doi:10.5128/ERYa18.15

POS-TAGGING TARTU CORPUS OF ESTONIAN LEARNER ENGLISH WITH CLAWS7

Liina Tammekänd, Reeli Torn-Leesik

Abstract. The aim of the study is to examine whether the CLAWS7 tagger is a suitable tool for tagging the Tartu Corpus of Estonian Learner English (TCELE). Extracts were tagged manually and automatically, and the results were compared to calculate the error rate and reveal the possible causes for tagger errors. The error rate was 4.01%. The tagger expectedly experienced some of the disambiguation problems outlined in the CLAWS7 post-editing guide, yet certain tagger errors were also triggered by learner errors.*

Keywords: Estonian learner English, TCELE, POS-tagging, tagger errors, corpus linguistics

1. Introduction

Learner language is a foreign language that is spoken by language learners and that is not an official language in their home country (Granger 2008: 260). Learner language is also known as interlanguage (Selinker 1972, 1992, Corder 1981) and represents a language system that the learner builds on the basis of linguistic input from the language they are learning. It is not a steady-state product but rather is dynamic in nature and exhibits variation. Learner language reflects the stage at which learners find themselves on their way to internalising target language norms. It can be described as a "transitional system reflecting the learner's current L2 knowledge" (Ellis 1994: 16).

Earlier research on learner language has often been based on data that are drawn from highly controlled language tests and that have been collected from a limited number of learner groups (Granger et al. 2015). Unlike such data, learner corpora that consist of "electronic collections of texts produced by language learners" (Granger 2008: 259) are large and contain samples from many learners. Their electronic format allows for speed and ease of analysis and makes the data suitable for many types of studies. The results of learner corpus research help shed light on the characteristics of learner language, make a contribution to second language

* The authors of the article are listed in alphabetical order.

acquisition theory as well as to pedagogical methods and tools that meet language learners' needs (Granger 2008).

In the last two decades, several specialised corpora for learner language study have been compiled in Estonia – The Estonian Interlanguage Corpus of Tallinn University (EIC[1]; Eslon 2014), the learner language corpus of the University of Tartu[2] (Sõrmus, Lepajõe 2014) and a smaller corpus of learner Spanish (The Tartu Learner Corpus of Spanish as a L3+; Kruse 2018). Still, Estonian learner English remains a largely unexplored field. This study is the first in what will hopefully become a series of research papers on the subject.

The aim of the study at hand is to determine whether the error rate of the automatic CLAWS7 (Constituent Likelihood Automatic Word-tagging System) tagger allows it to be recommended as a suitable tool for tagging Estonian learner English. The questions that motivate the study are: What is the error rate of CLAWS7 in Estonian learner English? What are the main causes for tagging errors?

The paper is divided into two main parts. The first part gives an overview of POS-taggers, problems in POS-tagging, evaluation and accuracy of taggers, the CLAWS word tagging system and its most common issues. The second part describes the process of tagging the Tartu Corpus of Estonian Learner English (TCELE) with the CLAWS7 tagger and discusses tagger errors and their possible reasons.

2. Automatic POS-taggers and problems in POS-tagging

Corpus annotation means adding interpretative, linguistic information to an electronic spoken or written corpus (Leech 2013). This type of annotation allows otherwise unavailable information to be extracted from the corpus. POS-tagging is a sub-type of corpus annotation and is typically undertaken automatically by means of a computer program (POS-tagger) which assigns each word a "tag" that identifies the part-of-speech category that the word belongs to and collects other grammatical category information regarding it without input from the user (Newman, Cox 2020, Gries, Berez 2017, van Rooy 2015, Jurafsky, Martin 2008: 123–172). There are three main types of POS-taggers. Rule-based POS-taggers use hand-written disambiguation rules when assigning POS-tags to words. Such taggers are TAGGIT (Green, Rubin 1971), TOSCA (Oosdijk 1991), Constraint Grammars and EngCG (Voutilainen 1994, Karlsson et al. 1995), and AMBILIC (de Yzaguirre et al. 2000). Stochastic taggers – a category to which the CLAWS (Garside et al. 1987) tagger belongs – are trained on an already tagged corpus to calculate the probability of a word having a particular tag in a specific context. Hybrid taggers use both handwritten disambiguation rules and probability calculations. An example of such taggers is Brill (1992).

