figure 5).

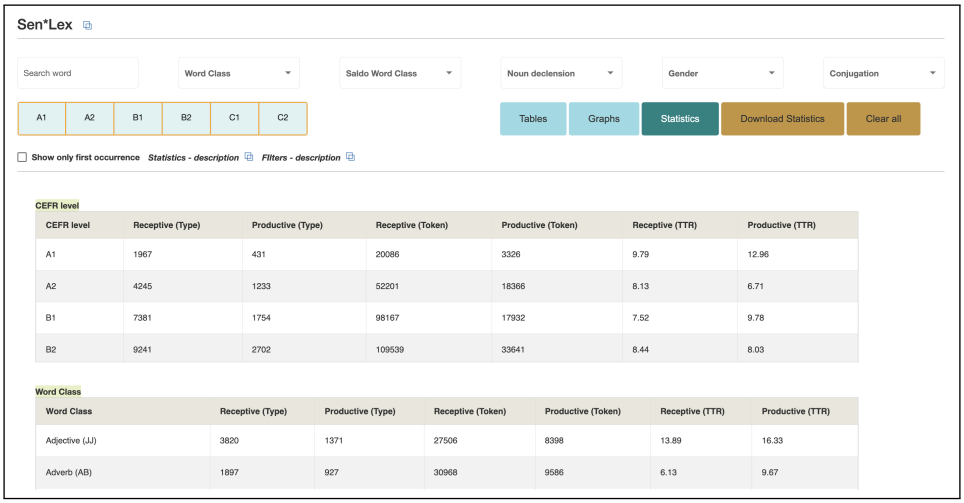


Figure 4: Statistical view showing entries per CEFR level A1–B2 and entries for two word classes (Lexical Profile, SweL2P)

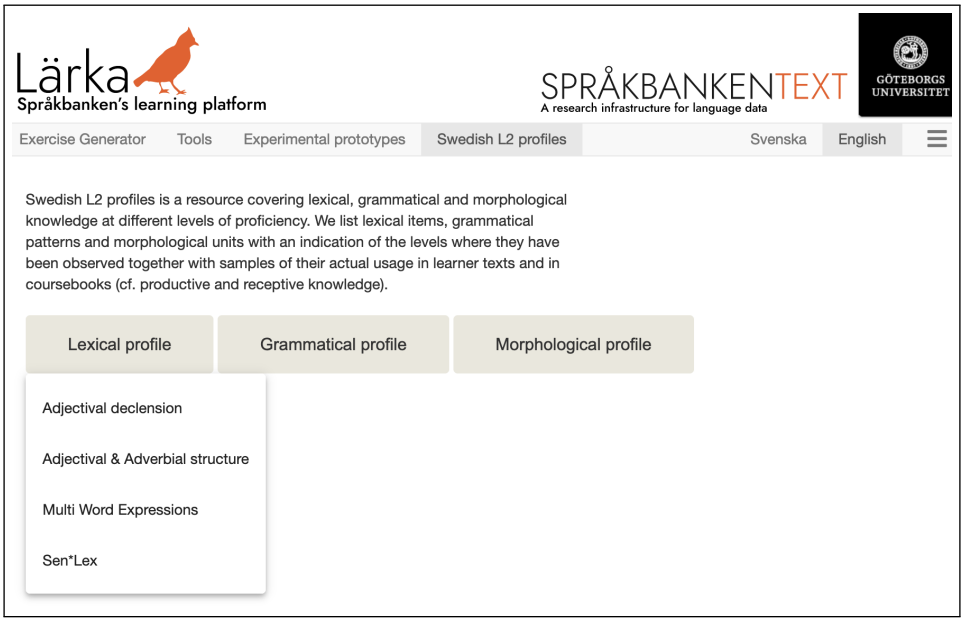


Figure 5: Overview of the possible selections under the Lexical profile in the SweL2P graphical user interface.

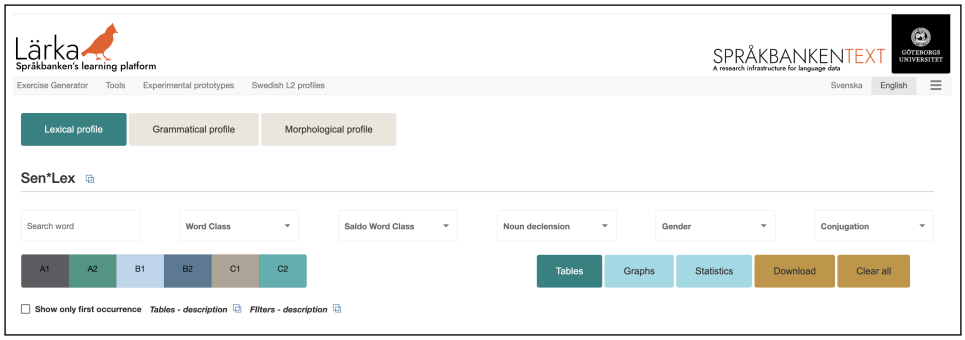


Figure 6: Buttons on the right for selection of the three different viewing options: Tables, Graphs or Statistics
(SweL2P)

The whole lexical profile is, in fact, based on Sen*Lex which is a sense-based word list which we introduce further in Section 4.1. However, rather than including everything as filters in the word list we provide some separate subprofiles for ease of access.

For each of the subprofiles of the Swedish Lexical L2 profile the user can choose various settings which will filter the results and provide different viewing alternatives. The views are presented as buttons on the right-hand side (see Figure 6) and the default setting is the table view, the alternatives being graphs and statistics. On this side the user also has the possibility of choosing to download the results, or to clear all the previous settings in that part of the profile.

Each part of the profile also offers a choice of showing only first occurrences. This is checked by default so that you only see instances according to their first level of occurrence. However, since for vocabulary this is currently based only on the registered first occurrence in course books it is often better to uncheck this.

4.1 *Sen*Lex – a CEFR-graded sense-based vocabulary list*

Traditional word frequency lists – and especially corpus-derived word lists – are typically based on form, grouping all inflected forms of a word under the same lemma, and making no distinction between different senses of a word. From a second language learning perspective, this is problematic for at least two different reasons: first, not all word forms are learned at the same proficiency level (Capel 2010, 2012, 2015). For example, learners typically first learn simple verb forms such as the Swedish present tense, before learning periphrastic forms such as the Swedish perfect or conditional

forms (Philipsson 2007, Yamaguchi et al. 2022). Second, polysemous words are typically subsumed under the same lemma – and by extension they share the same frequency in traditional wordlists. However, not all meanings of polysemous words are learned at the same proficiency level (Alfter, Cardon et al. 2022).

In order to address the polysemy problem, we have compiled new sense-based versions of SVALex (François et al. 2016) and SweLLex (Volodina, Pilán, Llozhi et al. 2016). The original corpora used to compile these lists were re-tagged with Sparv (Borin et al. 2016, Hammarstedt et al. 2022), including automatic word sense disambiguation. This allows us to separate different senses of words and assign them separate frequency values (Alfter 2021: 31–32). We call these sense-based lists SenSVALex and SenSweLLex, or Sen*Lex to refer to both lists simultaneously. In the online SweL2P resource, the two are available as a joint list and hence called Sen*Lex in the graphical user interface.

To prepare the data, we combined automatic processing (Borin et al. 2016, Nieto Piña 2019) and manual annotation (Volodina et al. 2021, Lindström Tiedemann et al. 2024a, Lindström Tiedemann et al., in preparation).

The manual annotation of Sen*Lex included various categories such as morphological analysis (Volodina et al. 2021[8], cf. Sköldberg et al. 2019), multi-word classification (Lindström Tiedemann et al. 2024a), and adjective and adverb type using the dedicated lexicographic annotation interface Legato (Alfter et al. 2019). In order to alleviate the human effort, the data was also linked to Lexin (Hult et al. 2010), Saldo-Morphology (Borin et al. 2013), and Swesaurus (Borin & Forsberg 2014) to (semi-)automatically enrich the data. Out of 15 categories, 8 were automatically enriched. In a pilot study where annotators checked 100 randomly sampled items, it was found that the automatic annotations for nominal declension type, nominal gender, and verb conjugation type were of sufficiently high quality that these categories were not manually checked in Legato. However, at a later stage further checks were carried out of all nouns and verbs in relation to these categories and some corrections were made before merging the information with Sen*Lex in the SweL2P. Some of the remaining automatically enriched categories were manually checked through Legato (adjective/adverb structure, adjectival declension) and others are still pending further manual annotation (synonyms, topics/domain, transitivity). Potential additional categories for manual annotation were also discussed but have not yet been included.

8 Additional information in the form of guidelines for the morphological analysis are available as https://docs.google.com/document/d/1G5PEfeDEKg4dAZaupj6FmUUWBGiegiqagzXgTA3cDSY/edit?usp=sharing

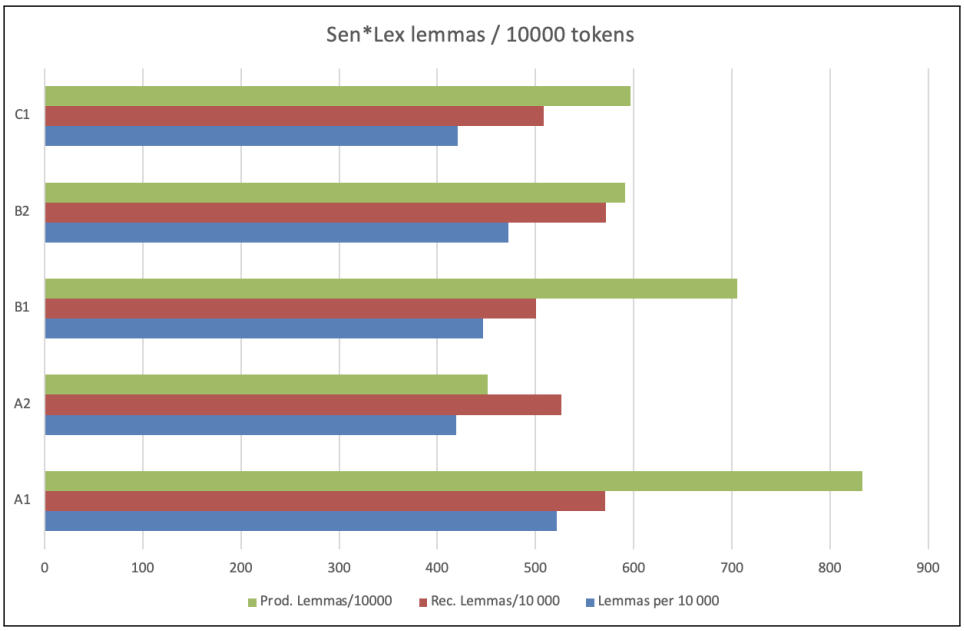


Figure 7: Comparison of the number of lemmas in relation to the total number of tokens in Sen*Lex presented as a total as well as separately for receptive and productive data.

4.2 *General overviews from the lexical profile in SweL2P*

Using the lexical SweL2P the user can choose to download all items (or first occurrences only) and use this data to see how many different lemmas there are per level and also how frequent they are in receptive (L2 Swedish course books based on the COCTAILL corpus) and productive (L2 Swedish learner essays based on the SweLL pilot corpus) data. As usual for such a comparison it is important to do this as relative frequencies, i.e. in relation to the data size of the respective data sets, if we want to compare the two datasets or different CEFR-levels. This is important since the data sizes at the different levels and in the two corpora as a whole are different.

