Reliability of Automatic Linguistic Annotation : Native vs Non-native Texts

Volodina, Elena; Alfter, David; Lindström Tiedemann, Therese; Lauriala, Maisa Susanna; Piipponen, Daniela Helena

Monachini, Monica; Eskevich, Maria

2022-07

http://hdl.handle.net/10138/346483

Volodina, E, Alfter, D, Lindström Tiedemann, T, Lauriala, M S & Piipponen, D H 2022, Reliability of Automatic Linguistic Annotation : Native vs Non-native Texts. in M Monachini & M Eskevich (eds), Selected papers from the CLARIN Annual Conference 2021. Linköping Electronic Conference Proceedings, vol. 189, Linköping University Electronic Press, Linköping, pp. 151-167, CLARIN Annual Conference , 27/09/2021. https://doi.org/10.3384/ecp18914

Reliability of Automatic Linguistic Annotation: Native vs Non-native Texts

Elena Volodina, David Alfter
University of Gothenburg, Sweden
name.surname@gu.se

Therese Lindström Tiedemann, Maisa Lauriala, Daniela Piipponen
University of Helsinki, Finland
name.surname@helsinki.fi

Abstract

We present the results of a manual evaluation of the performance of automatic linguistic annotation on three different datasets: (1) texts written by native speakers, (2) essays written by second language (L2) learners of Swedish in the original form and (3) the normalized versions of learner-written essays. The focus of the evaluation is on lemmatization, POS-tagging, word sense disambiguation, multi-word detection and dependency annotation. Two annotators manually went through the automatic annotation on a subset of the datasets and marked up all deviations based on their expert judgments and the guidelines provided. We report Inter-Annotator Agreement between the two annotators[1] and accuracy for the linguistic annotation quality for the three datasets, by levels and linguistic features.

1 Introduction

In the current project, *Development of grammatical and lexical competences in immigrant Swedish*,[2] we explore profiling of lexical and grammatical competences among second language (L2) learners of Swedish based on two corpora. The coursebook corpus, COCTAILL (Volodina et al., 2014), and the L2 Swedish learner corpus, SweLL-pilot (Volodina et al., 2016), are used for qualitative and quantitative analysis of lexical and grammatical categories that L2 learners are exposed to or produce themselves. The texts in the two corpora have been automatically annotated with linguistic information using the Sparv-pipeline (Borin et al., 2016) which is an essential part of the CLARIN infrastructure for the Swedish language. Sparv, in turn, relies on the gold annotation standards from the Stockholm Umeå Corpus (SUC) (Ejerhed et al., 1997) and on the theoretical framework in the Saldo lexicon (Borin et al., 2013). Since the process of linguistic annotation is performed automatically, we need to evaluate to which degree we can expect the results of the annotation to be reliable, so that our theoretical generalizations and conclusions about language learning can factor that in. For this reason, we performed a manual "annotation quality check" of Part-of-Speech (POS) tagging, lemmatization, dependency annotation, identification of multi-word expressions (MWE) and word sense disambiguation (WSD) which we report in this paper.

Previous work suggests that performance of automatic pipelines trained on native language models is non-optimal on L2 language due to a large number of non-words, deviating syntactic patterns and statistical distributions in L2 production (Štindlová et al., 2012). Rubin (2021) shows that the performance of two independent parsers for Dutch drops by ≈7–8% on L2 learner data compared to first language (L1) data. Krivanek and Meurers (2013) have similar results for L2 German, with ≈6% drop in LAS (labeled attachment scores) for dependency parsing of L2 German. Ott and Ziai (2010) have observed that not all L2 deviations have an equally drastic impact on automatic linguistic annotation, e.g. deviations in morphology and word order do not influence the accuracy of POS tagging or syntactic parsing, whereas omission of syntactically important relations, such as subjects and verbs, yields incorrect parses. Meurers and Wunsch (2010) discuss the need for theoretical analysis of linguistic features in learner language with implications for automatic L2 annotation. For example, the three criteria for assigning a part

[1] Note, though, that dependency annotation was checked by one annotator only

[2] Riksbankens jubileumsfond P17-0716:1, project homepage: <https://spraakbanken.gu.se/en/projects/l2profiles>

Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Lauriala and Daniela Piipponen 2022. Reliability of Automatic Linguistic Annotation: Native vs Non-native Texts. *Selected papers from the CLARIN Annual Conference 2021*. Linköping Electronic Conference Proceedings 189: 189 151–167.

of speech (POS) - lexico-semantic, morphological and distributional - are not always applicable to the L2 data, e.g. I was **choiced* for a job; He walked **rapid*; where one or more of the criteria are not followed. However, there is no common consensus (and very little discussion) about which automatic fallback strategies should be preferred in case of automatic annotation of non-native language or whether the principles of L2 annotation should be more drastically revised (Meurers and Wunsch, 2010).

A very dangerous trap in annotation of learner language is to start encoding what the learner meant (which is subjective in nature) rather than objectively describing what has been used. To ensure objectivity in L2 POS-tagging, it might be best if all the three criteria could be encoded separately. This would mean that for the word **rapid* in *He walked *rapid*, three POS codes could be assigned: lexical-POS: adjective; morphological-POS: adjective; and distributional-POS: adverb.

To our knowledge, no evaluation of automatic linguistic annotation on Swedish L2 data has been done yet. In the present paper we present the results of such an evaluation for the Sparv pipeline and our conclusions regarding the applicability of the Sparv pipeline for analysis of L2 data. This experiment complements and extends several investigations of the Sparv pipeline where Sparv has been analyzed from the point of view of automatic tools, models and modules (Ljunglöf et al., 2019), and its performance has been *automatically* evaluated in relation to *native* language (L1) varieties (Berdicevskis, 2020a; Berdicevskis, 2020b), whereas we examine the reliability of annotations *manually* and on several types of language – L1, L2 original and L2 normalized (i.e. corrected). We analyze the performance of the tool by categories and subcategories, as well as in relation to different L2 proficiency levels.

Despite Swedish being the focus of this experiment, we expect our findings to be generalizable to other languages and to the performance of other pipelines on non-native language samples. It is an important study for CLARIN since it evaluates how well part of the CLARIN infrastructure works for both L1 and L2 Swedish, thereby assessing the need for improvements to the current pipeline for Swedish.

2 Notes on Linguistic Terminology

Notion of Lexical Items The way researchers operationalize the construct of a "word" influences the way word statistics and frequency counts are collected and the way different aspects of individual items are analyzed. This has a direct impact upon the application of the collected statistics (Gardner, 2007). One of the most common ways to work with words is based on *lemmas* (=base forms of a word, e.g. `file`) and its derivative version *lemgrams* (=base form + POS, e.g. `file, verb`). There are different ways to define the notion of lemgrams. In our case we rely on the operationalization of lemgram in the Saldo lexicon (Borin et al., 2013) which is used in the Sparv-pipeline (Borin et al., 2016, p.1): "A lemgram is a lexical identifier which refers to an inflection table in the SALDO lexicon (Borin et al., 2013), which provides linkages between lemgrams and word sense identifiers, although the relation is many-to-many." This means that Sparv can differentiate between words of the same part of speech if they belong to different inflectional paradigms, e.g. between the verb `hang-hanged` and the verb `hang-hung`; however, if both the base form **and** the inflectional paradigm are shared, homographical items are not automatically differentiated, e.g. `fil` 'file on a computer' vs `fil` 'driving lane'. Sparv therefore provides a pointer to several possible senses of each identified lexical item (e.g. to different senses of all possible "file" nouns with the same inflectional paradigm). Even word senses are derived from the Saldo lexicon, with regards to their identifiers, descriptors and number of senses per lemgram, and are used in the module for *word sense disambiguation* in the Sparv pipeline.