The process of POS-tagging consists of three phases. In the first phase, a tokeniser divides the text into tokens (words, punctuation marks and utterance boundaries). Then, a lookup module uses a lexicon and a guesser to assign possible tags to each word. Finally, a disambiguation module selects a tag, using contextual (word-tag sequences) and statistical information (Voutilainen 1999, 2003).

[1] https://evkk.tlu.ee (30.10.2021).

[2] https://korpused.keeleressursid.ee/emma (30.10.2021).

The third phase tends to present the most problems (Voutilainen 2003). Although English has many words that are unambiguous and easily tagged correctly, many frequently used English words present ambiguities. Thus, 11.5% of English word types in the Brown corpus and 40% of Brown tokens are ambiguous and will be considered for several different tags by a POS-tagger (Jurafsky, Martin 2008: 123–172). Jurafsky and Martin (*ibid.*) identify three main sources of ambiguity in the POS-tagging of English texts: a) prepositions, particles and adverbs often overlap; b) it is difficult to tag common nouns, proper nouns and adjectives when they modify nouns; c) it is difficult to differentiate between participles and adjectives. A POS-tagger needs to resolve these ambiguities successfully (Voutilainen 2003).

The accuracy of POS-tagging depends on several factors: the nature of corpus language and its morphological features, the complexity of the texts in the corpus, the size and POS-tagging accuracy of the training corpus and the size of the tagset, etc (Griez, Berez 2017). Also, taggers tend to perform worse in tagging learner language because it has errors and features structures that do not occur in the language of the training corpus (van Rooy 2015). At the same time, the fact that learner language is simple in nature and that errors (e.g., semantic issues) have no significant impact on the automatic POS-tagging process (*ibid.*) still allows taggers to perform adequately.

According to Nagata et al. (2018), POS-tagging errors in learner English are mainly caused by three factors. Learner English texts have many unknown words that are caused either by spelling or grammar errors and are unlikely to occur in the training corpus. Compared to native-speaker data, learner English has different POS-distributions. For example, the word *concentrate* is typically used as a noun in newspaper texts but often appears as a verb in academic learner English. Learner English has characteristic POS-sequences, some of which may depend on learners' L1 and some of which seem to be universal among many English learners. Thus, Aarts and Granger (1998) found that English learners with French, Dutch and Finnish L1 overuse sentence-initial connectives, adverbs, auxiliaries and pronouns and underuse patterns with prepositions, sentence-initial nouns, conjunctions+nouns and prepositions+-*ing*-verbs. These learner preferences might have an adverse effect on the outcome of automatic POS-tagging.

Linguists are mostly interested in taggers' accuracy, the metrics of which are precision, recall, ambiguity and error rate/correctness (Voutilainen 2003). Precision measures how many of the tokens tagged X were tagged correctly. Recall measures how many of the tokens that should have been tagged X have indeed been so tagged. Ambiguity counts the average number of tags each token gets. Error rate/correctness measures how many tokens receive a contextually appropriate tag (van Halteren 1999) and is evaluated by comparing the "Gold Standard" – a manually tagged test text – to the tagger's output of the same text (Jurafsky, Martin 2008: 123–172). It should be noted that the Gold Standard itself might have a 3–4% error rate (*ibid.*).