Figure 7 shows that at the group level the number of lemmas in relation to the token frequency (cf. type/token ratio) is quite stable in the material as a whole, between 400–500+/ 10000 tokens at each CEFR level. Similarly, in the receptive data there are no big differences really between the different levels, the frequency is between 500–600 / 10 000 tokens. But as one would expect, the productive data shows clear differences between the levels. Interestingly, and at first somewhat surprisingly, there is a much higher number of lemmas per 10000 tokens at A1 (c. 800 / 10000 tokens). This means more lemmas

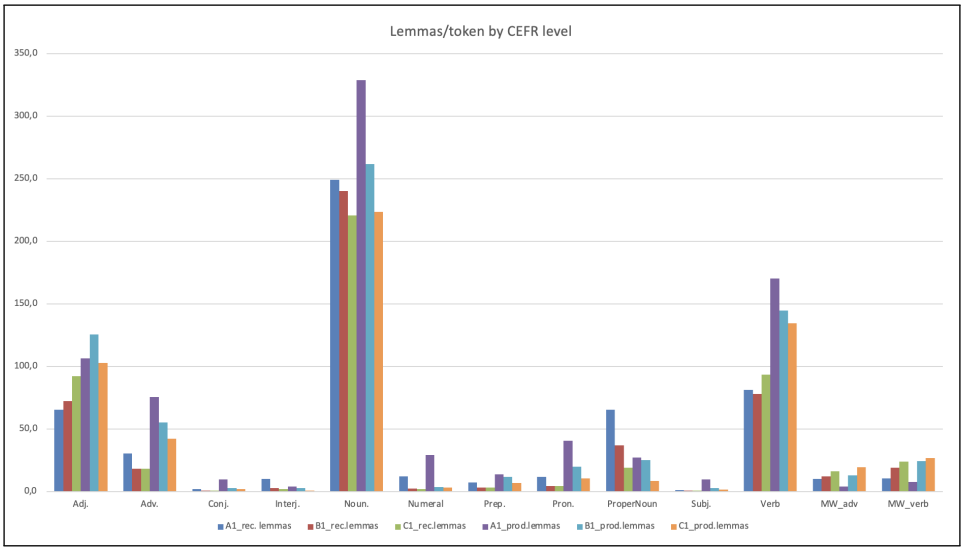


Figure 8: Lemmas per POS in relation to token frequency per CEFR level based on SweL2P

that are rarely used. The number of lemmas in relation to tokens is also quite high at B1-level (c. 700 / 10000 tokens). At A2-level it is only c. 450 and at B2 and C1 level nearly 600 and hence very similar to the receptive data. It could be that this is a result of the number of tasks which have been used to collect the data for the different levels, or it could be a result of words that are misused or misspelt, but this needs further analysis.

If we inspect the lemmas according to Saldo word class (SaldoPOS) (cf. Borin et al. 2013), we see that nouns are the most common word class, followed by verbs and adjectives (Figure 8). Nouns become less common in both receptive and productive texts at higher CEFR levels. Verbs remain fairly stable in receptive data but become less common at higher proficiency levels in productive data. Interestingly, we can also see that according to the relative frequency, learners use more verb lemmas, i.e. they vary their verb usage more at the group level, than the verb proportion used in the receptive texts. Adjectives, unlike nouns and verbs, show an increase in lemma proportion and hence vary more at higher CEFR levels in the receptive data (at the group level, i.e. based on all course books). However, in the productive data there is first an increase followed by a drop and A1 and C1 levels are found to be almost equal. However, notably, the proportion of adjectives is higher in productive texts than in receptive data which is something that should be explored further. It is quite possible that this is related to the topics of the texts.

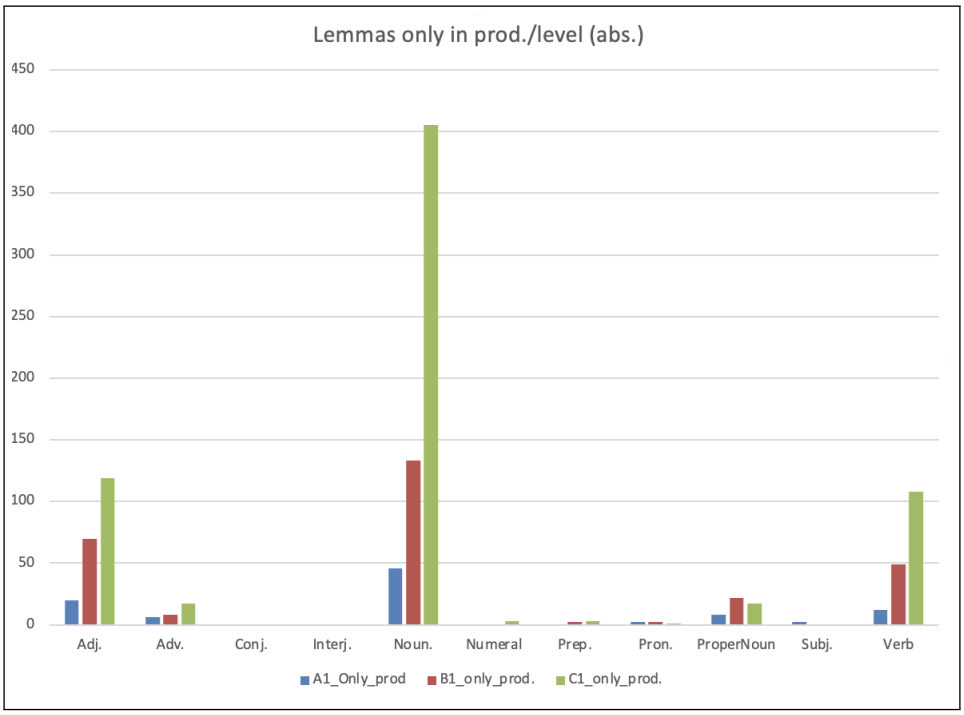


Figure 9: Lemmas exclusive to the productive data at that level based on SweL2P

Nouns, verbs, adjectives and adverbs all show higher lemma proportions in the learner production than in the receptive data at the levels A1, B1 and C1 (Figure 8), except nouns at C1-level which are used as much in receptive data as in productive data. A closer look at the word lists from the SweL2P show that this is clearly also shown in the number of lemmas which only occur in the productive data set at a certain level (Figure 9). This poses the question of how these lemmas were learnt. Were they learnt based on their frequency in course books at the level before this or from other texts? One way of exploring this issue further is by looking at whether vocabulary items appear to be core items or more peripheral at a certain level, in the sense of whether it is something that is common to many course books and/or learners (core) or rather something that is only covered by some and might not be necessary for everyone to know at that level (peripheral). This is something which we have explored in Volodina, Alfter & Tiedemann (2022).

5 *Research and teaching use scenarios*

The SweL2P resource is currently primarily useful for researchers and teachers of the Swedish language. Researchers can use the resource either:

1. to identify new research questions in a bottom-up manner by comparing receptive (course books) and productive (learner essays) occurrences across CEFR levels, or
2. to search for data based on a hypothesis, e.g. based on SLA theories or previous research on Swedish as a second language.

Teachers can similarly use the resource in at least two different ways:

1. to identify potentially challenging constructions for learners at a certain proficiency level, for instance by looking for structures which are rare in the course books and comparing this to what they have noticed that the learners find challenging. Based on this teachers have more information to use when they decide how to treat these structures in their teaching,
2. to plan data driven learning exercises where the teacher uses the resource to find concordance examples in corpora. Since the teacher can do this by CEFR-level there should be potential to find appropriate examples from course books and learner texts that could be adapted for classroom exercises (bearing in mind copyright as well as citation restrictions for learner essays), or the teacher can go one step further and look for similar examples in other corpora (cf. Section 6.2).

We now present a case study of how the lexical profile can be used in relation to linguistic research on adjectives (Section 6), followed by a case study in relation to language teaching where we look at adjectives in relation to their declension and adjectival MWEs (Section 7).

6 *Case study 1: linguistic research with the lexical SweL2P including hands-on guidelines*

Previous research (see Section 2.4) has shown that adjectives are used more at higher proficiency levels and that their semantic fields change (Axelsson 1994). We showed above (Section 4.2) that adjectives are the third most common POS in our data.

In this case study we look at the frequency of adjectives across CEFR levels and compare their receptive and productive frequencies based on

course-book data and learner essays, respectively. We start by looking at adjectives through Sen*Lex (Section 6.1) and compare the most common adjectives to previous studies. After this we explore how to access the actual occurrences in the corpora we have used, and how we can compare the results to other Swedish corpora (Section 6.2). In this way, we show how we can make a comparison similar to that in Axelsson (1994) but with access to actual corpus data for comparison rather than a static frequency list based on one genre, such as Allén (1970, 1971) which was based on newspapers. Furthermore, we also know that all of the data has been annotated in exactly the same way since the corpora are all annotated with the same pipeline (Sparv, Borin et al. 2016). This way, our method can potentially be used to compare learner language to different genres of first language writing and different potential conventions and/or norms (cf. Section 2.3).

In order to use the SweL2P you go to https://spraakbanken.gu.se/larka/svlp. The lexical profile is selected by clicking on the box Lexical profile and then selecting one of the alternatives from the drop-down menu: Adjectival declension, Adjectival-adverbial structure, Multi word expressions or Sen*Lex (see Figure 5). All of these are based on Sen*Lex (see Section 4.1 for more information), however, as described earlier, Sen*Lex has been augmented through manual annotation.

6.1 *Using Sen*Lex*

Sen*Lex (see Section 4.1) has a number of different filtering options which allow the user to get a bird's eye view of the lexis in course books and/or learner essays at different CEFR levels or to focus on a particular word class, gender, nominal declension, or verbal conjugation across all CEFR levels or to focus only on one or two (See Figure 10 and Table 3 for more details).

Sen*Lex can either be browsed directly online through the SweL2P resource, or the user can choose to download all items in the word list or all that have been filtered by clicking on Download on the righthand side (see Figure 10), which is usually better for researchers, in particular.

As an example of how Sen*Lex can be used we here use SweL2P to study the frequency of adjectives at different CEFR levels and compare these not only between the receptive and the productive side in the resource but also with previous research for both L2 and L1 Swedish (see Axelsson 1994, Löhndorf 2021, respectively). Zooming in further on adjectives we can compare the most frequent adjectives at each level in the receptive and productive data. We can do this either inside the resource or by downloading data from the resource as described above. Let us take a closer look at the most common adjectives in Sen*Lex for three levels (A1, B1, C1).

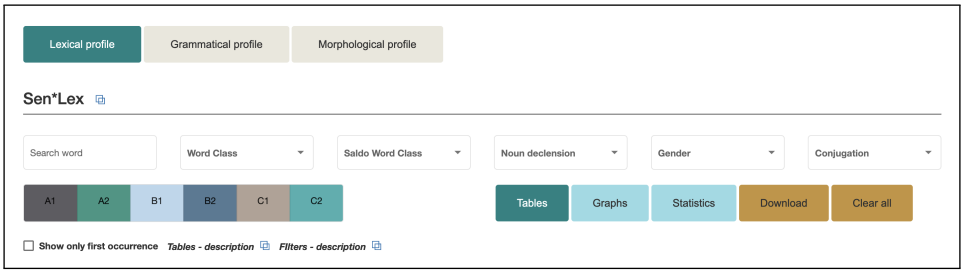


Figure 10: Filtering options in the Sen*Lex wordlist in the lexical profile, SweL2P.