Notion of a Single-Word Lexical Item *Lemgram* is usually understood as a set of word forms having the same base form and belonging to the same POS, e.g. all occurrences of the word forms *flicka, flickas, flickan,* etc. are counted together since they have the same base form *flicka* 'girl' and the same part-of-speech *noun*. The Sparv annotation takes this a step further, where lemgrams are also differentiated based on inflectional paradigms encoded in Saldo, so that *val* (noun, -et; the neuter gender, 6th declension; 'election; choice') and *val* (noun, -en, -ar; the uter gender, 2nd declension; 'whale') count as two different items in frequency statistics. Besides, due to the recent development in word sense disambiguation approaches for Swedish (Nieto Piña, 2019), it is now possible to collect triples of identifying

information for each lexical item, namely lemma+POS+sense. Thus, for the lexical item *gräva*, verb, we are able to collect frequencies separately for the sense *dig a hole* and for the sense *do research*.

Notion of a Multi-Word Expression The concept of Multi-word expression (MWE) is both broad and vaguely defined. The literature abounds in different terms with similar meanings: *collocations* (Bhalla and Klimcikova, 2019), *phraseological units* (Paquot, 2019), *lexicalized phrases* (Sag et al., 2002), *formulaic sequences* (Wray, 2005), etc. The definition of multi-word expressions in our study is inherited from the Saldo lexicon which is used in the Sparv annotation pipeline, where Saldo forms the lexical knowledge-base. The Saldo definition of MWEs is based on semantic-orthographic principles, i.e. an MWE consists of two or more orthographically defined lexical items, while exhibiting a certain (varying) extent of semantic non-compositionality (Borin, 2021). Each MWE is a lemgram of its own, can have several senses and falls into one of the three structurally-defined broad categories: contiguous, non-contiguous or constructions (Borin, 2021, p.223). However, constructions, which by definition contain open placeholder e.g. *på X bekostnad* 'on X's account', are not yet fully integrated into Saldo, and are therefore not yet automatically processed by the Sparv pipeline either. We accept the Saldo definition of MWEs at face value for this particular investigation limiting ourselves to the first two types of MWEs (see, however, Alfter et al. (2021) for our more refined taxonomy developed within the context of the current project based on the first two MWE types in Saldo).

Part-of-Speech Categories As is clear from the descriptions above, we are focusing on the analysis of annotation tags (and their interpretation) present in the Sparv annotation output, even though they do not always reflect the way we may want to define the categories which they represent. The same concerns part-of-speech (POS) categories.

There are two POS taxonomies used in the output of Sparv: one coming from the model trained on SUC (Gustafson-Capková and Hartmann, 2006), a gold-annotated corpus, with 22 POS categories[3]; and the other based on the Saldo lexicon (Borin et al., 2013), with 37 POS categories[4]. The analysis in this experiment is focused on the SUC-based POS tags (see Appendix A for an overview). There is an option to convert SUC-based POS tags into the universal tagset[5] (Petrov et al., 2011), but the conversion is not fully reliable. Not all POS categories used in the Sparv output correspond to the part-of-speech defined in the Swedish Academy Grammar (SAG) (Teleman et al., 1999), which is the most authoritative description of Swedish grammar. The difference is especially notable in relation to *determiners* which are used in the SUC tagset, but are not among POS categories in SAG. Another difference concerns adverbial usage of neuter adjectives (e.g. `högt`) which in SUC are treated as adverbs but as adjectives in neuter form in SAG (i.e. adjective `hög` + neuter inflection `-t`). The conflicting theoretical views on POS categories may have prompted unnecessary corrections by the annotators.

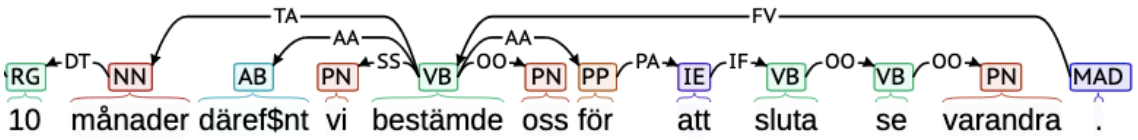

Figure 1: Syntactic tree based on Sparv annotation.

Dependency Relations Categories used by Sparv come from the MAMBA tagset[6] used in the Swedish treebank Talbanken (Nivre et al., 2008). The Mamba tagset contains sixty-five (65) tags including fourteen (14) tags describing punctuation (see Appendix B for a full taxonomy). The dependency relations (DepRels) are split into Root (or head) and Relations (or syntactic functions), e.g. *subject, finite verb,*

[3] `https://spraakbanken.gu.se/korp/markup/msdtags.html`
[4] `https://spraakbanken.gu.se/en/resources/saldo/tagset`
[5] `https://universaldependencies.org/u/feat/index.html`
[6] `https://cl.lingfil.uu.se/~nivre/swedish_treebank/dep.html`

direct object, agent. No conversion to the Universal Dependency Relations[7] (De Marneffe et al., 2014) is offered by Sparv. DepRel tags are used to build syntactic trees (see Figure 1) where syntactic relations are shown through arrows, while POS tags are shown in squares.

3 Experiment Setup

Figure 2 shows the main steps in the experimental setup. We started with three main hypotheses (subsection 3.1), selected three datasets appropriate for testing our hypotheses (Section 3.2), processed all the datasets with the Sparv pipeline (Section 3.3), and manually checked the automatic annotation (Section 3.4). The choice of evaluation metrics and quantitative analysis of the results are given in Section 4, followed by a qualitative analysis in Section 5.

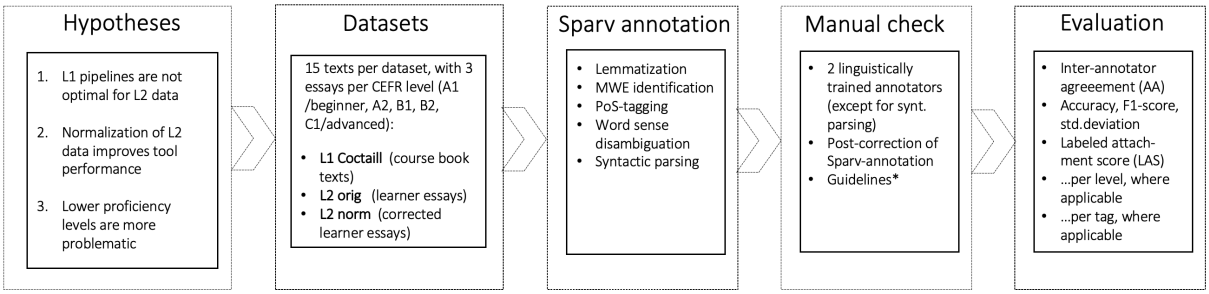

Figure 2: Overview of the experiment setup.
(*)Guidelines: https://tinyurl.com/bdhsukys

3.1 Hypotheses

As is obvious from the short description of the automatic linguistic annotation of learner language given in the Introduction, there is a need to explore the reliability of automatic pipelines further, to assess the needs to adapt the pipelines for L2, and to discuss the implications of the results for L2 theoretical studies and practical applications. Our hypotheses for this experiment are:

1. Pipelines trained on a standard language (L1) do not perform as well on non-standard language varieties such as learner language (e.g. L2 learner production).