POS-taggers trained on native English and French texts achieve an accuracy of 96% when tagging native texts in the language of training (van Rooy 2015). The accuracy rate of tagging learner English is below 90%, but tends to increase by about 6% when spelling errors are corrected (van Rooy, Schafer 2002, 2003 as cited in van Rooy 2015). De Haan (2000) reports a learner English tagging accuracy rate of 95%.

3. CLAWS word-tagging system and its most common disambiguation problems

The CLAWS POS-tagging system is one of the first POS-taggers that uses statistical calculations and achieves an accuracy of 95–98% in tagging native English texts, depending on text type (Garside 1996, UCREL[3] Team 1996). The first version of CLAWS was developed as a joint project by researchers of Lancaster University, the University of Oslo and and the Norwegian Computing Centre for the Humanities, (Bergen) in 1981–1983, and was used to tag the million-word Lancaster-Oslo-Bergen Corpus (Garside 1987). The tags of CLAWS1 were based on the Brown Corpus tagset (Garside 1987). The British National Corpus (BNC) was tagged with the C5[4], or a main tagset of just over 60 tags. The C7[5] set of 137 tags was used to tag the "core" corpus sample of 2 million words (Garside 1996).

The CLAWS tagging system assigns one or more tags from its tagset using the following resources:

a) a lexicon, which consists of about 12,000 words; each word in the lexicon has 1–6 possible tags (UCREL Team 1996). 65–70% of all words get their potential tags from the lexicon (Garside 1996);

b) a suffix list that links common or predictable word endings with possible tags;

c) an idiom list that features multiword units whose syntactic role in the sentence might differ from the roles of unit constituents;

d) probability data: potential POS-tags are assigned to each word according to rules based on the word's orthography and suffix endings; following that, statistical calculations are conducted to choose the most probable tag (UCREL Team 1996).

The CLAWS tagger goes through several steps to produce a POS-tagged text. First, it tokenises the text and assigns one or more possible POS-tags to each word with the help of its lexicon. The words that are not in the lexicon are tagged with the help of the suffix list. Then, the word and tag patterns on the "idioms list" and in the template libraries which provide contextual cues are compared to patterns in the text and changes are made in the assigned tags, if necessary. The words that have more than one tag are inspected, and statistical calculations are conducted to choose the most probable tag in the given context (Garside 1987, 1996, UCREL Team 1996).

Despite the resources and processes available to the CLAWS tagger, it still faces a number of specific disambiguation issues. According to the UCREL Team (1996), the most problematic ones are as follows.

The tagger finds it difficult to differentiate between comparative after-determiners (DAR)[6] and comparative general adverbs (RRR). For instance, the tag DAR should be assigned to noun-phrase-like uses of the word *more* (*You should spend more_DAR*)[7], while the tag RRR should be assigned to adverbial uses, e.g. (*You should relax more_RRR*).

The tagger also has problems differentiating between general prepositions (II) and locative adverbs (RL), as well as general prepositions (II) and prepositional

[3] UCREL is a research centre of Lancaster University.

[4] http://ucrel.lancs.ac.uk/claws5tags.html (30.10.2021).

[5] http://ucrel.lancs.ac.uk/claws7tags.html (30.10.2021).

[6] See the Appendix.

[7] All examples in this section have been taken from *A Post-Editor's Guide to Claws7 Tagging* by UCREL Team (1996).

adverbs or particles (RP). Errors in tagging may occur in relation to stranded prepositions whose NP complements have been fronted or elided as, for example, in relative clauses, passives or questions (*Which car did you arrive in_II?*).

General adjectives (JJ) and singular common nouns (NN1) are another source of difficulties. Words ending in *-ing* may receive both the NN1 (*new_JJ spending_NN1 reductions_NN2*) or JJ (*working_JJ mother_NN1*) tag. The example *working_JJ mother_NN1* means *mother who works*, i.e. the noun *mother* is the notional subject of the verb *work* and *working* should get the JJ tag. In other cases, the *-ing*-word should receive the NN1 tag.