Table 3: Filters and filtering options in Sen*Lex

Filters (Menus)	Nr categories	Filtering options
Word class	22	acc. to the SUC taxonomy, (Källgren et al. 2006: Table 12)
Saldo word class	37	acc. to the word classes used as part of lemgrams in Saldo, see https://spraakbanken.gu.se/resurser/saldo/taggmangd
Noun declension	6	*-or* (*blomma, blommor* 'flower, flowers'); *-ar* (*älv, älvar*, 'river, rivers'); *-er* (*dans, danser; bok, böcker* 'dance, dances; book, books'); *-r* (*sko, skor* 'shoe, shoes'); *-n* (*bo, bon* 'nest, nests'); - (*bär, bär* 'berry, berries')
Gender (nouns only)	6	always plural, no gender (*avgaser* 'exhaust'); neuter (*ett ägg, ett år* 'egg, year'); common (*en älv, en bil* 'river, car'); common or neuter (*bacon* 'bacon'); no gender (*gång på gång* 'time and again'); not applicable (*Afrika* 'Africa')
Conjugation (verbs only)	6	*-ar, -as* (*-ade(s)*); *-er, -es* (*-de(s), -te(s)*); *-r, -s* (*-dde(s)*); strong verbs (ablaut -); irregular (*dra – drog* 'tug – tugged'); no conjugation (refers to all non-verbal word classes)

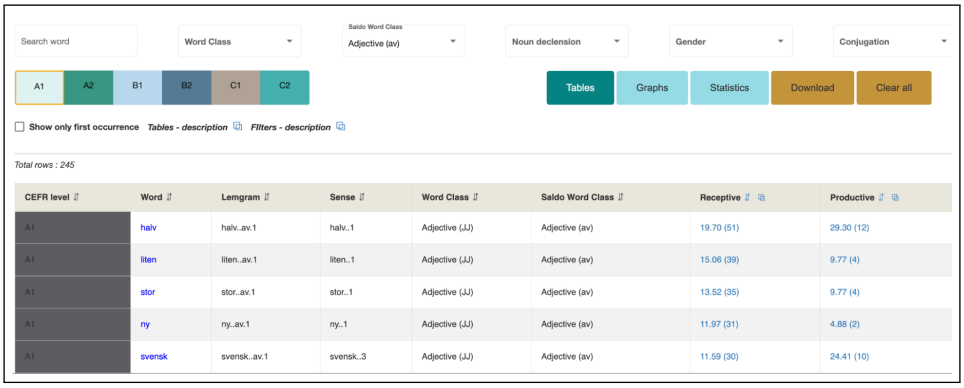


Figure 11: Most common adjectives in receptive data according to Saldo word class at A1-level, SweL2P

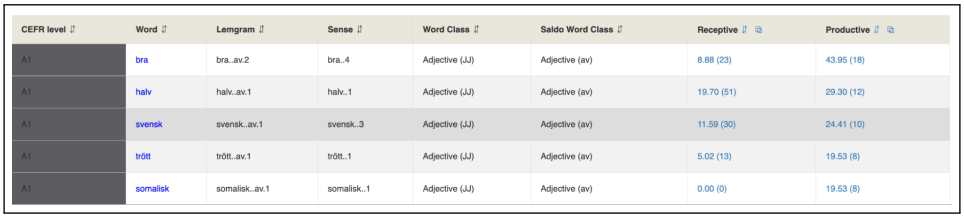


Figure 12: Most common adjectives in productive data according to Saldo word class at A1-level, SweL2P

Choosing the filters *Saldo word class: Adjective* and *CEFR-level: A1* we can then click on the arrows at the top of the columns marked receptive or productive to order the hits according to the least or the most common in either category (cf. the columns at the right in Figure 11). The most common adjectives at A1-level in the receptive data are *halv* 'half', *liten* 'little', *stor* 'big', *ny* 'new' and *svensk* 'Swedish'. In the productive data there are few overlaps among the most common adjectives, as there *bra* 'good', *halv* 'half', *svensk* 'Swedish', *trött* 'tired', *somalisk* 'Somali' are the most common adjectives (Figure 12). The differences here are clearly partly due to the limited topics which A1-level learners are asked to write about and hence do not necessarily give a very good idea of their lexical proficiency. It also makes it clear that adjectives connected with national, cultural or linguistic background are likely to be learnt quite early which also relates well to previous research (Axelsson 1994), and is in accordance with the A1 "can-do" statement (Council of Europe 2020).

For comparison, let us compare the most frequent adjectives at all CEFR levels in the productive data in SweL2P to the 13 adjectives listed as the most common by Axelsson (1994) based on her spoken learner data (Table 4).

Table 4: The most frequent adjectives in productive data in SweL2P as compared to Axelsson (1994).

	Axelsson (1994)	CEFR level in SweL2P				
		A1	A2	B1	B2	C1
1	*bra* 'good'	**bra** 'good'	**bra** 'good'	**olik** 'different'	*god* 'good'	**stor** 'large'
2	*svår* 'difficult'	**halv** 'half'	**svensk** 'Swedish'	*viktig* 'important'	*viktig* 'important'	*psykisk* 'psychological'
3	*stor* 'large'	**svensk** 'Swedish'	**olik** 'different'	**bra** 'good'	**stor** 'large'	*olik* 'different'
4	*bättre* 'better'	*somalisk* 'Somali'	*snäll* 'kind, nice'	**ny** 'new'	*andra* 'other'	*viktig* 'important'
5	*svensk* 'Swedish'	*trött* 'tired'	*glad* 'happy'	**stor** 'large'	**bra** 'good'	*andra* 'other'
6	*liten* 'small'	**gammal** 'old'	**gammal** 'old'	*god* 'good'	**olik** 'different'	*hög* 'high'
7	*lång* 'long'	**olik** 'different'	**ny** 'new'	*andra* 'other'	**ny** 'new'	**bra** 'good'
8	*olika* 'different'	**liten** 'small'	**liten** 'small'	*hel* 'whole'	*hel* 'whole'	*arbets-relaterad* 'work-related'
9	*inte bra* 'not good'	**stor** 'large'	*rolig* 'fun(ny)'	*glad* 'happy'	*stressad* 'stressed'	*stress-relaterad* 'stress-related'
10	*ny* 'new'	**svensk** 'Swedish'	**stor** 'large'	*ensam* 'alone'	*negativ* 'negative'	*stressad* 'stressed'
11	*lätt* 'easy'	*andra* 'other'	**halv** 'half'	**svår** 'difficult'	**svår** 'difficult'	**ny** 'new'
12	*halv* 'half'	*dålig* 'bad'	*fin* 'pretty'	*dålig* 'bad'	*arbetslös* 'unemployed'	*arbetslös* 'unemployed'
13	*gammal* 'old'	*engelsk* 'English'	*hel* 'whole'	**gammal** 'old'	*viss* 'certain'	*karolinsk* 'Carolingian'

Items that occur at two adjacent levels have been shaded. The items that are also on the Axelsson list have been bolded in Table 4. We can see that 7–8 of the items at A1 are also present in Axelsson even though her data is from spoken interviews. *Svensk* appears twice in the A1 frequency list but with different meanings according to the annotation.[9] Eight of the A2 items are also on Axelsson's list and to a large extent they are the same as on A1.

9 In the sense disambiguation, lemmas are linked to different senses in Saldo (Borin et al. 2013). Svensk 1 is the adjective linked to Sweden and hence nationality, while Svensk 3 is linked instead to Swedish as a language.

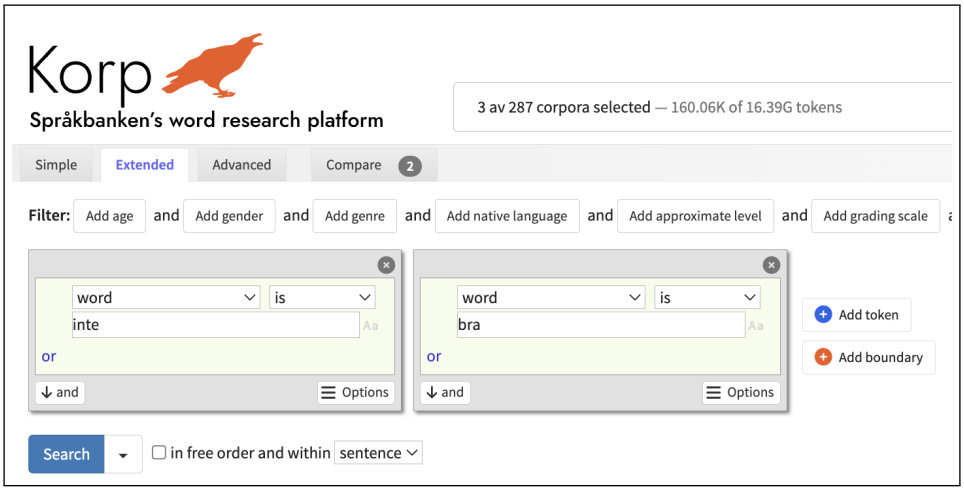


Figure 13: Querying the SweLL-pilot for occurrences of *inte bra* 'not good' in Korp

On B1 there are only 6 items which are the same as in Axelsson and this is where *svår* 'difficult' makes an entrance among the top-13 in the written learner essays. At B2 the similarity is 5 items, at C1 only 3 items. Clearly worth noting is also the fact that *dålig* 'bad' is relatively more common than in the earlier study, and as we saw in Section 2.4 *dålig* 'bad' is considered a semantic prime in Goddard & Karlsson (2008). It was also a word that was more common in easy language versions of novels than in the originals according to Arle (2018).

Instead of *dålig* 'bad' Axelsson (1994) included *inte bra* 'not good'. This has not been included in Sen*lex as an adjectival entry of its own, nor have any other combinations of the negation *inte* 'not' and an adjective. This is because it has not been seen as a MWE but rather as two independent single words. Nevertheless, in Korp we can query the SweLL pilot corpus (see Figure 13) for the collocation *inte bra*. This shows that the collocation occurs on all levels but it is much less common than *bra* is on its own, although, in comparison to *olika* 'different' and *ny* 'new', i.e. the items next to *inte bra* 'not good' on Axelsson's list, the frequency is similar to *olika* 'different' at A1, but at the other levels *olika* is much more common. *Ny* 'new' is less common at A1, similar at A2 and after that more common than *inte bra* 'not good'.