2. Normalization of non-standard language, e.g. through error correction, improves tool performance.

3. The need for normalization (cf "correction") is especially critical for L2 texts written by learners at lower proficiency levels since they are likely to contain a higher level of misspellings, wrong words and syntactic discrepancies in comparison to the standard.

Even though some of the claims above appeal to common sense, they need to be confirmed explicitly and there is a need for an estimate of how well, or how poorly, the automatic annotation works in order to know how it can be reused for research, CALL and other scenarios.

3.2 Datasets

To address the hypotheses above, we selected 15 texts per language variety which we are interested in – namely, native language used in L2 Swedish course books (*L1 Coctaill*), L2 essays (*L2 orig*) and corrected L2 essays (*L2 norm*) – so that they represent five levels of proficiency with three texts per level for each dataset. The levels are defined in accordance with CEFR, the Common European Framework of Reference (Council of Europe, 2001), in our datasets covering five of the six levels: A1 (beginner), A2, B1, B2 and C1 (advanced). C2 was excluded due to a lack of data in the source corpora.

Care was taken to select texts of different genres and topics to avoid biases. Only texts containing at least one MWE according to the Sparv annotation were selected. Learner essays in the *L2 orig* dataset

[7] https://universaldependencies.org/u/dep/index.html

represent speakers of different native languages – namely: Chinese, English, Finnish, Flemish, Lithuanian, Macedonian, Persian, Romanian, Serbian, Somali, Spanish, Tigrinya, Vietnamese – to avoid potential influence of L1 on L2 usage. Detailed statistics over the datasets are available in Appendix C.

L1 Coctaill *L1 Coctaill* is a dataset representing native language and contains 2190 tokens (incl. punctuation) per 15 texts. These texts comprise various genres (narrations, facts, evaluation, dialogues, letters, poems) and different topical domains (traveling, languages, culture and traditions, relations with other people, etc.). The dataset is based on *COCTAILL* – a corpus of course books (Volodina et al., 2014), where each chapter has been marked with the level of proficiency at which it could be used in the teaching of L2 Swedish. The CEFR levels are represented by three texts per level.

L2 orig *L2 orig* is a dataset that contains 4012 tokens (incl. punctuation) per 15 essays, with three essays per CEFR level, covering several genres (narration, evaluation, argumentation, etc.) and topical domains (personal identification, daily life, travel, house and home, culture and traditions, etc.). *L2 orig* is a subset of the *SweLL-pilot* – a corpus of learner-written essays (Volodina et al., 2016) collected from three different schools/test bodies, and also marked with CEFR levels.

L2 norm *L2 norm* is a dataset containing 3955 tokens (incl. punctuation) per 15 essays and consists of the same essays (or in two cases of comparable essays, e.g. of the same topic and genre) as in *L2 orig*, but normalized for errors and deviations to reflect the current norms of the target language. The normalization was performed using the SVALA tool (Wirén et al., 2019), by a linguistically trained L1 Swedish speaker following the normalization guidelines from the SweLL-project (Rudebeck et al., 2021).

3.3 Sparv Pipeline

The Sparv pipeline[8] (Borin et al., 2016), consists of several modules, sequentially applied to the Swedish data input. In version 3.0, analyzed by us, for *lemmatization*, the Saldo lexicon (Borin et al., 2013) returns lemgrams including potential MWEs and a list of associated senses. *Senses* are disambiguated using an algorithm developed by Nieto Piña (2019) based on Saldo senses. For *POS tagging*, Sparv uses HunPos (Halácsy et al., 2007) trained on the SUC 3.0 corpus (Ejerhed et al., 1997). For *syntactic annotation*, the MaltParser (Nivre et al., 2007) is used, trained on the Swedish Talbanken (Nilsson et al., 2005).

A new Sparv version (4.0) was released for public use in 2021,[9] where the POS tagger and syntactic annotation are changed to Stanza (Qi et al., 2020) with new models. According to Berdicevskis (2020a) and Berdicevskis (2020b), annotation for syntactic relations and POS tagging in versions 4.0 and above should have a higher accuracy than previous versions. The newer versions of Sparv continue using models trained on SUC and Talbanken, which means that the tagsets for both POS and DepRels are still the same.

3.4 Manual Check

Two linguistically trained assistants, one an L1 Swedish speaker and one an advanced L2 Swedish speaker (L1 Finnish), manually analyzed the automatic tags of the three datasets, introducing corrections where necessary. Assistants were equiped with guidelines[10] and were in regular contact with one of the researchers for discussions, which cleared up uncertainties and led to clarifications in the guidelines. They performed the check using separate spreadsheet files to avoid influencing each other. Instructions were specific for each linguistic feature.

The rule of thumb for the annotation check was to start from a *positive assumption* that the Sparv-pipeline's suggestions are correct, and introduce corrections only if necessary and motivated. With regards to annotation of learner essays, it meant *disregarding* the perspective of "what the learner meant" and assessing the output of the pipeline from a formal point of view, i.e. what it had been fed.

Most problems arose from the conceptual interpretation of the task in relation to the *L2 orig* dataset, namely, what to consider correct or incorrect output from the pipeline. Consider the following example:

[8] History of Sparv-releases: `https://github.com/spraakbanken/sparv-pipeline/releases`

[9] `https://github.com/spraakbanken/sparv-pipeline/releases/tag/v4.0.0`

[10] `https://docs.google.com/document/d/1W9gcwRwFJ7-DsAC6cf6BHUoEivt73r-XWCV1oKS6xV8/edit?ts=5f3518d7#`

[1] `Jag tycker om spela fotbol , sinima , cykler och TV-speL .` (A1 level)
'I like to play footbal , ?swinim / ?sinima , bykes[11] and TV-gameS .' (translation tries to replicate the errors in the original variant. `sinima` may be an attempt to write *simma* 'to swim', but since the learner lists hobbies it could also be an attempt to write the English word 'cinema' in a more Swedish way instead of the corresponding Swedish word *bio*.)

The word `cykler`[12] 'bikes' could be interpreted as either the present tense of the verb "to bike"[13], *cyklar*, or the plural form of the noun "a bike", *cyklar*. The use of `TV-speL` 'TV-gameS' suggests that a noun is a possible alternative in a list together with the noun `TV-speL`. However, a verb is also a fully legitimate alternative, as a part of a list together with the verb `spela (*fotbal)`, 'play (football)'. The assistants have annotated this output differently – one correcting the Sparv-suggested *noun*-tag for `cykler` with a *verb*-tag, the other accepting the *noun*-suggestion as the right one. Similarly, the misspelled word `sinima` was automatically tagged as a noun, and accepted as such by one of the assistants, but changed to a verb by the other. These examples show the problems of dealing with learner data and potential reasons for disagreements between the annotators. Both interpretations above are equally possible and equally close to the original.

The example below is easier to interpret and does not cause disagreement between the annotators. One learner produced a misspelling of the preposition *enligt* 'according to' which was tagged as an *adjective*, most probably due to its morphological form in conjunction with the position in the sentence:

[2] `Engligt ungmedia.se, är...` (C1 level)
'Accordin to ungmedia.se, is...' (translation tries to preserve the errors in the original variant.)