Tagging errors may occur when the tagger encounters a general adjective (JJ), general adverb (RR), general comparative adjective (JJR) or general comparative adverb (RRR). Ambiguities arise if the word appears after a verb or an object (*they arrived tired_JJ and hungry_JJ*; *Peter sang out loud_RR and clear_RR*).

The tagger finds it difficult to tag general adjectives (JJ) and *-ing* participles of lexical verbs (VVG), and general adjectives and past participles of lexical verbs (VVN). An *-ing*-word should receive the VVG tag after the verb *be* (*the man was dying_VVG*) and the JJ tag after nouns (*the dying_JJ man*). When the *-ing* or *-en*/*-ed* word is part of a phrase premodifying a noun, it is tagged VVG/VVN (*interest_NN1 earning_VVG account*). If a NN1-VVG/VVN sequence is hyphenated, it may be tagged as JJ. With event verbs, the JJ refers to a resultant state (*Bill was married_JJ* = not single) and the VVG/VVN to an event (*Bill was married_VVN to Sarah last week*).

Degree adverbs (RG) vs general adverbs (RR) also represent ambiguity. Intensifiers (also known as adverbs of degree, e.g. *very*, *so*, and *as* in comparatives) modifying a word or phrase should receive the RG tag. Adverbs that have many other functions besides intensification are usually tagged with the more general RR tag following the general-specific ambiguity rule, according to which the general tag within a category is selected instead of a specific one in the same category to avoid the proliferation of tagging ambiguities. Words which may be tagged RG or RR are *so*, *too*, *quite*, and *rather* (*she is so_RG attractive*; *I would think so_RR*). (UCREL Team 1996)

4. Tagging the Tartu Corpus of Estonian Learner English with CLAWS7

4.1. Material

The aim of the study is to examine whether the CLAWS tagging system, one of the most popular taggers that is freely available online – and its C7 tagset, which has been used to tag the BNC – represent a good choice for tagging Tartu Corpus of Estonian Learner English (TCELE). TCELE is a learner English corpus still being compiled at the Department of English Studies of the University of Tartu. TCELE consists of essays written as part of the University of Tartu's English Language and Literature BA programme entrance exam and currently has 75,818 words. The essays generally run to 250–300 words (although there are exceptions to the length) and supposed to represent on a short journalistic text. Writing the essay is timed and the assumed proficiency level of exam essays is CEFR B2. Out of the corpus, 10 texts of about 200 words (below, the "tagged mini-corpus") were chosen randomly. The average length of the essays was 268.5 words. The shortest essay was 194 and the longest 413 words. The tagged mini-corpus consisted of 2658 words.

4.2. Method

Two linguists, whose L1 is Estonian and who have received no training in the CLAWS tagging system and its architecture but are expert users of English and work with the language on an everyday basis, manually tagged the randomly chosen essays in a double-blind arrangement using the C7 tagset. The same essays were then automatically tagged using CLAWS. Finally, the automatic and the manual tagging output were compared to calculate the tagger's error rate and shed light on possible causes for errors in automatic tagging.

Before the error rate could be calculated, a series of determinations had to be made concerning divergences to be considered tagger errors. For instance, such determinations had to be taken in situations when the tagger assigned a wrong tag to a word because no correct tag was available in the tagset – for instance, when tagging the reciprocal pronoun *each other* and the relative pronouns *that* and *which*. The tagger has no tag for reciprocal pronouns and therefore *each other* is consistently tagged as a reflexive pronoun (*each_PPX221 other_PPX222*), which it is not but which could be considered the tagger's closest match. The tagger also has no tag for relative pronouns and tags *that* and *which* as, respectively, a conjunction and a determiner. As it is difficult to understand the reasons behind these analyses, such instances were counted as tagger errors.