Out of the top five receptive and productive adjectives in SweL2P (see Figure 14, 15) three are the same and the two productive adjectives which are not among the most common receptive ones are clearly related to the type of tasks which forms the basis of the data: *psykisk* 'psychological' is a word that appears most in the tasks about stress which form an important

CEFR level	Word	Lemgram	Sense	Word Class	Saldo Word Class	Receptive	Productive
C1	stor	stor..av.1	stor..1	Adjective (JJ)	Adjective (av)	22.82 (359)	30.46 (160)
C1	andra	andra..av.1	andra..2	Adjective (JJ)	Adjective (av)	16.27 (256)	18.66 (98)
C1	ny	ny..av.1	ny..1	Adjective (JJ)	Adjective (av)	11.63 (183)	11.61 (61)
C1	olik	olik..av.1	olik..1	Adjective (JJ)	Adjective (av)	10.49 (165)	23.80 (125)
C1	liten	liten..av.1	liten..1	Adjective (JJ)	Adjective (av)	9.98 (157)	8.57 (45)

Figure 14: Most common adjectives in receptive data according to Saldo word class at C1-level, SweL2P

CEFR level	Word	Lemgram	Sense	Word Class	Saldo Word Class	Receptive	Productive
C1	stor	stor..av.1	stor..1	Adjective (JJ)	Adjective (av)	22.82 (359)	30.46 (160)
C1	psykisk	psykisk..av.1	psykisk..1	Adjective (JJ)	Adjective (av)	0.25 (4)	25.89 (136)
C1	olik	olik..av.1	olik..1	Adjective (JJ)	Adjective (av)	10.49 (165)	23.80 (125)
C1	viktig	viktig..av.1	viktig..1	Adjective (JJ)	Adjective (av)	8.33 (131)	21.70 (114)
C1	andra	andra..av.1	andra..2	Adjective (JJ)	Adjective (av)	16.27 (256)	18.66 (98)

Figure 15: Most common adjectives in productive data according to Saldo word class at C1-level, SweL2P

part of the data (cf. Caines & Buttery 2018). A task effect in relation to the task about stress is also visible in other words which show a high frequency such as *stressad* 'stressed', *stressrelaterad* 'stress-related' and also *arbetslös* 'unemployed' and *psykisk* 'psychological'. Similarly, the frequency of *viktig* 'important' is affected not by topic but by the fact that tasks at this proficiency level usually consist of argumentative essays.

6.2 *Browsing the data in Korp and comparing with reference corpora*

Axelsson (1994) used Allén (1970, 1971) for comparisons, a frequency word list (*Nusvensk frekvensordbok*, 'Frequency Dictionary of Present-Day Swedish') based on newspaper texts, since at the time that was the best resource for statistical comparisons. Nowadays, thanks to Språkbanken Text[10] and the Language Bank of Finland,[11] we have access to many different Swedish corpora which have been annotated in the same way and which can provide excellent statistical and qualitative data for comparison through Korp (Borin et al. 2012). This gives us better potential to empirically predict what learners can learn at the different stages of acquiring Swedish as Davidson et al. (2008: 138) have claimed that "a broad sample of many different use domains should be used" if we wish to try to predict learner usage.

10 https://spraakbanken.gu.se/
11 https://www.kielipankki.fi/

If we click on the number in either the receptive or the productive column in the SweL2P this will take us to the data in Korp, where we can see how the words have been used in the context of at least one sentence. The receptive data is open to everyone, whereas the productive data is restricted. The receptive data can only be accessed at the sentence level, whereas the productive data can also be accessed as full texts by clicking on a link under text attributes, to the right in Korp.

Let us look at the word *psykisk* 'psychological, mental' and see how it was used. In COCTAILL, *psykisk* appears in sentences such as (1a) and (1b) and it only appears in one course book.

(1) a. Jag har faktiskt också rätt – förutsatt att jag inte lider av en så allvarlig psykisk sjukdom att jag kvalificerat mig för inlåsning på en psykiatrisk klinik – att krossa mitt huvud mot marken.
'Actually, I also have the right – as long as I am not afflicted by a serious psychological disease which has qualified me to be locked up at a psychiatric clinic – to beat my head to pieces against the ground.' (Språkporten – svenska som andraspråk 1 2 3)

b. Barn behöver närvarande föräldrar – inte bara fysiskt utan också andligt och psykiskt.
'Children need parents who are present – not only physically but also spiritually and psychologically.'
(Språkporten – svenska som andraspråk 1 2 3)

All of the productive examples of *psykisk* at the C1 level are from learner essays on the topic of *Stress in current society*, e.g. (2a) and (2b).

(2) a. Det är inte bara vår fysiska hälsa som vi måste oroa sig för utan också psykiska.
'It is not only our physical health that we should worry about, but also our psychological [health].' (T100TT1)

b. Socialstyrelsen har i sin studie bevisat att stress på jobbet kan vara anledningen till psykisk sjukdom.
'The National Board of Health and Welfare have proven in their study that stress at work can be a reason for mental disease.'
(T95TT1)

Using the same query as the one that opens in Korp when we choose to click on the number in one of the columns, we can remove the CEFR-level (Figure 16) and then pick another corpus which we would like query

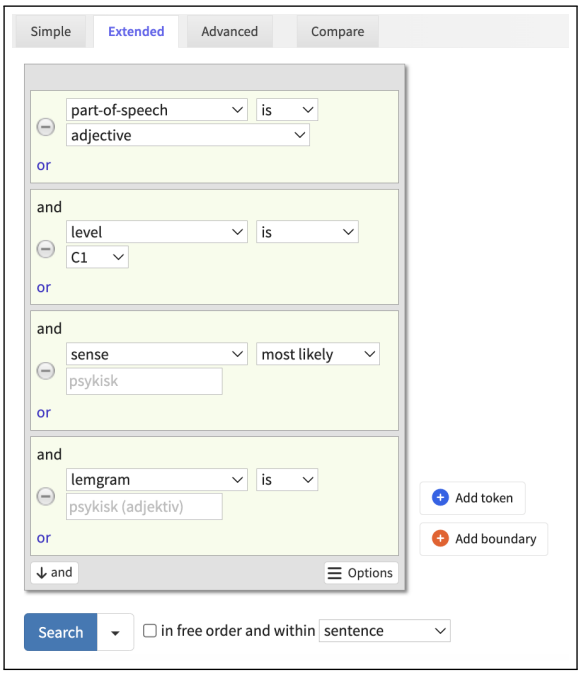


Figure 16: The adjustable query which opens in Korp.

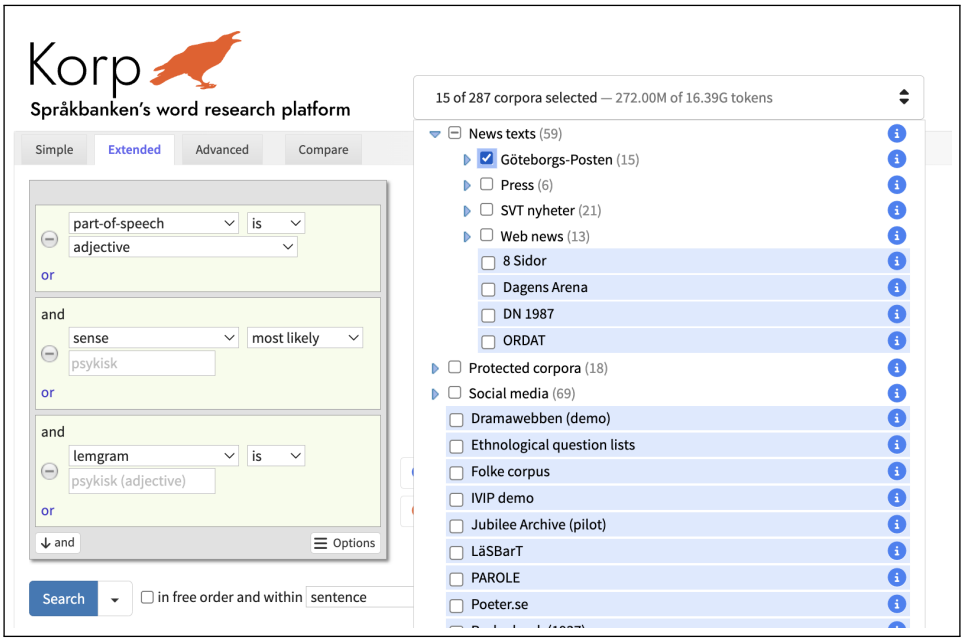


Figure 17: Choosing a new corpus to query in Korp

by using the drop-down corpus menu (Figure 17), e.g. news texts – Göteborgsposten.[12] The statistics from the search for *psykisk* as an adjective show a relative frequency 30.4 per 1 million token in *Göteborgsposten*, compared to 25.9 / 10 000 token, or 2589 / 1 million token, which is highest in student writings (C1 level). However, this word is nowhere near as common in course books, where the highest relative frequency is 0.26 / 10 000 tokens at B2 and 0.25 / 10 000 tokens at C1. Hence, learners who are using this are using it more than is common in either course books or newspapers. It seems unlikely that they have learnt the word only based on the course book, and the comparison with newspaper texts indicates possible overuse. This could possibly mean that they have copied the prompt, and it is hard to say how well they understand what the word means. To investigate this we would first of all need to have a closer look at the way learners use the word. This is something one can do if granted access to the restricted SweLL-pilot corpus.

7 *Case study 2: the lexical profile when teaching adjectives*

Teachers can of course also be interested in browsing the lexical frequencies as we showed in the previous section where we focused on research. We will now look at another aspect of the lexical profile which can interest both teachers and researchers, but this time focusing on a teacher perspective.

When teaching adjectival inflection, it can be of interest to know how often students are likely to read and also use adjectives from the different declensions, since according to usage-based theories this may affect how easily different forms can be learnt. In this section we therefore show how adjectives can be browsed by declension (Section 7.1), but also according to how they are used in comparisons (Section 7.2). Similarly, it can also be of interest to have more information in relation to the teaching of adjectival MWEs, and therefore we finish with an example of how to examine adjectival MWEs through the MWE profile (Section 7.3).

7.1 *The adjectival declension profile*

Teaching adjectives to L2 Swedish learners involves teaching them to inflect the adjectives correctly according to the gender and number of the noun, as well as whether the noun phrase is definite or indefinite. In addition, the inflection also depends on the syntactic position since the inflection is

12 Göteborgsposten, also known as GP, is a newspaper on the west coast of Sweden in the town of Gothenburg.

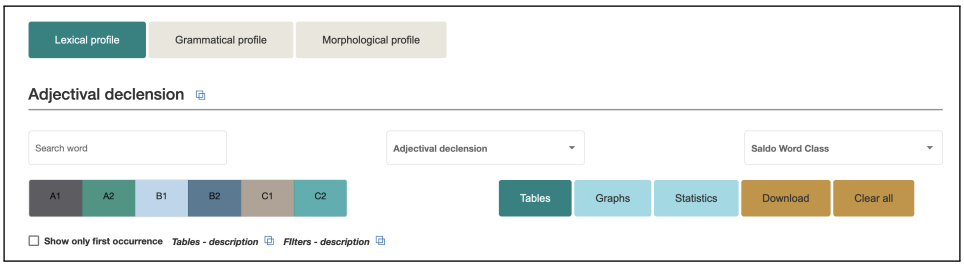


Figure 18: Adjectival declension and its filtering options, SweL2P

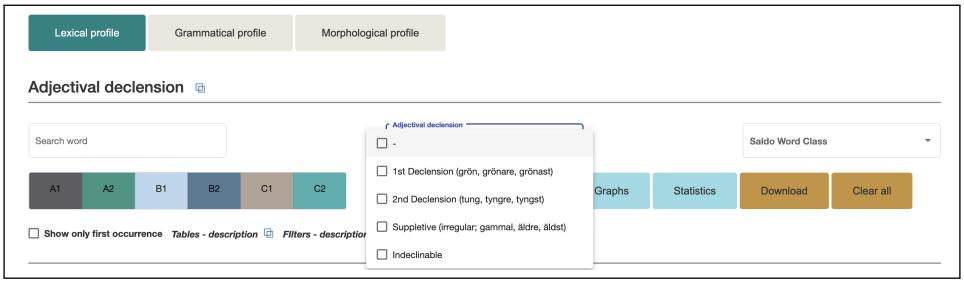


Figure 19: The adjectival declension filtering options in the subsection Adjectival declension, SweL2P

not quite the same when the adjective is part of a noun phrase (nominal modifier, Sw. *attribut*) as when it is an adjectival phrase on the clausal level (predicative, Sw. *predikativ*).