Both annotators corrected Sparv-suggested tag *adjective* to *preposition*. In standard Swedish, the first position of a sentence is most likely to contain a subject, often consisting of a noun phrase which can contain an adjective. However, to a human annotator, the similarity of `engligt` to the word `enligt` was obvious; it was also obvious that there is no adjective that is similar to this. This motivated the correction to the pipeline's output and suggests a need to check for lexical similarity in POS-tagging.

4 Results

On completion of the check, we analyzed the number of deviations discovered during the manual check and inspected their nature per linguistic category in each dataset and in relation to the proficiency level and tagset, where appropriate. Below, we report these results using precision, F1-score and LAS measures (averaged over the two annotators for all tasks except syntactic parsing/DepRel annotation which was checked by only one annotator). For word sense disambiguation, we have additionally computed a baseline using the first sense in all cases.

Inter-Annotator Agreement To put the reported results into perspective, we calculated inter-annotator agreement (IAA) for the two annotators using Krippendorff's alpha (Krippendorff, 2004) for MWEs, and pairwise agreement for Lemma, POS and Sense, see Table 1. Pairwise agreement is calculated on a token basis, and we count only whether a change has been made to the original annotation or not.

Corpus	Lemma	POS	Sense	MWE
L1 Coctaill	0.95	0.97	0.88	0.85
L2 orig	0.94	0.95	0.88	0.74
L2 norm	0.95	0.97	0.90	0.89

Table 1: Pairwise agreement for Lemma, POS and Sense; Krippendorff's alpha for MWE

The agreement lies over 0.8 for most of the datasets and denotes high agreeement. We see that values for *L2 orig* is nearly always lower than for the other datasets; reasons for that have been briefly touched

[11] since the original *cykler* is a misspeling, we mock a misspelling in the English version of the word *bike*
[12] Note that *cykler* is a misspelling, too.
[13] While "to cycle" might be a more idiomatic translation, we want to illustrate the homonymy between word classes here

upon in Section 3.4. Most disagreements appear in the evaluation of the MWE identification, with the lowest at 0.74 for *L2 orig*. The intersection of corrections introduced by both annotators is high. Still we see that one annotator is better at noticing grammatical MWEs (e.g. `trots att` 'even though' and the other is better at spotting light verb constructions (e.g. `få barn` 'have a child/children') and this causes disagreement, but enriches the results of the check.

Corpus	Lemma	POS	DepRel
L1 Coctaill	**0.93 (0.0)**	**0.98 (0.0)**	**74.49**
A1	0.96 (0.02)	0.98 (0.0)	75.93
A2	0.98 (0.03)	0.97 (0.02)	72.51
B1	0.94 (0.0)	0.97 (0.0)	76.65
B2	0.89 (0.0)	0.97 (0.01)	71.05
C1	0.92 (0.01)	0.97 (0.0)	76.31
L2 orig	**0.90 (0.02)**	**0.95 (0.0)**	**63.01**
A1	0.89 (0.01)	0.92 (0.0)	51.66
A2	0.89 (0.0)	0.94 (0.0)	57.18
B1	0.91 (0.02)	0.96 (0.01)	60.42
B2	0.92 (0.04)	0.97 (0.01)	67.53
C1	0.92 (0.03)	0.97 (0.01)	69.18
L2 norm	**0.93 (0.02)**	**0.97 (0.0)**	**69.02**
A1	0.95 (0.0)	0.98 (0.01)	67.23
A2	0.92 (0.0)	0.96 (0.0)	69.30
B1	0.95 (0.02)	0.98 (0.01)	70.53
B2	0.92 (0.03)	0.98 (0.01)	71.52
C1	0.92 (0.02)	0.97 (0.0)	66.80

Table 2: Lemmatization and POS tagging: precision and standard deviation; Dependency: LAS

4.1 Automatic Lemmatization, POS-tagging and Dependency Annotation

Table 2 summarizes the results regarding the quality of the automatic annotation (i.e. how often the two annotators corrected automatically assigned tags) for lemmatization, POS tagging and Dependency Relations. Lemmatization and POS-tagging are evaluated in terms of precision (number of correct items by total number of items), averaged over the two annotators. Dependency annotation is evaluated using micro-averaged (i.e. token-based) Labeled Attachment Score (LAS) (Kübler et al., 2009).[14]

Automatic Lemmatization Results for automatic lemmatization show that it is very successful, with 93% precision on average for *L1 Coctaill*. As expected, the number decreases in *L2 orig* in comparison to *L1 Coctaill*, resulting in 90% precision; and after normalization it increases to 93% in *L2 norm*, the same level as in *L1 Coctaill*. We also see the expected tendency of quality increase in the *L2 orig* by proficiency level. As learners become more proficient they write in a way that can be expected to be closer to L1, a language containing less discrepancies and hence easier to annotate automatically with tools trained on L1 data. The fact that we do not see the same increase in *L1 Coctaill* is probably due to the fact that language presented as reading materials to learners at more advanced levels can contain more specialized vocabulary, some of which might not be in Saldo. It is interesting that similarly we also do not see an increase over all levels in the *L2 norm*. But this correlates with the L1 data and it is notable that C1-level in this data is as well lemmatized as the L1 data.

Part-of-Speech Tagging Results for POS-tagging are systematically high across all datasets, with the average top 98% for *L1 Coctaill* and the lowest average result of 95% in *L2 orig*. Normalization of learner

[14]Note that the dependency annotation was checked by one assistant, while the rest of the annotation was checked by two.

essays improves the results for POS by 2 points. Just like for lemmatization there is a clear improvement in the POS-tagging on higher levels in the *L2 orig*, which reaches as high precision as the L1 data on B2 and C1-levels. The *L2 norm* has as high precision as the L1 data had at its best, with 98% precision on A1, B1, B2. However the precision drops on A2 to 96% and also on C1 where it is on the same level as L1 and L2 orig, 97%.

Dependency Annotation Our results show that dependency annotation is less reliable even for L1, with a preserved tendency of quality loss on *L2 orig* as in the lemmatization and POS-annotation. In this case, however, the performance drops by 11 points, from 74.5% to 63%. Normalization improves performance of the Sparv-tool by 6 points, from 63% to 69%. Level of proficiency seems to have a direct effect on the improvement of annotation of *L2 orig*, and for dependency relations also of *L2 norm* except for C1 level. The results for *L1 Coctaill* are in line with previous results reporting a LAS score of 78.39 on L1 text (Berdicevskis, 2020a) in automatic evaluation of dependency-relation annotation with Sparv (v.3.0).

We see that our general assumptions are confirmed: the performance of the automatic annotation on learner essays (*L2 orig*) has lower accuracy than on native (*L1 Coctaill*) or normalized (cf corrected) (*L2 norm*) texts, even though only marginally for lemmatization and POS tagging. This echoes the results obtained in the automatic evaluation of the Sparv POS-tagging on in-domain L1 texts versus out-of-domain Internet texts (accuracy 0.98 vs 0.93) (Berdicevskis, 2020b) and partially for dependency annotation (Berdicevskis, 2020a). While dependency relation is only moderate in quality, the automatic lemmatization and POS tagging are reliable enough to base further generalizations about L2 development.

4.2 Automatic Detection of MWEs

The purpose of the MWE check in our experiment was to find out whether MWEs: (1) were correctly identified; (2) failed to be identified; (3) were incompletely identified; or (4) were incorrectly identified in the different datasets. Table 3 shows precision, recall and F1 score per resource, as well as a breakdown over the different CEFR levels. These values are calculated relative to the number of automatically and manually identified MWEs ($\approx$ the total correct number of MWEs) and not on a token basis. Numbers are averaged over the two annotators, with standard deviation indicated in parentheses.