There were also cases where the tagger assigned two tags to one word or one tag to two words, requiring a determination on whether these count as one or two errors. For instance, the tagger analysed the compound *science-based* as two words and assigned separate tags (*science_NN1 and based_VVN*) to its components. We analysed the word as a single one and tagged it as an adjective (*science-based_JJ*). Another example concerned the case of a learner error (*persons* for *person's*), which the tagger analysed as a plural noun (*persons_NN2*) and not the genitive of a singular one (*persons_NN1+_GE*) as should have emerged from the context. In both cases the incorrect tag was counted as one error, not two.

Occasionally, the tagger identified the general category correctly, but was unsuccessful in deciding which tag to assign to a word within that category. Such cases required a determination on whether to consider them tagger errors. For instance, the tagger tagged words such as *information* and *Tartu* as singular nouns (NN1) and *media* as a common noun neutral for number (NN). We decided to use specific categories for these words (*information_NN*, *Tartu_NP1*, *media_NN2*), considering that since issues within the verb category are counted as tagger errors, issues within the noun category should also be counted as such. A similar problem arose when the tagger tagged *more* as a comparative general adverb (*more_RRR*) instead of a comparative degree adverb (*more_RGR*). Since the post-editing guide (UCREL Team 1996) explicitly allows this type of general-specific ambiguity, we did not count such instances as tagger errors.

Tagging the determiner phrase *a lot of* as individual words (*a_AT1 lot_NN1 of_IO*) seemed puzzling as well. Yet, such an analysis of *a lot of* appears reasonable if the tagger does not have the phrase in its idioms list as a multi-word unit and therefore tags the words separately. As the tagger analysed each word in the phrase correctly, corresponding instances were not considered tagger errors.

4.3. Tagger errors

Having resolved the issues outlined above, the following results were obtained (see Table 1). Out of 2685 tagged words in the tagged mini-corpus, 110 words had been mistagged, either because of a tagger disambiguation problem, a random tagger error or by a tagger error caused by a learner error. Out of 110 tagging errors, 17 coincided with learner errors, and 15 of these had an adverse effect on the tagging process. Even without removing learner spelling errors (van Rooy, Schäfer 2002), the tagger's error rate was 4.01%. This corresponds to van Rooy and Schäfer's (2002) and De Haan's (2000) findings who both report accuracy rates of above 95%.

Table 1. Error rate

Words in the mini-corpus	Mistagged words		Error rate
	Tagger errors	Learner errors	
2685	93	15	4.01%

4.3.1. Errors caused by disambiguation issues

A separate type of errors caused by tagging rules were those related to disambiguation issues (see Section 3). Words with multiple word class potential and formal overlap were often the cause of incorrect tagging. One instance of this is ambiguity between nouns and adjectives. For example, the word *English*, which can function as an adjective or a noun depending on the context, occurred 64 times in the tagged mini-corpus, out of which an incorrect tag was assigned to it on 10 occasions. In most of these cases, the tagger incorrectly analysed *English* as an adjective although the word functioned as a noun. A similar tagging error occurred with the word *Estonian*.

Other words that posed tagging problems were those that can function as determiners, adverbs, prepositions or conjunctions depending on their context. For instance, the words *more* and *much* occurred 17 and 5 times respectively in the tagged mini-corpus. Both can function as determiners in a noun phrase or degree adverbs in front of adjectives. While *more* received a wrong tag only once, *much* was tagged incorrectly on 4 occasions. Similar tagging problems occurred with the word *as*, which can be tagged as a preposition (II), conjunction (CSA), general adverb (RR) and degree adverb (RG). There were 26 instances of *as*, 8 of which involved the word occurring as a constituent of 4 *as…as*-structures. Two of these structures were wrongly tagged. The manual instructs to tag the first *as* in the *as…as*-structure as a degree adverb (RG) and the second *as* a conjunction (CSA). It seems that tag assignment may be influenced by the distance between the first and the second *as*.