This part of the lexical profile is based on all adjectives in Sen*Lex (according to the part-of-speech tag). These adjectives have been manually annotated with information regarding declension in our annotation tool, Legato (Alfter et al. 2019).

The subprofile can be filtered according to CEFR-levels through the buttons on the left-hand side (see Figure 18), just as all other subprofiles in the lexical profile, cf. the MWE subprofile in Lindström Tiedemann et al. (2024a) and the morphological profile in Volodina, Alfter & Tiedemann (2024).

It is possible to choose only to view first occurrences or all items, where first occurrences are based on first occurrence in course books. The user can choose to search for a specific adjective, or filter based on adjectival declension (see Figure 19), or Saldo word class (see Figure 20). Saldo word class might seem counter-intuitive seeing as these are items which have been classed as adjectives. However, we have two different types of POS-annotation and they sometimes yield different results.

If, for instance, we select A1 and graphs (see Figure 21) we can see the proportion of adjectives of different declensions in both course books

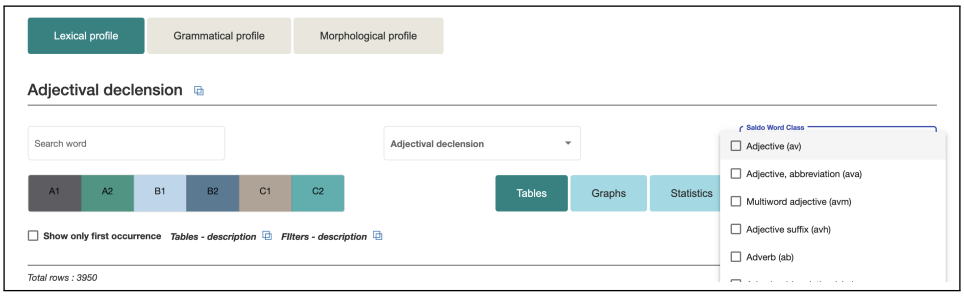


Figure 20: The Saldo word class filtering options in the subsection adjectival declension, SweL2P

Figure 21: Adjectival declension. Filter selected: A1. View selected: Graphs, showing all occurrences, SweL2P.

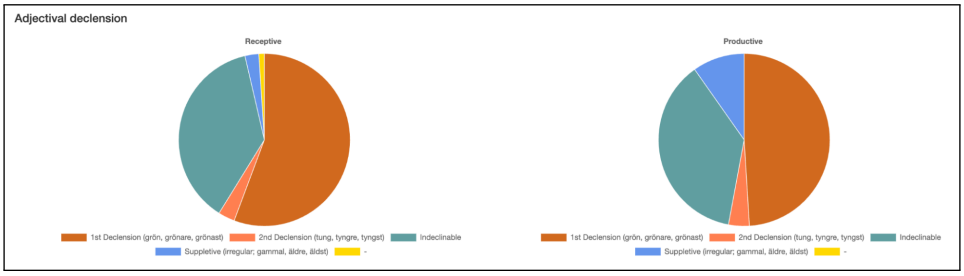


Figure 22: Adjectival declensions at A1-level, SweL2P

(receptive, left-hand side) and learner texts (productive, right-hand side) from the A1-level (see Figure 22).

This shows that the regular first declension *grön, grönare, grönast* 'green, greener, greenest' is clearly the most common in both receptive and productive texts, but quite closely followed by indeclinable adjectives (e.g. *svensk* 'Swedish', *alfabetisk* 'alphabetic'; see Figure 22).

With this much exposure to the first declension adjectives in course-book input and also the high level of usage in the productive texts, we might hypothesise that learners may be able to use the first declension quite well. However, the second most common group is nearly as common and consists

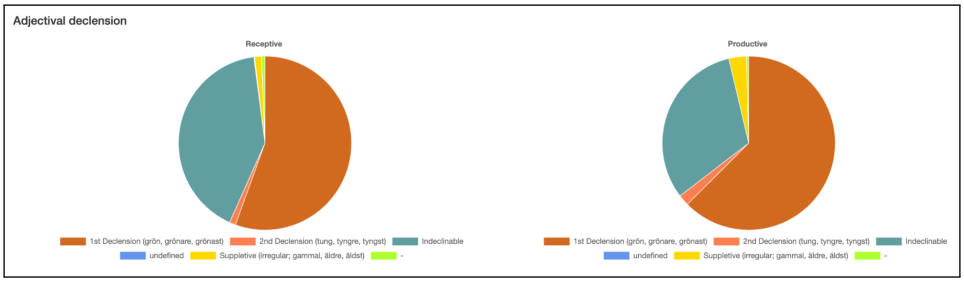


Figure 23: Adjectival declensions at B1-level, SweL2P

of indeclinable adjectives. Hence, learners meet both of these in nearly equal proportion, and that might complicate learning to distinguish the two. Furthermore, it must be noted that this only gives us information about the lemmas and their declension and not how often the learners actually read or use the comparative and superlative forms from each declension. If a teacher has seen that their students seem to have some problems with adjectival inflection, seeing these frequencies they can decide to try to provide more examples of the different declensions and additional exercises.

At B1 level (Figure 23) we see that the first declension continues to be the most common in both datasets, however, its proportion in the productive data increases. Suppletive adjectives can be seen to decrease clearly in the productive data and less clearly in the receptive data. At C1 level, finally, the proportion of first declension and indeclinable adjectives seems fairly stable in comparison to B1.

Interestingly, suppletive adjectives such as *dålig, sämre, sämst* 'bad, worse, worst' are more common in the productive texts than in the course books at A1 level. However, the suppletive adjective proportion has been reduced somewhat in the productive data at C1 level. Nevertheless, as noted for the A1 level, all levels would require more information regarding the contexts and inflectional forms in which the adjectives are used to determine how much exposure the learners get to the different patterns which the inflectional paradigms show.

Notably the adjectives which are listed as indeclinable are only indeclinable as far as comparison goes, and even in this sense they might sometimes be metaphorically compared (e.g. *Kan man bli **svenskare** än så?* 'Can one become **more Swedish** than that?'). Just like adjectives from the first and the second declension, these "indeclinable" adjectives also have to agree with the nominal correlate in terms of gender and number, and with the noun phrase for definiteness: *dålig, dåligt, dåliga* 'bad'. The first declension has very regular inflection to show agreement with the nominal correlate *grön,*

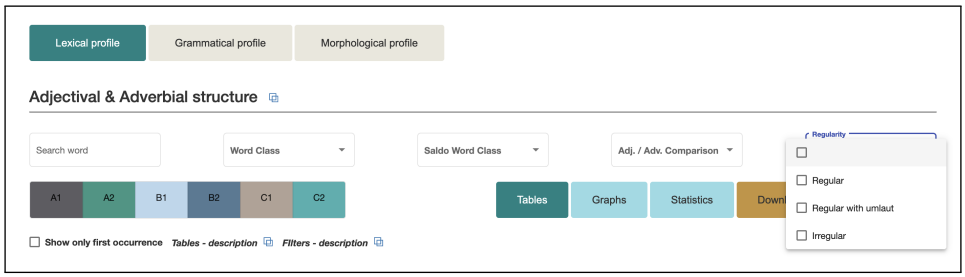


Figure 24: Filtering options under regularity in the adjectival and adverbial structure subsection, SweL2P

grönt, gröna 'green'. The same is true of the second declension: *tung, tungt, tunga* 'heavy'.

Sadly, the current view in the lexical profile does not tell us how common the different inflectional forms are and previous research has shown that learners tend to overuse non-inflected forms. To assess the difficulty of learning the difference of the first declension and the indeclinable adjectives, we need to take a closer look at the frequency and use of the different inflectional forms. We can do this by querying the corpora as shown above (Section 6.2).

Since suppletive adjectives are so common on the lower proficiency levels, it is very important that adjectival inflection and comparison is not only taught on the basis of the regular first declension, but includes suppletive adjectives from the start. It is also important to consider how often inflected suppletive verbs are used in the texts that learners read and whether noticing (Schmidt 1990) techniques might be needed to help learners learn the different ways in which adjectives can be inflected to show comparison.

7.2 *The adjectival and adverbial structure profile*

Adjectives can also be browsed through the option called Adjectival and adverbial structure (see Figure 24). Like adjectival declension, this has been based on all lemmas which have been part-of-speech tagged as adjectives, but this also includes all adverbs, and the lemmas have been manually annotated in Legato (Alfter et al. 2019) according to the type of comparison that can be used with them: morphological (*-are, -ast* '-er, -est') or periphrastic (*mer, mest* 'more, most'). In addition, manual annotations also marked whether the morphological inflection is regular. This means that this profile to some extent overlaps with the subprofile for adjectival declension (see Section 7.1), although, it provides other opportunities for filtering and includes two parts of speech.

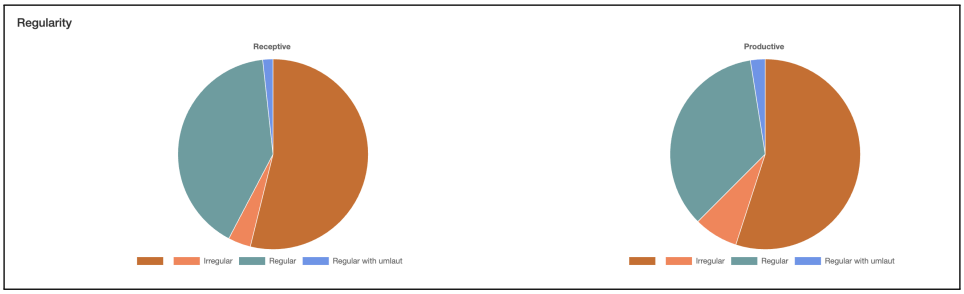


Figure 25: Adjective regularity at A1-level including whether stem change with umlaut, SweL2P

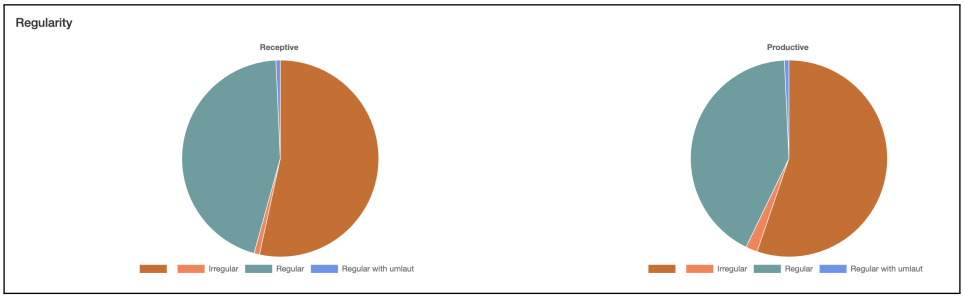


Figure 26: Adjective regularity at C1-level including whether stem change with umlaut, SweL2P

In this subprofile it is also possible to filter adjectives that have regular inflection with a stem change with umlaut such as *stor, större, störst* 'big, bigger, biggest'. These adjectives are clearly relatively rare at all levels (see Figures 25 and 26). It is important in teaching to remember how rare these types of adjectives are in comparison with others, since this is likely to make it more difficult to learn when to use the umlaut characters when writing Swedish. Initial studies on Swedish as a second language in Finland have shown that Swedish spelling connected with the umlaut letters, ä, ö, can pose challenges to learners even if their first language contains the same characters (Lindström Tiedemann et al. 2024b).