F1-scores in Table 3 follow the same tendency as the features described earlier: the pipeline performs best on *L1 Coctaill*, the performance drops on *L2 orig* (in this case by 12 points), and improves on *L2 norm* (by 4 points). We cannot see any clear tendency across proficiency levels, the increases and decreases seem to be idiosyncratic and depend on other factors than levels of proficiency, e.g. text genres, topic or task types. Still the results in Table 3 indicate that we can expect that out of 10 MWEs, 7–8 are correctly captured, 2–3 are missed and a small percentage of noise is introduced in the form of suggestions of MWEs that are not actually in the text or that are incomplete MWEs. In nine of the missed cases (45%) an MWE entry is also missing in the Saldo lexicon. However, there are also cases where the MWEs did exist in Saldo but were still missed. All in all, results of this evaluation suggest that we can trust the automatic MWE identification, even though we need to be aware of possible misses.

4.3 Automatic Word Sense Disambiguation (WSD)

The goal of this check was to find out how often: (1) sense was correctly identified; (2) no sense was assigned at all; (3) a lemgram for the correct sense was missing in Saldo; and (4) the correct sense was missing in Saldo. Table 4 shows the results of the WSD annotation checks. In all three datasets the accuracy of WSD is high, with very slight fluctuations between the datasets. Counter to our expectations, we do not see any radical improvement in performance following normalization of L2 data, nor is there any distinct tendency for poorer WSD quality on the lower proficiency levels. The check shows that some senses are missing in Saldo; sometimes even lemgrams are missing. Most challenging are function words, like *som, mången, än* 'as, much, yet', that have very few (sense-based) entries in Saldo, and often in combination with a POS that does not match POS tagging based on SUC. For example, for the word `som` 'as, like', the SUC taxonomy used in Sparv contains two POS - *conjunction (KN)* and *relative pronoun (HP)*, whereas in Saldo, `som` is listed as *subjunction (SN)* and *adverb (AB)*, leaving no overlap

Resource	Identified	Correct	Partial	Incorrect	Missed	Precision	Recall	F1
L1 Coctaill	59	**50.5 (1.5)**	**4.0 (1.0)**	**4.5 (0.5)**	**13.5 (3.5)**	**0.85 (0.02)**	**0.79 (0.04)**	**0.82 (0.03)**
A1	8	6.5 (1.5)	1.0 (1.0)	0.5 (0.5)	1.0 (1.0)	0.81 (0.18)	0.85 (0.14)	0.83 (0.16)
A2	7	6.0 (0.0)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.85 (0.0)	0.85 (0.0)	0.85 (0.0)
B1	18	16.5 (0.5)	0.0 (0.0)	1.5 (0.5)	4.5 (0.5)	0.91 (0.02)	0.78 (0.02)	0.84 (0.02)
B2	17	14.0 (1.0)	2.0 (0.0)	1.0 (1.0)	6.5 (1.5)	0.82 (0.05)	0.68 (0.03)	0.74 (0.00)
C1	9	7.5 (0.5)	1.0 (0.0)	0.5 (0.5)	0.5 (0.5)	0.83 (0.05)	0.93 (0.06)	0.88 (0.05)
L2 orig	81	**56.0 (1.0)**	**1.0 (0.0)**	**24.0 (1.0)**	**21.0 (2.0)**	**0.69 (0.01)**	**0.72 (0.01)**	**0.70 (0.00)**
A1	5	5.0 (0.0)	0.0 (0.0)	0.0 (0.0)	4.5 (1.5)	1.00 (0.0)	0.53 (0.08)	0.69 (0.07)
A2	4	3.0 (0.0)	0.0 (0.0)	1.0 (0.0)	3.5 (0.5)	0.75 (0.0)	0.46 (0.03)	0.57 (0.02)
B1	15	5.0 (0.0)	0.0 (0.0)	10.0 (0.0)	5.5 (0.5)	0.33 (0.0)	0.47 (0.02)	0.39 (0.00)
B2	22	15.0 (0.0)	0.0 (0.0)	7.0 (0.0)	3.5 (0.5)	0.68 (0.0)	0.81 (0.02)	0.74 (0.00)
C1	35	28.0 (1.0)	1.0 (0.0)	6.0 (1.0)	4.0 (1.0)	0.80 (0.02)	0.87 (0.02)	0.83 (0.00)
L2 norm	98	**68.5 (1.5)**	**3.0 (2.0)**	**26.5 (0.5)**	**17.5 (0.5)**	**0.69 (0.01)**	**0.79 (0.00)**	**0.74 (0.01)**
A1	8	6.5 (0.5)	0.5 (0.5)	1.0 (0.0)	4.5 (0.5)	0.81 (0.06)	0.59 (0.04)	0.68 (0.05)
A2	8	6.0 (0.0)	0.0 (0.0)	2.0 (0.0)	3.5 (0.5)	0.75 (0.0)	0.63 (0.03)	0.68 (0.01)
B1	23	13.5 (0.5)	1.5 (1.5)	8.0 (2.0)	3.5 (1.5)	0.58 (0.02)	0.79 (0.07)	0.67 (0.04)
B2	33	21.0 (1.0)	0.0 (0.0)	12.0 (1.0)	2.0 (0.0)	0.63 (0.03)	0.91 (0.00)	0.74 (0.02)
C1	26	21.5 (0.5)	1.0 (0.0)	3.5 (0.5)	4.0 (1.0)	0.82 (0.01)	0.84 (0.03)	0.83 (0.02)

Table 3: Number of correctly identified MWEs including precision, recall and F1 score: Averages (and standard deviations)

between the two resources. Besides, the sense inventory for such words is too limited in Saldo to cover all the possible contexts where they are used, which we can see in the fact that Svensk ordbok (SO) lists six senses for `som`.

Checking the quality of the automatic WSD on our three datasets has shown that we can expect that in 80–90 percent of the cases the word sense is correctly assigned. Despite the fact that the WSD in Sparv is not bullet-proof, we consider it reliable enough to build our vocabulary resource (L2 lexical profile) on the sense level using lemma+POS+sense as our main entry.

For WSD, a frequently used baseline is the *most frequent sense* baseline which assigns the most frequent sense observed in the training data to each word (Mihalcea, 2007, p. 123). Saldo senses are not ordered by frequency (Borin et al., 2013), thus such a baseline is difficult – albeit not impossible – to calculate (before calculating frequencies, one would need to clarify and justify which corpora to use, etc.). Sense distinctions in Saldo simply indicate that there is a *difference* in sense (Borin et al., 2013). We therefore calculate a simplified version of the *most frequent sense* baseline – the *first sense* baseline – which assigns each word the *first* sense in Saldo. Table 5 shows the number of correct word senses according to the baseline calculation, in comparison with the annotations by annotators 1 and 2, the total number of tokens per dataset, and the mean accuracy and standard deviation. With an average accuracy of about 75%, this baseline is clearly outperformed by the WSD in Sparv. The results for WSD by Sparv are, thus, very encouraging.

5 Qualitative Analysis

In this section we take a closer look at the POS-annotation and the dependency relations in the three datasets. Grammatical annotation such as this can prove very useful both in research and applications in relation to L2 acquisition, but we need to know exactly which tags that are reliable enough.