The words *today* and *tomorrow* can function as adverbs or nouns. *Today* occurred three times and *tomorrow* once in the tagged mini-corpus. All these instances were analysed by the tagger as time adverbs, although several of the contexts clearly pointed to noun function, as can be seen in (1). It may be the case that the tagger always tags these words as adverbs.

(1) in_II **today_RT** `_" s_ZZ1 ever_RR changing_JJ world_NN1 today_NN 's_GE

As mentioned in Section 4.2, the tagger always analyses demonstratives *that* and *this* as determiners and not demonstrative pronouns – even if they clearly display pronominal function as in (2a) and (2b). The reason lies in the fact that the tagset has only one tag for *this* and *that* – regardless of whether they function as a pronoun or determiner.

(2a) But_CCB what_DDQ exactly_RR will_VM be_VBI consequences_NN2 of_IO **that_DD1**

(2b) **This_DD1** also_RR encourages_VVZ people_NN to_TO move_VVI to_II

That appeared 65 times and *this* 25 times in the tagged mini-corpus. While *that* was tagged as a determiner (DD1) or a conjunction (CST), *this* was always tagged as a determiner. *That* in conjunctive function was incorrectly tagged as a determiner 3 times as illustrated in (3a), yet there were also 2 instances of *that* tagged as a conjunction although it actually appeared as a relative pronoun, as in (3b).

(3a) I_PPIS1 think_VV0 **that_DD1** nature_NN1 and_CC primary_JJ purpose_NN1 of_IO)

(3b) a_AT1 place_NN1 **that_CST** can_VM help_VVI turn_NN1 life_NN1 around_RP)

As already mentioned, the tagger has no tag for relative pronouns and can make a choice only between tagging one as a determiner or a conjunction. This is, however, misleading because *that* in these two instances clearly belonged to different word classes. The absence of a tag for pronouns is also unfortunate because it makes studies on relative clauses in learner English more difficult. Exactly the same problem arises with the relative pronoun *which*. The latter occurred 7 times in the tagged mini-corpus and was tagged as a *wh*-determiner (DDQ), as illustrated in (4). As the tagset has no tag for relative pronouns, the word is analysed as a determiner.

(4) by_II giving_VVG them_PPHO2 knowledge_NN1 **which_DDQ** is_VBZ a_AT1 tool_NN1.

The tagger also made errors within the category when assigning a tagger specification. For instance, nouns such as *mathematics*, *engineering*, *usage*, *information*, *solving* – all of which occurred once in the tagged mini-corpus – were tagged NN1 (singular common nouns). The same analysis was applied to the words *communication* and *extinction*, occurring 6 and 3 times respectively. All of these words being uncountable nouns, the tag for a common noun, neutral for number (NN) would have made more sense.

As to verbs, the tagger had occasional problems differentiating non-finite bare infinitive forms from finite base forms in the sequence of several V-NP-V structures that omit the first V by ellipsis. The structure is illustrated in (5). In the first VP *help us govern ourselves*, the verb *govern* is correctly tagged as an infinitive (VVI), but in the following VPs *understand our development* and *argue for doing it better* the first verb *help* is elided and the tagger mistakenly tags the verbs *understand* and *argue* as VV0 (finite base forms).

(5) Firstly_RR ,_, the_AT Shape_NN1 subjects_NN2 help_VV0 us_PPIO2 govern_VVI ourselves_PPX2 ,_, **understand_VV0** our_APPGE development_NN1 over_II time_NNT1 and_CC **argue_VV0** for_IF doing_VDG it_PPH1 better_RRR

The tagger makes some errors with complex transitive verb structures in which the adjective complementing the direct object is analysed as an adverb. This error is illustrated in examples (6a) and (6b).