7.3 *The multiword expression subprofile and the study of adjectival MWEs*

For more advanced learners it can be of interest (see e.g. Pawley & Syder 1983, Wray 2002) to practise multiword expressions as they have been proven to be important for idiomatic language use (Paquot 2019). In the subprofile called

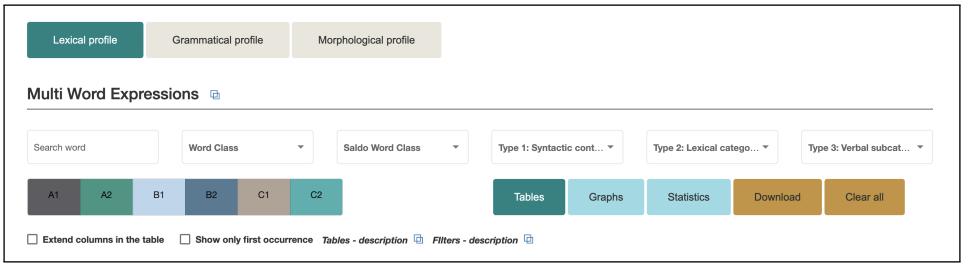


Figure 27: Filtering options for multiword expressions, SweL2P

Figure 28: Automatically annotated adjectival MWEs according to the Saldo word class filter over CEFR levels, SweL2P

Multiword expressions (MWEs), you also find adjectival information.[13] This subprofile provides a possibility to filter automatically annotated multiword expressions (MWEs) based on the manual annotation according to different types (1–3) as well as the two different POS tagsets (see Figure 27).

We can filter the MWEs by automatically annotated adjectival MWEs (see Figure 28). This shows us that the frequency of adjectival MWEs varies at the different CEFR-levels and that there are clear differences between the receptive and productive data.

The first MWE filter (Type 1) separates MWEs that can be non-contiguous, i.e. that allow insertion of other words due to word order restraints or other factors, e.g. *fatta beslut* 'make decision'), from contiguous ones, i.e. the ones that cannot be separated, e.g. *Vita Huset* 'White House'. There are no non-contiguous adjectival MWEs in the SweL2P (Figure 29).

13 The manual annotation of multiword expressions has been carefully described in Lindström Tiedemann et al. (2024a).

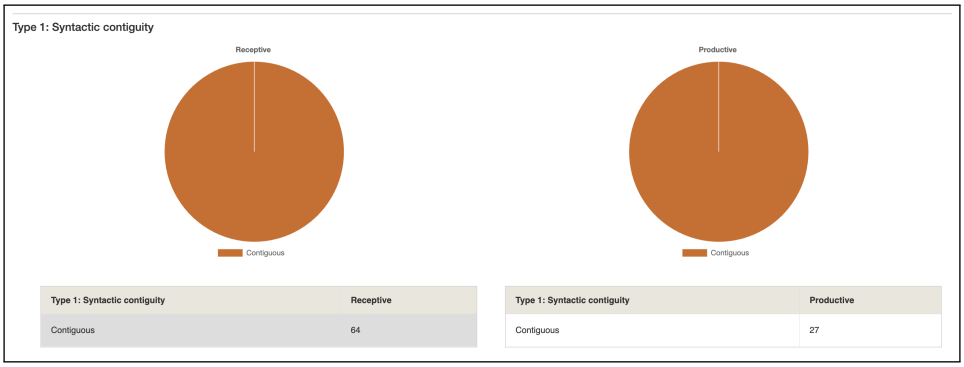


Figure 29: Contiguity of adjectival MWEs in SweL2P

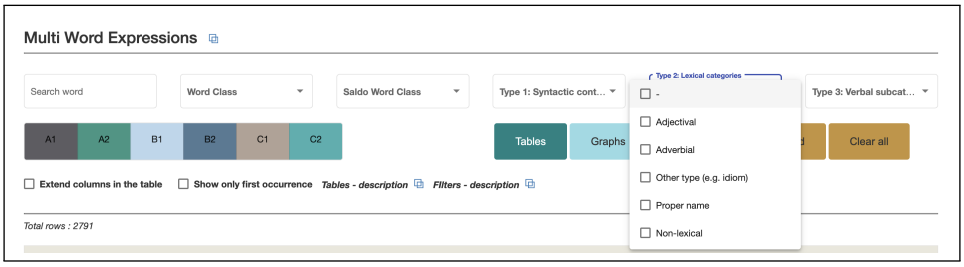
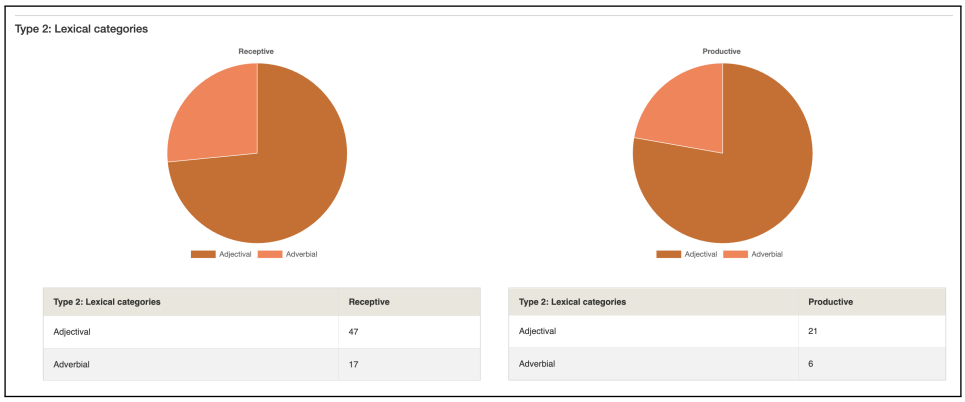


Figure 30: Lexical subcategories under Type 2 filter, SweL2P

The second filter (Type 2, Figure 30) includes subcategorisation by lexical class: adjectival, adverbal, interjections, nominal, non-lexical, other types (i.e. idioms), proper names, verbal according to manual annotation (for more information see Lindström Tiedemann et al. 2024a). Even though we have checked the reliability of the automatic annotation (Volodina, Alfter, Lindström Tiedemann et al. 2022), in many ways this is an important way to provide users with a possibility to see that different analyses may be possible in certain contexts. By comparing the Saldo word class and the manual annotation under type 2, we can also check the reliability of the automatic annotation of MWEs by Sparv. We see that there is a certain difference but this does not necessarily mean that Sparv should be seen as unreliable with regards to adjectival MWEs, but rather that the classification can depend on the guidelines for POS categorisation. Still in relation to teaching it is valuable to see that some of the adjectival MWEs are used primarily in typically adjectival functions, wheras others are used more in adverbial functions (Figure 31) and therefore categorised as adverbial MWEs in the manual annotation. This might cause challenges to learners and should be studied more closely.



Type 2: Lexical categories	Receptive
Adjectival	47
Adverbial	17

Type 2: Lexical categories	Productive
Adjectival	21
Adverbial	6

Figure 31: Adjectival MWEs according to the manual annotation of lexical categories, SweL2P

We now select (1) Saldo word class: Multiword adjective and (2) Type 2: adverbial, so that we can see which MWEs that are used adverbially. The first to appear is sadly a mistaken MWE annotation by Sparv, *i sin* 'in his/her/their' (A2, B1, C1, C2), which is not used as an MWE. Instead this is the preposition *i* 'in', and the pronoun *sin* 'his, her, their'. However, from the B2-level there are several adjectival MWEs which are in fact commonly used adverbially: e.g. *i högsta hugg* 'at the ready', *i blodet* 'in the blood', *i sken* 'bolting, running away', *i tur* 'in turn', *till hands* 'at hand'. These are often not completely transparent or compositional and they can often pose challenges even to advanced learners (cf. Alfter, Tiedemann et al. 2022, Lindström Tiedemann et al. 2024a).

The final MWE filter (Type 3) categorises verbal MWEs further into subcategories, such as support verb MWEs, particle verbs and reflexive verbs and hence will not be treated here any further.

8 *Downstream NLP tasks with the lexical profile*

The SweL2P resource is also useful for research and development within the area of Natural Language Processing (NLP). Historically, Sen*Lex and its predecessors – SVALex and SweLLex – have been extensively used for multiple downstream NLP tasks, which we shortly outline below: Development of machine learning algorithms for text and sentence classification, including CEFR-grading of learner essays, readability classification of reading comprehension texts and complexity prediction on a sentence level (e.g. Pilán, Volodina & Zesch 2016, Pilán, Volodina & Borin 2016). Linguistic fea-

tures, such as graded vocabulary lists, have been used for training predictive systems. One of the most important insights from this work is that lexical features have proven to be the most reliable predictors among a variety of other linguistic features (Pilán & Volodina 2018). These experiments have resulted in an online demo-tool for classification of texts (Volodina, Pilán & Alfter 2016). Recently, the work on essay classification has been resumed, this time relying on the Large Language Models (LLMs) where linguistic features are of less importance (Muñoz Sánchez, Alfter et al. 2024, Muñoz Sánchez, Dobnik et al. 2024).

Automatic complexity prediction of single lexical items (Alfter & Volodina 2018), where the two graded vocabulary lists (SVALex and SweLLex) have been used to train automatic systems to classify, at which level a learner can be expected to know a word (receptively or productively). Experiments on different approaches to predicting the difficulty of a lexical item (especially important for items that are not in Sen*Lex and hence have not been linked to a CEFR level through course books or learner essays) have resulted in an experimental suite of tools, CEFR-tools (Alfter 2021). Lately, the abilities of Language Language Models to predict lexical complexity have also been tested, where SVALex and similar graded lists have been used as a test set (Alfter 2024).

The same vocabulary lists for Swedish have been very important input for two methodological studies on crowdsourcing linguistic annotations from second language learners. In one study the object of annotation was multiword expressions (Alfter, Tiedemann et al. 2022), which was later analysed more qualitatively in Lindström Tiedemann et al. (2022). Another study focused on single lexical items (Volodina, Alfter & Tiedemann 2022). In both cases the task was to rank the items from easy to difficult using comparative judgment settings (best-worst scaling). A very encouraging conclusion from the two experiments was that, when annotation is solicited in a comparative setting, second language learners are as reliable in predicting the difficulty of lexical items as are experts, such as language teachers and assessors.

Another very promising direction, where the lexical profile in SweL2P with its lexicographically enriched annotation has extreme potential, is the development of teaching materials, such as vocabulary exercises for language learners within the context of Intelligent Computer-Assisted Language Learning (ICALL), similar to Volodina, Pilán, Borin et al. (2014), Volodina & Pijetlovic (2015), Alfter et al. (2018). In the future, there is a lot to be explored. For example: using the SweL2P, we can filter relevant items of interest (e.g. the most frequent adjectives at the A1 level, productively and receptively) and extract sentences containing those adjectives.