5.1 Qualitative Analysis of the POS Check

Most of the POS have a precision between 1–0.9 in most of the three datasets and according to both annotators, hence POS-tagging generally provides a very good basis for both research and applications

	# tokens excl punct	Correct sense	Incorrect sense	No sense	Lemgram missing in Saldo	Sense missing in Saldo	Accuracy (correct/ total)
L1 Coctaill	**1900**	**1619.5 (66.5)**	**192.0 (61.0)**	**46.5 (3.5)**	**25.0 (1.0)**	**17.0 (1.0)**	**0.85 (0.03)**
A1	434	399.5 (5.5)	26.5 (4.5)	2.0 (0.0)	5.0 (2.0)	1.0 (1.0)	0.92 (0.01)
A2	101	84.0 (7.0)	15.5 (6.5)	0.5 (0.5)	1.0 (1.0)	0.0 (0.0)	0.83 (0.07)
B1	554	466.0 (22.0)	58.5 (21.5)	14.0 (2.0)	6.5 (1.5)	9.0 (0.0)	0.84 (0.04)
B2	488	409.0 (15.0)	47.5 (13.5)	18.5 (1.5)	8.0 (0.0)	5.0 (0.0)	0.84 (0.03)
C1	324	262.0 (17.0)	44.0 (15.0)	11.5 (0.5)	4.5 (0.5)	2.0 (2.0)	0.81 (0.05)
L2 orig	**3635**	**3000.0 (178.0)**	**326.5 (163.5)**	**201.5 (13.5)**	**25.5 (3.5)**	**81.5 (2.5)**	**0.83 (0.05)**
A1	301	243.5 (11.5)	33.0 (10.0)	22.5 (2.5)	1.0 (0.0)	1.0 (1.0)	0.81 (0.04)
A2	481	405.5 (14.5)	39.5 (13.5)	27.5 (0.5)	4.0 (1.0)	4.5 (0.5)	0.84 (0.03)
B1	814	657.0 (51.0)	78.5 (42.5)	68.0 (7.0)	5.0 (0.0)	5.5 (1.5)	0.81 (0.06)
B2	886	737.5 (39.5)	76.5 (36.5)	45.0 (1.0)	4.5 (2.5)	22.5 (0.5)	0.83 (0.04)
C1	1153	956.5 (61.5)	99.0 (61.0)	38.5 (2.5)	11.0 (0.0)	48.0 (2.0)	0.83 (0.05)
L2 norm	**3565**	**2963.5 (109.5)**	**372.5 (108.5)**	**123.5 (1.5)**	**30.5 (2.5)**	**75.0 (3.0)**	**0.83 (0.03)**
A1	323	271.5 (7.5)	40.5 (8.5)	6.0 (0.0)	1.0 (0.0)	4.0 (1.0)	0.84 (0.02)
A2	499	426.0 (5.0)	45.0 (6.0)	15.0 (0.0)	7.5 (0.5)	5.5 (0.5)	0.85 (0.01)
B1	852	718.5 (36.5)	92.0 (33.0)	23.5 (2.5)	6.5 (0.5)	11.5 (0.5)	0.84 (0.04)
B2	1159	966.5 (32.5)	107.0 (31.0)	54.0 (1.0)	7.5 (2.5)	24.0 (0.0)	0.83 (0.03)
C1	732	581.0 (28.0)	88.0 (30.0)	25.0 (0.0)	8.0 (0.0)	30.0 (2.0)	0.79 (0.04)

Table 4: Overview of the automatic sense annotation in the three datasets: Averaged counts (and standard deviation)

Resource	Correct (Annotator 1)	Correct (Annotator 2)	Total	Accuracy (std)
L1 COCTAILL	1469	1401	1900	75.52 (1.79)
L2 orig	2800	2588	3635	76.42 (2.34)
L2 norm	2808	2641	3565	74.11 (2.92)

Table 5: WSD first sense baseline

even when based on learner data. *Participles (PC)* have low precision according to both annotators in *L2 orig*, and this is the only time both annotators are in clear agreement that the pipeline is wrong (precision 0.5 and 0.38). However, only eight tokens have been annotated with PC in this dataset so the figures are hardly reliable. Still we know that participles have been problematic in several ways. They can be lemmatized as verbs, adjectives or as their own POS (participles). Both in Saldo (Borin et al., 2013) and in SAG (Teleman et al., 1999) they are treated as an individual POS. A complicating fact, though, is the ability of Swedish past participles to agree with their noun phrase antecedents, which makes their behavior similar to adjectives, e.g. plural Stolarna är **täckta** med snö 'The chairs are **covered** in snow' vs singular Bordet är **täckt** med snö 'The table is **covered** in snow'. Note also that many adjectives in Swedish are historically derived from participles, e.g. *nöjd* 'content, happy' from the verb *nöja sig* 'be content with'. All these factors combined make distinguishing participles from verbs and adjectives complicated, especially in learner language.

Other POS with low precision by one annotator can have moderate to excellent precision from the other annotator. This is because there are very few tokens which have been tagged with some POS, e.g. *Interjection* – 2 items in L2 norm (precision 1 and 0.5) or *Ordinal number* – 5 items in L1 Coctaill (1 and 0.4). This particular case clearly shows that Saldo can contribute to disagreement between annotators. The two ordinals annotated here, första 'first' and tredje 'third', are adjectival lemmas in Saldo. Other ordinals such as fjärde 'fourth' appear twice in Saldo, once as an adjectival lemma but also as a form in the morphological paradigm for fyra 'four'. Comparing the datasets we see that första, tredje received no lemma automatically. Both annotators inserted lemmas according to Saldo, but only one of them adjusted the POS-tag from *RO* to *JJ* in agreement with Saldo. When the token is the ordinal fjärde this is lemmatized as fyra and neither annotator corrects this in *L2 orig*, but in *L2 norm* it is corrected by one to the lemma fjärde, but the POS-tag *RO* is left untouched. Disagreements like these

are not errors and can only be avoided by specifying how to treat these in the guidelines. However, even the researchers who are used to working with Saldo were not aware that this was a difference that existed in Saldo and hence could not take it into account in writing the guidelines. Instead, the check has helped to spotlight an inconsistency in Saldo that should be taken into consideration for future developments, but which may be unavoidable to some extent due to differences in how different the ordinal forms are in relation to the cardinal numbers. Since this also appears to bear a direct affect on the success of lemmatization this is clearly of importance to the performance of the pipeline.

Twenty-three tokens in *L2 orig* have been tagged as *particles*, but the precision differs between 0.96 and 0.43. This appears to be related to the definition of particles. They can be seen as a POS of their own, or as e.g. *adverbs* or *prepositions*; and *particle (adverbial)* is sometimes instead seen as a syntactic function. This is the way SAG views them. It seems one annotator followed SAG more closely and this caused disagreement. IAA could here have been improved by a stricter guideline with regards to how to treat *particles*.

Interestingly, *adjectives* also have quite low precision according to one annotator in both *L2 orig* and *L2 norm*. Most cases (61.6%) have been corrected to *determiner*, a category which this annotator seems to be more familiar with than the other annotator and a category which is not normally included in Swedish grammar, nor is it a POS category in SAG (Teleman et al., 1999). Half of the items are the word `många` 'many' which is classed as a pronoun by *Svensk ordbok* and also by Saldo, but which is normally used as a prenominal modifier for quantity and hence could according to some theories be classed as *determiner*. However in SUC 3.0 `många` is annotated as *adjective* (76%) or *pronoun*. *Determiner* is used for similar words like `några` 'some' or `alla` 'all'. This is a clear example where it is hard to decide when the pipeline should be considered correct.