(6a) makes_VVZ communication_NN1 between_II companies_NN2 **easier_RRR**
(6b) knowing_VVG one_MC1 very_RG popular_JJ language_NN1 makes_VVZ travelling_VVG **easier_RRR** and_CC more_RGR safe_JJ

The tagger also makes prediction errors when tagged words can occur in several possible structures. For instance, in (7a), the word *before* is analysed as a subordinating conjunction (CS) although the context shows it to be an adverb. The same error is shown in (7b) where the word *after* is analysed as a subordinating conjunction and not a preposition. As both *before* and *after* can function as subordinating conjunctions, the tagger seems to expect them to be followed by a subordinating clause.

(7a) having_VHG studied_VVN here_RL **before_CS** I_PPIS1 have_VH0 become_VVN to_TO
(7b) decided_VVN on_II applying_VVG again_RT **after_CS** any_DA2 years_

4.3.2. Random errors

The analysis also revealed errors that appeared random and were therefore difficult to explain in the context of their occurrence. Several examples are shown below. In examples (8a) and (8b), the tagger misanalyses verbs as nouns. In (8a) and (8b), *changes* and *means* are tagged as a plural noun (NN2) and common noun neutral in number (NN) respectively, yet in respect of both an analysis assigning them the function of a verb would have been more logical since the tagger's analysis leaves the clause without a finite verb. In (8c), *will* is analysed as a noun (NN1) although it clearly appears as a modal verb in the verb phrase *will be speaking*. Again, the reasons for the tagger's choice are unclear.

(8a) every_AT1 language_NN1 **changes_NN2** constanly_RR
(8b) Speaking_VVG a_AT1 language_NN1 **means_NN** having_VHG an_AT1 opportunity_NN1 to_TO understand_VVI (8c) whether_CSW we_PPIS2 all_DB **will_NN1** one_MC1 day_NNT1 be_VBI speaking_VVG

In (9) the noun *stem* in the noun phrase *stem subjects* is analysed as a verb (VV0), which should have been ruled out as improbable since the phrase is the subject of the finite verb *are*.

(9) humanities_NN2 and_CC **Stem_VV0** subjects_NN2 are_VBR quite_RG equal_JJ

In (10) the adverb *overall* is analysed as a noun.

(10) their_APPGE future_JJ career_NN1 ,_, but_CCB also_RR life_NN1 **overall_NN1**

Tagger errors also include some instances of incorrect tagging of the genitive construction and of contracted negative forms. The mini-corpus included 3 genitive constructions and 2 contracted negatives, both being incorrectly tagged on one occasion. Because of failing to tag what follows the apostrophe in these constructions, the tagger makes a mistake also in tagging the word before the apostrophe as in (11a) and (11b). The tagger had no problems with tagging contracted tense forms.

(11a) we_PPIS2 **aren_NN1** `_" t_ZZ1 all_RR so_RG different_JJ
(11b) in_II **today_RT** `_" s_ZZ1 ever_RR changing_JJ world_NN1

4.3.3. Errors caused by learner errors

The mini-corpus also included 15 tagger errors caused by learner errors. The latter can be assigned to the categories of spelling errors (6 instances), morphological errors (6 instances), grammar errors (3 instances), and punctuation errors (2 instances). Learner errors are illustrated in examples (12a) and (12b). In (12a), a learner's word *determinate* for the word *determine* causes the tagger to misanalyse the word as an adjective (JJ). It may be the case that the tagger analyses the suffix *-ate* as an adjectival one. In (12b), the tagger is unable to assign a correct tag for the misspelt word *litirature* (literature).

(12a) gets_VVZ harder_RRR to_II **determinate_JJ** whose_DDQGE language_NN1 harder_JJR to_TO
(12b) Why_RRQ I_PPIS1 choose_VV0 English_JJ language_NN1 and_CC **litirature_VV0**

Although the number of such errors is not large, all of these (except punctuation errors) caused the tagger to assign an incorrect tag.

5. Concluding remarks

The aim of the study was to examine whether the CLAWS7 tagger can be considered a suitable tool for tagging Estonian learner English, more specifically Tartu Corpus of Estonian Learner English. The questions posed in this study were: What is the error rate of CLAWS7 in Estonian learner English? What are the main causes for tagging errors?