Using the Lärka[14] facilities, we can easily create stepwise vocabulary training:

1. First, learners work with gapped sentences, where the target adjectives have been removed, trying to fill in the gaps with appropriate adjectives (training productive knowledge), see the mock-example below.

```
TASK: Fill in the missing adjectives
-----------------------------------
Du är en [b.......] fotograf !
Klockan är [h.......] åtta på kvällen .
Det finns en [g.......] soffa , två fåtöljer och ett bord i
    vardagsrummet .
```

2. Second, they listen to the sentences and fill in the gaps based on what they can hear (training listening and spelling).

```
TASK: Listen to the sentence and fill in the word you hear (an
    adjective)
-----------------------------------
[play >] Du är en [.......] fotograf !
[play >] Klockan är [.......] åtta på kvällen .
[play >] Det finns en [.......] soffa , två fåtöljer och ett
    bord i vardagsrummet .
```

3. Finally, they get a list of missing adjectives ("word bank") that should be matched with the gaps (training graphical form recognition, as well as grammatical and semantic constraints).

```
TASK: Use one of the words to fill in the gaps. One word is
    ``odd'' :
halv , bra , gammal , nästa
-----------------------------------
Du är en [.......] fotograf !
Klockan är [.......] åtta på kvällen .
Det finns en [.......] soffa , två fåtöljer och ett bord i
    vardagsrummet .
```

4. Each of the exercise types can be used to log reaction times (or similar functionality) to permit use of the generated exercises for experimental research into cognitive processes and learning.

14 A language learning platform developed at Språkbanken Text, Sweden: https://spraakbanken.gu.se/larka

It is worth mentioning that we are collecting available scripts that could be reused by other researchers or students, in a github repository.[15] Among others, we provide a script for the generation of frequency lists from COCTAILL and SweLL essays (or resources with similar annotation); for interlinking such frequency lists with other lexicographic information; for generation of best-worst scaling mini-sets of four items using a redundancy-reducing algorithm. New scripts are contiuously being added, therefore, the current description might not provide full coverage.

All of the datasets that are relevant for the Swedish (Lexical) Profile are available for download from Språkbanken's resource webpage and are citable as resources:

- COCTAILL (Volodina & Pilán 2014)
- SweLL-pilot (Volodina, Pilán, Enström et al. 2016b)
- SVALex (Volodina, Pilán & François 2016)
- SweLLex (Volodina & Pilán 2016)
- Sen*Lex (Alfter et al. 2023)
- L2Lex-Adj (Lindström Tiedemann et al. 2023a)
- L2Lex-AdjAdv (Lindström Tiedemann et al. 2023b)
- Swe-MWELex (Lindström Tiedemann et al. 2023c)

9 *Final remarks*

The lexical part of the Swedish L2 profile (SweL2P) provides many opportunities for teachers and language researchers to explore L2 Swedish from different perspectives in innovative empirical ways. It can be used as a way to identify possible needs for further teaching and for new teaching materials, including functioning as a way to design DDL exercises. Furthermore, users can use SweL2P for research and to identify new research questions in a data driven manner. In addition to this, the Swedish L2 profile can be used to provide new CALL/ICALL applications which are sensitive to CEFR levels.

The lexical part of the Swedish L2 profile has been carefully designed in a way that ascertains its comparability with reference corpora available through Korp. Since the corpus queries used for the Swedish L2 profile are completely open, users can reuse them on other corpora which have been annotated in the same way and compare L2 usage and L2 course-book usage to the use in a variety of L1 genres, thus supporting e.g. CIA analyses.

15 https://github.com/spraakbanken/L2_profiles

The Swedish L2 profile is open to everyone and unlike other similar resources the development team has made sure that the user does not only have access to information about what sources the different lists are based on, but is also given the opportunity to access the empirical data. Our data is openly available, apart from the data from the learner texts which require further authentication. The fact that SweL2P facilitates comparison with other Swedish corpora means a great potential in connection with research and teaching. It can help us compare frequencies, contexts and also find examples that can be used in designing exercises for teaching. This gives the resource a more descriptive angle than previous profiles such as the English Vocabulary Profile which may appear more prescriptive and which only provides occasional examples from the empirical data. It also makes our profile easier to use as a basis for further research.

It is a fact that languages and research domains that can boast rich data collections are subject to more empirical and data-intensive research (Perc 2014, Søgaard 2022). This makes us believe that now, with the Sen*Lex vocabulary list and the lexical profile in the Swedish Second Language Profile (SweL2P) available for research and development, the field of Swedish as a second language, research on L2 Swedish as well as ICALL for Swedish and other related research fields will receive a boost.

In the future, we are expecting both short-term and long-term impact from the Swedish Second Language Profile on the fields of Swedish as a Second Language, Learner Corpus Research (nationally and internationally), and NLP- and AI-based approaches to L2 Swedish, both with respect to research and teaching.

Acknowledgments

Work on the Swedish L2 Profile has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *Development of lexical and grammatical competences in immigrant Swedish*, P17-0716:1 (2018–2021). Work on the article for the first author has been supported by the University of Helsinki (Finland); the last author has been supported by Nationella språkbanken and HUMINFRA, both funded by the Swedish Research Council (2018–2024, contract 2017-00626; 2022–2024, contract 2021-00176) and their participating partner institutions.

References

Alfter, David. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Data Lingvistica 31, University of Gothenburg.

Alfter, David. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2024)*.

Alfter, David, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann & Elena Volodina. 2018. From language learning platform to infrastructure for research on language learning. In *CLARIN Annual Conference 2018*, 53.

Alfter, David, Rémi Cardon & Thomas François. 2022. A dictionary-based study of word sense difficulty. In Rodrigo Wilkens, David Alfter, Rémi Cardon & Núria Gala (eds.), *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, 17–24. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.readi-1.3.

Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2019. LEGATO: A flexible lexicographic annotation tool. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 382–388.

Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2023. *Sen*Lex (updated: 2023-04-20)*. [*Data set*]. DOI: 10.23695/b417-1j26.

Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2022. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. *Northern European Journal of Language Technology* 7(1).

Alfter, David & Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the thirteenth workshop on innovative use of NLP for Building Educational Applications*, 79–88.

Allén, Sture. 1970. *Nusvensk frekvensordbok baserad på tidningstext* [Frequency dictionary of present-day Swedish]. Vol. 1: *Graford. Homografkomponenter.* [Graphic words. Homograph components]. Stockholm: Almqvist & Wiksell.

Allén, Sture. 1971. *Nusvensk frekvensordbok baserad på tidningstext* [Frequency dictionary of present-day Swedish]. Vol. 2: *Lemman* [Lemmas]. Stockholm: Almqvist & Wiksell.

Arle, Solveig. 2018. *Hur används primord för att skriva lättläst?: En undersökning av tre återberättade romaner*. University of Helsinki. (MA dissertation).

Axelsson, Monica. 1994. *Noun phrase development in Swedish as a second language: A study of adult learners acquiring definiteness and the semantics and morphology of adjectives*. Stockholm University. (Doctoral dissertation).

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

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Louise Holmer & Arild Matsson. 2025. 10 korp: Språkbanken's word research platform. In *Sixty years of Swedish computational lexicography*. Dana Dannélls, Kristian Blensenius & Lars Borin (eds.). De Gruyter. 175–194. DOI: 10.1515/9783111577234-010.

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

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

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp-the corpus infrastructure of Språkbanken. In *LREC*, vol. 2012, 474–478.

Boulton, Alex & Nina Vyatkina. 2021. Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning and Technology* 25(3). 66–89.

Caines, Andrew & Paula Buttery. 2018. The effect of task and topic on opportunity of use in learner corpora. In *Learner corpus research: New perspectives and applications*. Vaclav Brezina & Lynne Flowerdew (eds.). London & New York: Bloomsbury Publishing Academic. 5–27.

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

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

Capel, Annette. 2015. The English Vocabulary Profile. In Julia Harrison & Fiona Barker (eds.), *English Profile in practice*, 9–27. Cambridge University Press.

Clahsen, Harald. 1985. Profiling second language development: A procedure for assessing L2 proficiency. In *Modelling and assessing second language acquisition*. Kenneth Hyltenstam & Manfred Pienemann (eds.). (Multilingual Matters 18). Clevedon: Multilingual Matters. 283–331.

Collberg, Philippe. 2021. Grammatik i skolan: Till hjälp för skrivande? In Johan Brandtler & Mikael Kalm (eds.), *Nyanser av grammatik: Gränser, mångfald, fördjupning*. Studentlitteratur.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe Publishing.

Crystal, David, Paul J Fletcher & Michael Garman. 1976. *The grammatical analysis of language disability: A procedure for assessment and remediation*. London: Edward Arnold.

Davidson, Douglas J, Peter Indefrey & Marianne Gullberg. 2008. Words that second language learners are likely to hear, read, and use. *Bilingualism: Language and Cognition* 11(1). 133–146.

François, Thomas, Elena Volodina, Ildikó Pilán & Anaïs Tack. 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)*, 213–219.

Gilquin, Gaëtanelle & Sylviane Granger. 2010. How can data-driven learning be used in language teaching? In Michael McCarthy & Anne O'Keeffe (eds.), *The Routledge Handbook of Corpus Linguistics*, 359–370. New York: Routledge.

Goddard, Cliff. 2012. Semantic primes, semantic molecules, semantic templates: Key concepts in the NSM approach to lexical typology. *Linguistics* 50(3). 711–743.

Goddard, Cliff & Susanna Karlsson. 2008. Re-thinking THINK in contrastive perspective: Swedish vs. English. In Cliff Goddard (ed.), *Cross-linguistic semantics*, 225–240. Philadelphia & Amsterdam: John Benjamins Publishing Company.

Granfeldt, Jonas & Malin Ågren. 2014. SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing* 31(3). 285–305.

Granfeldt, Jonas, Pierre Nugues & Emil Persson. 2005. Direkt Profil: A system for evaluating texts of second language learners of French based on developmental sequences. In *43rd Annual Meeting of the Association of Computational Linguistics*, 53–60.

Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer K. (ed.), *Languages in contrast. Text-based cross-linguistic studies*, 37–51. Lund University Press: Lund. http://hdl.handle.net/2078.1/75847.

Granger, Sylviane. 2002. A bird's eye view of learner corpus research. In Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 3–33. Second Language Acquisition & Foreign Language Teaching. Amsterdam & Philadelphia: Benjamins.

Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. *Corpora and language teaching* 33. 13–32.

Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1). 7–24. DOI: 10.1075/ijlcr.1.1.01gra.

Hammarstedt, Martin, Anne Schumacher, Lars Borin & Markus Forsberg. 2022. *Sparv 5 user manual*. Göteborg.

Hawkins, John A & Luna Filipović. 2012. *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*, vol. 1. Cambridge University Press.

Hult, Ann-Kristin, Sven-Göran Malmgren & Emma Sköldberg. 2010. Lexin - a report from a recycling lexicographic project in the north. In Anne Dykstra & Tanneke Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress* (*Leeuwarden, 6-10 July 2010*), 800–809. Leeuwarden/Ljouwert, The Netherlands: Fryske Akademy. https://euralex.org/category/publications/euralex-leeuwarden-2010/.

Hultman, Tor & Margareta Westman. 1977. *Gymnasistsvenska*. Liber-Läromedel.

Källgren, Gunnel, Sofia Gustafson-Capková & Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Department of Linguistics, Stockholm University. https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf.