To summarize, IAA is easily severely damaged if there are few items that are being evaluated. Lexical items with clear morphological paradigms with many different forms are easier to classify by POS. But lexical items with morphological paradigms which are hardly used for agreement (e.g. `mången – många`) and which in comparison show suppletive forms which can be interpreted as independent lemmas (e.g. `mången, många – flera, flest`) the morphological paradigms cause problems for the annotation.

5.2 Qualitative Analysis of the Check of Dependency Relations

Dependencies can be problematic for linguists because – even though they may be familiar with main categories (e.g. subject, object, finite verb) – checking the dependency annotation entails understanding of what should be seen as correct according to that particular dependency grammar (which in our case includes 65 tags). Since the dependency parser has been trained on Talbanken our annotator was instructed to consult the annotations in Talbanken for comparison when uncertain. In addition, she discussed complicated cases in detail with one of the researchers. Another complicating factor turned out to be that L1 data included some lyrics and poems which were difficult for the parser since sentences were not marked as usual. Similar problems can often be seen in L2 language at low proficiency levels. Unfortunately, few of the dependency labels have a precision above 0.9, only eight in *L1 Coctaill*, four in *L2 orig* and four in *L2 norm*. In addition, several labels have been assigned to very few tokens and hence the accuracy is not really reliable as shown above. There are only three categories with a precision between 0.9–0.95 and 39 tokens or more.

It is only *Infinitive Verb phrase minus infinitive marker* (IF) and *Negation adverbial* (NA) that have a precision of 0.9 or more in all the three data sets. In *L1 Coctaill* this is based on very little data, but in L2 datasets it is based on 50–63 tokens which is reassuring. Unfortunately, these particular dependency labels do not give that much additional power to L2 research or applications since they are highly correlated with specific words or morphological forms, the negation `inte` 'not' and the infinitive. The correct *NA*-labels are always correlated with the lemma `inte`. Out of all *NA* in Talbanken 697/742 = 94% are attached to `inte`. And out of all the `inte` in Talbanken 697/720 = 97 % are *NA*. Of course looking at the actual dependency tree it is of interest to see that this dependency relation is related to the correct nodes in the tree since this can affect the semantic interpretation of the sentence.

Nominal adjectival pre-modifiers (AT) are rather well annotated in all datasets. In *L2 orig* the precision is at its lowest at 0.85 (based on 100 tokens) and increases to 0.92 (based on 83) in *L2 norm*. Neither as high as the L1 data, precision 0.95, but the L1 data is based on only 39 tokens and could therefore be seen as less certain. *AT* are interesting for L2 acquisition in relation to both agreement and definiteness. Hence these are important to capture for assessment purposes, CALL and research. Moreover, being able to use extended noun phrases with adjectival premodifiers can be seen as a first step to increased proficiency even if the forms are incorrect.

In comparison to the pre-nominal adjectival modifiers it would be interesting to also be able to catch predicative complements well since they also show agreement to some extent and this type of agreement is more difficult to a learner according to Pienemann's processability theory (Pienemann and Håkansson, 1999) since it crosses phrase boundaries. *Predicative complements* have somewhat lower precision. It is moderate for *subjective predicative complements (SP)* and there is a clear improvement from *L2 orig* to *L2 norm*, but interestingly it does not quite reach L1 precision.

Finally, one last dependency which receives reasonably good scores and is also based on a fair number of tokens is *determiner (DT)*, 0.83 (*L1 Coctaill*, 196), 0.87 (*L2 orig*, 357) and 0.92 (*L2 norm*, 339). It is interesting that here L1 has the lowest precision and we see a clear improvement from *L2 orig* to *L2 norm*. *DT* has been attached to tokens which vary quite a lot. Their POS-tags include: conjunctions (KN), determiners (DT), adjectives (JJ), nouns (NN).

6 Conclusions

To summarize, we have seen than lemmatization, POS-tagging and word sense disambiguation are the least sensitive to being applied to non-native data instead of L1. Most affected are dependency annotation and identification of multi-word expressions. All of the annotation steps perform better when applied to normalized learner data instead of the original.

Comparing a non-standard text to a standard text is complicated, and such an evaluation is affected by the type of texts which are used in the evaluation, including the levels of text complexity and the proficiency levels of the essay writers. One complication in evaluating learner texts is that mistakes can be on many different levels. A word might have been used in the wrong context but annotated correctly based on the morphological principles, disregarding semantic and syntactic principles.

Despite the challenges and varying results per linguistic features, we find that our hypotheses have been generally confirmed:

1. Pipelines trained on standard language do not perform equally well on non-standard deviating language. The performance drop varies between different linguistic features, and in certain cases it is relatively negligible (e.g lemmatization and POS tagging).

2. We have shown that normalization of the learner language improves the performance of the automatic pipeline for all linguistic features, but sometimes only marginally.

3. Proficiency levels have no systematic influence on the performancee of the automatic pipeline, apart from in *L2 orig* where there are improvements for each of the linguistic features with growing proficiency levels. This may be due to the fact that automatic pipelines are more sensitive to incorrect language typical of L2 original data than to length of the sentences, lexical and syntactic complexity in normlike written texts of different genres and levels.

All in all, the results of our evaluation are very encouraging, especially with regards to lemmatization, POS tagging, and word sense disambiguation. MWE identification seems to be a cognitively more challenging task. Further, we have strong indications that automatic dependency relation annotation is relatively unreliable with the exception of the labels IF, NA, AT, DT, SS and ROOT, and to some extent FS if we disregard the low precision in L1 data because it is based on so few instances. We should therefore be selective in which categories we use for theoretical geeneralizations and practical implementations. However, the new version of the Sparv pipeline may perform reliably enough for our purposes for all categories.

Acknowledgements

This work has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *Development of lexical and grammatical competences in immigrant Swedish*, P17-0716:1, and by *Nationella språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We also wish to thank the anonymous reviewers for their valuable comments on a previous version.

References

Alfter D., Lindström Tiedemann T., and Volodina E. 2021. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. *Northern European Journal of Language Technology*.

Berdicevskis A. 2020a. Choosing a new dependency parser for Sparv. Technical report, University of Gothenburg, Department of Swedish, 2020-06-03.

Berdicevskis A. 2020b. Choosing a new POS-tagger for Sparv: Update. Technical report, University of Gothenburg, Department of Swedish, 2020-05-12.

Bhalla V. and Klimcikova K. 2019. Evaluation of automatic collocation extraction methods for language learning. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*.

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

Borin L., Forsberg M., Hammarstedt M., Rosén D., Schäfer R., and Schumacher A. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *Proceedings of Swedish Language Technology Cconference (SLTC)*. Umeå University.

Borin L. 2021. Multiword expressions–a tough typological nut for swedish framenet++1. In Dannells D., Borin L., and Friberg Heppin K., editors, *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, volume 14, pages 221–262. John Benjamins Publishing Company.

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

De Marneffe M.-C., Dozat T., Silveira N., Haverinen K., Ginter F., Nivre J., and Manning C. D. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Ejerhed E., Källgren G., and Brodda B. 1997. Stockholm Umeå Corpus version 1.0, SUC 1.0. *Department of Linguistics, Umeå University*.

Gardner D. 2007. Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied linguistics*, 28(2):241–265.

Gustafson-Capková S. and Hartmann B. 2006. Manual of the stockholm umeå corpus version 2.0. *Unpublished Work*.