Keßler, Jörg-U & Mathias Liebner. 2011. Diagnosing L2 development: Rapid profile. In Manfred Pienemann & Jörg-U Keßler (eds.), *Studying processability theory: An introductory textbook*, 133–148. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Lindström Tiedemann, Therese, David Alfter, Yousuf Ali Mohammed, Daniela Helena Piipponen, Beatrice Silen & Elena Volodina. 2024a. Multiword expressions in Swedish as a second language: Taxonomy, annotation, and initial results. In *Multiword expressions in lexical resources*. Voula Giouli & Verginica Barbu Mititelu (eds.). (Phraseology and Multiword Expressions). Germany: Language Science Press. 309–348. DOI: 10.5281/zenodo.10998647.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2022. CEFR-nivåer och svenska flerordsuttryck. In Siv Björklund, Marianne Haagensen Bodil Nordman & Anders Westerlund (eds.), *Svenskan i Finland 19: Föredrag vid den nittonde sammankomsten för beskrivningen av svenskan i Finland, Vasa den 6–7 maj 2021*, 218–233. Svensk-Österbottniska Samfundet.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2023a. *L2LexAdj* (*updated: 2023-04-20*).[*Data set*]. DOI: 10.23695/z8wj-1r23.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2023b. *L2LexAdjAdv* (*updated: 2023-04-20*).[*Data set*]. DOI: 10.23695/3kcx-s405.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2023c. *Swedish MWELex* (*updated: 2023-04-20*).[*Data set*]. DOI: 10.23695/352q-wa92.

Lindström Tiedemann, Therese, Yousuf Ali Mohammed & Elena Volodina. In preparation. *Swedish grammar profiling for empirical L2 research and teaching*.

Lindström Tiedemann, Therese, Daniela Piipponen, Lisa Södergård & Erik Axelson. 2024b. *Normalisering som analysredskap för andraspråkssvenska*. Unpublished presentation at Svenskan i Finland 21.

Löhndorf, Simone. 2021. *Development of adjectival use and meaning structures in Swedish students' written production*. Lund University. (Doctoral dissertation).

Mohsen, Mohammed Ali, Sultan Althebi, Rawan Alsagour, Albatool Alsalem, Amjad Almudawi & Abdulaziz Alshahrani. 2024. Forty-two years of computer-assisted language learning research: A scientometric study of hotspot research and trending issues. *ReCALL* 36(2). 230–249.

Muñoz Sánchez, Ricardo, David Alfter, Simon Dobnik, Maria Irena Szawerna & Elena Volodina. 2024. Jingle BERT, Jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT. In *Proceedings of the Workshop on Natural Language Processing for Computer-Assisted Language Learning* (*NLP4CALL 2024*).

Muñoz Sánchez, Ricardo, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann & Elena Volodina. 2024. Did the names I used within my essay affect my score? Diagnosing name biases in automated essay scoring. In *Proceedings of the workshop on computational approaches to language data pseudonymization* (*CALD-pseudo 2024*), 81–91. https://aclanthology.org/2024.caldpseudo-1.10/.

Nieto Piña, Luis. 2019. *Splitting rocks: Learning word sense representations from corpora and lexica*. Data Lingvistica 30, University of Gothenburg.

Paquot, Magali. 2019. The phraseological dimension in interlanguage complexity research. *Second language research* 35(1). 121–145.

Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication*, 191–226. London: Longman.

Perc, Matjaž. 2014. The Matthew effect in empirical data. *Journal of The Royal Society Interface* 11(98). 20140378. DOI: 10.1098/rsif.2014.0378.

Philipsson, Anders. 2007. *Interrogative clauses and verb morphology in L2 Swedish: Theoretical interpretations of grammatical development and effects of different elicitation techniques*. Centrum för tvåspråkighetsforskning. (Doctoral dissertation).

Pienemann, Manfred & Alison Mackey. 1993. An empirical study of children's ESL development and Rapid Profile. *ESL development: Language and literacy in schools* 2. 115–259.

Pilán, Ildikó & 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*, 49–58.

Pilán, Ildikó, Elena Volodina & Lars Borin. 2016. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues* 57(3). 67–91.

Pilán, Ildikó, Elena Volodina & Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2101–2111.

Ravid, Dorit & Ronit Levie. 2010. Hebrew adjectives in later language text production. *First Language* 30(1). 27–55. DOI: 10.1177/0142723709350529.

Römer, Ute. 2023. Usage-based approaches to Second Language Acquisition vis-à-vis Data-Driven Learning. *TESOL Quarterly*.

Schmidt, Richard W. 1990. The role of consciousness in second language learning. *Applied linguistics* 11(2). 129–158.

Sköldberg, Emma, Louise Holmer, Elena Volodina & Ildikó Pilán. 2019. State-of-the-art on monolingual lexicography for Sweden. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 7(1). 13–24.

Søgaard, Anders. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 conference on empirical methods in natural language processing*, 5254–5260.

Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska akademiens grammatik*. Svenska Akademien & Norstedts ordbok.

Tenfjord, Kari, Paul Meurer & Knut Hofland. 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. In *LREC'06*, 1821–1824.

Viberg, Åke. 1990. Svenskans lexikala profil. In Erik Andersson & Marketta Sundman (eds.), *Svenskans beskrivning 17*. Åbo Akademi.

Viberg, Åke. 1992. Tvärspråklig lexikologi med svenskan i centrum. *Nordiske studier i leksikografi* 1.

Viberg, Åke. 2006a. The typological profile of Swedish – an introduction. *STUF - Language Typology and Universals* 59. 3–11.

Viberg, Åke. 2006b. Towards a lexical profile of the Swedish verb lexicon. *STUF - Language Typology and Universals* 59. 103–129.

Viberg, Åke. 2013. Seeing the lexical profile of Swedish through multilingual corpora. The case of Swedish åka and other vehicle verbs. In Karin Aijmer & Bengt Altenberg (eds.), *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson*. John Benjamins.

Volodina, Elena. 2024. On two SweLL learner corpora–SweLL-pilot and SweLL-gold. In *Proceedings of the Huminfra Conference (HiC 2024)*, 83–94. DOI: 10.3384/ecp205012.

Volodina, Elena, David Alfter, Therese Lindström Tiedemann, Maisa Susanna Lauriala & Daniela Helena Piipponen. 2022. Reliability of automatic linguistic annotation: native vs non-native texts. In *Selected papers from the CLARIN Annual Conference 2021*.

Volodina, Elena, David Alfter & Therese Lindström Tiedemann. 2022. Crowdsourcing ratings for single lexical items: a core vocabulary perspective. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 10(2). 5–61.

Volodina, Elena, David Alfter & Therese Lindström Tiedemann. 2024. Profiles for Swedish as a second language: lexis, grammar, morphology. In *Proceedings of the Huminfra Conference (HiC 2024)*, 10–19. DOI: 10.3384/ecp205002.

Volodina, Elena, Yousuf Ali Mohammed & Therese Lindström Tiedemann. 2022. Lyxig språklig födelsedagspresent from the Swedish word family. In Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Markus Forsberg & Shafqat Virk (eds.), *Live and Learn: Festschrift in honor of Lars Borin*. GU-ISS-2022-03, 153–160. Department for Swedish, Multilingualism, Language Technology, University of Gothenburg.

Volodina, Elena, Yousuf Ali Mohammed & Therese Lindström Tiedemann. 2021. CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, 178–189.

Volodina, Elena, Yousuf Ali Mohammed & Therese Lindström Tiedemann. 2024. Swedish word family resource: Construction, applicability, strengths and first experiments. *ITL-International Journal of Applied Linguistics*. DOI: 10.1075/itl.22026.vol.

Volodina, Elena & Dijana Pijetlovic. 2015. Lark trills for language drills: Text-to-speech technology for language learners. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 107–117.

Volodina, Elena & Ildikó Pilán. 2014. *COCTAILL.* [*Data set*]. DOI: 10.23695/v8xs-p564.

Volodina, Elena & Ildikó Pilán. 2016. *SweLLex.* [*Data set*]. DOI: 10.23695/6h3v-zw25.

Volodina, Elena, Ildikó Pilán & David Alfter. 2016. Classification of Swedish learner essays by CEFR levels. *CALL communities and culture–short papers from EUROCALL* 2016. 456–461.

Volodina, Elena, Ildikó Pilán, Lars Borin & Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. In *LREC*, 3973–3978.

Volodina, Elena, Ildikó Pilán, Stian Rødven Eide & Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a second language. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*. Linköping University Press.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg & Monica Sandell. 2016a. Swell on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the 10th Language Resources and Evaluation Conference* (*LREC*). Portorož, Slovenia.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg & Monica Sandell. 2016b. *SweLL-pilot.* [*Data set*]. DOI: 10.23695/6h3v-zw25.

Volodina, Elena, Ildikó Pilán & Thomas François. 2016. *SVALex, v.01.* [*Data set*]. DOI: 10.23695/rwrq-1w38.

Volodina, Elena, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse & Thomas François. 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*, 76–84.

Warren, Martin. 2016. Introduction to data-driven learning. In Fiona Farr & Liam Murray (eds.), *The Routledge Handbook of Language Learning and Technology*, 337–347. London & New York: Routledge.

Wierzbicka, Anna. 1996. *Semantics: Primes and universals*. Oxford: Oxford University Press.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wulff, Stefanie. 2020. Usage-based approaches. In Nicole Tracy-Ventura & Magali Paquot (eds.), *The Routledge Handbook of Second Language Acquisition and Corpora*, 175–188. New York & Oxford: Routledge.

Yamaguchi, Nami, David Alfter, Kaori Sugiyama & Thomas François. 2022. Towards a verb profile: distribution of verbal tenses in FFL textbooks

and in learner productions. In *Proceedings of the 11th workshop on natural language processing for computer-assisted language learning* (*NLP4CALL 2022*), 123–142.

List of abbreviations

CALL Computer-Assisted Language Learning
CEFR Common European Framework of Reference for Languages
CIA Contrastive Interlanguage Analysis
COCA Corpus of Contemporary American English
COCTAILL COrpus of Coursebook Texts As Input for Language Learning
DDL Data Driven Learning
ICALL Intelligent Computer-Assisted Language Learning
L1 first language, cf. mother tongue
L2 second language, i.e. non-native language
L2Lex-Adj a language resource containing a L2 list of adjectives with associated information used as an input to a subsection of Swedish L2 lexical profile devoted to declension of adjectives
L2Lex-AdjAdv a language resource containing a L2 list of adjectives and adverbs with associated information used as an input to a subsection of Swedish L2 lexical profile devoted to adverbial and adjectivar structure
MWE Multi-Word Expressions
NLP Natural Language Processing
NSM Natural Semantic Metalanguage
POS part of speech
SAG Swedish Academic Grammar
SAOL Svenska Akademiens Ordlista
Sen*Lex sense-based lexicon for L2 Swedish
SenSVALex sense-based SVALex
SenSweLLex sense-based wordlist of Swedish as a second language based on learner essays (SweLL-pilot)
SLA Second Language Acquisition
SO Svenska ordbok
SVALex wordlist of Swedish as a second language based on course books (COCTAILL)
SweL2P Swedish Second Language Profile

SweLL	corpus of learner essays, Swedish Learner Language
SweLLex	wordlist of Swedish as a second language based on learner essays (SweLL-pilot)
Swe-MWELex	a language resource containing an L2 list of mult-word expressions with associated information used as an input for a subsection of Swedish L2 lexical profile devoted to multi-word expressions
TTR	type-token ratio

Corresponding authors

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

Elena Volodina
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
elena.volodina@svenska.gu.se

David Alfter
Department of Literature, History of Ideas, and Religion
University of Gothenburg
david.alfter@lir.gu.se