Halácsy P., Kornai A., and Oravecz C. 2007. HunPos-an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, United States. Association for Computational Linguistics.

Krippendorff K. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Krivanek J. and Meurers D. 2013. Comparing rule-based and data-driven dependency parsing of learner language. *Computational dependency theory*, 258:207.

Kübler S., McDonald R., and Nivre J. 2009. Dependency parsing. *Synthesis lectures on human language technologies*, 1(1):1–127.

Ljunglöf P., Zechner N., Nieto Piña L., Adesam Y., and Borin L. 2019. Assessing the quality of Språkbanken's annotations. Technical report, University of Gothenburg, Department of Swedish.

Meurers D. and Wunsch H. 2010. Linguistically annotated learner corpora: Aspects of a layered linguistic encoding and standardized representation. *Proceedings of Linguistic Evidence*, pages 1–4.

Mihalcea R. 2007. Knowledge-based methods for WSD. In *Word sense disambiguation*, pages 107–131. Springer.

Nieto Piña L. 2019. *Splitting rocks: Learning word sense representations from corpora and lexica*. PhD Thesis, Data Linguistica 30.

Nilsson J., Hall J., and Nivre J. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. *Proceedings from the special session on treebanks at NoDaLiDa 2005*, pages 119–132.

Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., and Marsi E. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95.

Nivre J., Megyesi B., Gustafson-Capková S., Salomonsson F., and Dahlqvist B. 2008. Cultivating a Swedish treebank. *Resourceful language technology: Festschrift in honor of Anna Sågvall Hein*, pages 111–120.

Ott N. and Ziai R. 2010. Evaluating dependency parsing performance on german learner language. In *Proceedings of the ninth international workshop on treebanks and linguistic theories*, volume 9, pages 175–186. NEALT Tartu.

Paquot M. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1):121–145.

Petrov S., Das D., and McDonald R. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Pienemann M. and Håkansson G. 1999. A unifiedapproach toward the development of swedish as l2: A processabilityaccount. *Studies in second language acquisition*, 21(3):383–420.

Qi P., Zhang Y., Zhang Y., Bolton J., and Manning C. D. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Rubin R. 2021. Assessing the impact of automatic dependency annotation on the measurement of phraseological complexity in l2 dutch. *International Journal of Learner Corpus Research*, 7(1):131–162.

Rudebeck L., Sundberg G., and Wirén M. 2021. SweLL normalization guidelines. Technical report, GU-ISS Research report series, Department of Swedish, University of Gothenburg. http://hdl.handle.net/2077/69432.

Sag I. A., Baldwin T., Bond F., Copestake A., and Flickinger D. 2002. Multiword expressions: A pain in the neck for nlp. In *Conference on intelligent text processing and computational linguistics*. Springer.

Štindlová B., Rosen A., Hana J., and Škodová S. 2012. CzeSL–an error tagged corpus of Czech as a second language. In *Corpus data across languages and disciplines*. Peter Lang.

Teleman U., Hellberg S., Andersson E., et al. 1999. Svenska akademiens grammatik. *Arkiv*, page 233.

Volodina E., Pilán I., Eide S. R., and Heidarsson H. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*. Linköping University Press.

Volodina E., Pilán I., Enström I., Llozhi L., Lundkvist P., Sundberg G., and Sandell M. 2016. 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.

Wirén M., Matsson A., Rosén D., and Volodina E. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *Proceedings of CLARIN 2018*.

Wray A. 2005. *Formulaic language and the lexicon*. Cambridge University Press.

A Appendix: POS taxonomy

Kod 'Code' *Ordklass* 'Word class'	**Svensk term** 'Swedish term'	**Engelsk term** 'English term'
AB	Adverb	Adverb
DT	Determinerare, bestämningsord	Determiner
HA	Frågande/relativt adverb	Interrogative/Relative Adverb
HD	Frågande/relativ bestämning	Interrogative/Relative Determiner
HP	Frågande/relativt pronomen	Interrogative/Relative Pronoun
HS	Frågande/relativt possessivuttryck	Interrogative/Relative Possessive
IE	Infinitivmärke	Infinitive Marker
IN	Interjektion	Interjection
JJ	Adjektiv	Adjective
KN	Konjunktion	Conjunction
NN	Substantiv	Noun
PC	Particip	Participle
PL	Partikel	Particle
PM	Egennamn	Proper Noun
PN	Pronomen	Pronoun
PP	Preposition	Preposition
PS	Possessivuttryck	Possessive
RG	Räkneord: grundtal	Cardinal Number
RO	Räkneord: ordningstal	Ordinal Number
SN	Subjunktion	Subjunction
UO	Utländskt ord	Forein Word
VB	Verb	Verb

Table 6: SUC-based part of speech categories (POS code, Swedish term, English term).

B Appendix: DepRel taxonomy

MAMBA Categories			
Tag	Meaning	Tag	Meaning
++	Coordinating conjunction	JR	Second parenthesis
+A	Conjunctional adverbial	JT	Second dash
+F	Coordination at main clause level	KA	Comparative adverbial
AA	Other adverbial	MA	Attitude adverbial
AG	Agent	MS	Macrosyntagm
AN	Apposition	NA	Negation adverbial
AT	Nominal (adjectival) pre-modifier	OA	Object adverbial
CA	Contrastive adverbial	OO	Direct object
DB	Doubled function	OP	Object predicative
DT	Determiner	PL	Verb particle
EF	Relative clause in cleft	PR	Preposition
EO	Logical object	PT	Predicative attribute
ES	Logical subject	RA	Place adverbial
ET	Other nominal post-modifier	SP	Subjective predicative complement
FO	Dummy object	SS	Other subject
FP	Free subjective predicative complement	TA	Time adverbial
FS	Dummy subject	TT	Address phrase
FV	Finite predicate verb	UK	Subordinating conjunction
I?	Question mark	VA	Notifying adverbial
IC	Quotation mark	VO	Infinitive object complement
IG	Other punctuation mark	VS	Infinitive subject complement
IK	Comma	XA	Expressions like "så att säga" (so to speak)
IM	Infinitive marker	XF	Fundament phrase
IO	Indirect object	XT	Expressions like "så kallad" (so called)
IP	Period	XX	Unclassifiable grammatical function
IQ	Colon	YY	Interjection phrase
IR	Parenthesis	**New Categories**	
IS	Semicolon	CJ	Conjunct (in coordinate structure)
IT	Dash	HD	Head
IU	Exclamation mark	IF	Infinitive verb phrase minus infinitive marker
IV	Nonfinite verb	PA	Complement of preposition
JC	Second quotation mark	UA	Subordinate clause minus subordinating conjunction
JG	Second (other) punctuation mark	VG	Verb group

Table 7: MAMBA categories for annotation of dependency relations (DepRel code, English term).

C Appendix: Statistics of the three datasets

Dataset	Level	# sent.	# tokens excl.punct
L1 Coctaill	**15 texts**	**196**	**1900**
	A1	57	434
	A2	19	101
	B1	57	553
	B2	32	488
	C1	31	324
L2 orig	**15 texts**	**287**	**3635**
	A1	42	301
	A2	60	481
	B1	61	814
	B2	63	886
	C1	61	1153
L2 norm	**15 texts**	**306**	**3565**
	A1	52	323
	A2	60	499
	B1	65	852
	B2	64	1159
	C1	65	732
Total	**45 texts**	**789**	**9100**

Table 8: Statistics over the three datasets