The error rate of the CLAWS7 tagger was 4.01%, which coincides with previous similar findings concerning the tagging of learner English (van Rooy 2015, van Rooy, Schafer 2002, de Haan 2000). The errors were mainly caused by disambiguation problems and by learner errors. Some errors could not be explained by their context.

As pointed out by the UCREL Team (1996), the CLAWS tagger indeed had problems in distinguishing determiners from adverbs, general adverbs and singular common nouns, as well as adjectives from adverbs. The tagger had additional difficulties in deciding how to assign a more specific tag in the categories of nouns and verbs.When tagging Estonian learner English, the tagger also experienced problems distinguishing adverbs from nouns, as well as conjunctions from adverbs. These specific problems might be caused by the peculiarities of Estonian learner English and their exact nature has yet to be studied. A major issue for the learner English researcher is that the C7 tagset lacks suitable tags for *this*/*that* when used as pronouns, and for relative pronouns. Use of relative clauses and referential constructions by learners of English of any native tongue, not only Estonian, is an interesting field of analysis, and the tagger's failure to identify certain classes of pronouns might convince the researcher to decide in favour of a different tagger.

Despite its shortcomings, the tagger performed well and can be used to tag TCELE. When conducting further analyses, the weaknesses outlined above have to be addressed.

References

Aarts, Jan; Granger, Sylviane 1998. Tag sequences in learner corpora: A key to interlanguage grammar and discourse. – Sylviane Granger (Ed.), Learner English on Computer. London: Routledge, 132–141. https://doi.org/10.4324/9781315841342-10

Brill, Eric 1992. A simple rule-based part of speech tagger. – DARPA: Proceedings of the Speech and Natural Language Workshop. Morgan Kauffman online, 112–116.

Corder, Pit 1981. Error Analysis and Interlanguage. Oxford: Oxford University Press.

de Haan, Pieter 2000. Tagging non-native English with the TOSCA–ICLE tagger. – Christian Mair, Marianne Hundt (Eds.), Corpus Linguistics and Linguistic Theory. Language and Computers 33. Amsterdam: Rodopi, 69–79. https://doi.org/10.1163/9789004490758_007

Ellis, Rod 1994. The Study of Second Language Acquisition. Oxford: Oxford University Press.

Eslon, Pille 2014. Eesti vahekeele korpus [‘Estonian interlanguage corpus’]. – Keel ja Kirjandus, 6, 436–451. https://doi.org/10.54013/kk679a3

Granger, Sylviane 2008. Learner corpora. – Anke Lüdeling, Merja Kytö (Eds.), Corpus Linguistics. An International Handbook. Vol 1. Berlin–New York: Walter de Gruyter, 259–275.

Granger, Sylviane; Gilquin, Gaëtanelle; Meunier, Fanny 2015. Introduction. Learner corpus research: Past, present and future. – Sylviane Granger, Gaëtanelle Gilquin, Fanny Meunier (Eds.), The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press, 1–5. https://doi.org/10.1017/CBO9781139649414

Gries, Stefan T.; Berez, Andrea L. 2017. Linguistic annotation in/for corpus linguistics. – Nancy Ide, James Pustejovsky (Eds.), Handbook of Linguistic Annotation. Dordrecht: Springer, 379–410. https://doi.org/10.1007/978-94-024-0881-2_15

Jurafsky, Daniel; Martin, James H. 2008. Speech and Language Processing. 2nd ed. Upper Saddle River: Prentice Hall.

Kruse, Mari 2018. La transferencia en personas plurilingües: los falsos amigos como un obstáculo y una oportunidad en la enseñanza y aprendizaje de lenguas extranjeras.

Dissertationes philologiae romanicae Universitatis Tartuensis 8. Tartu: Tartu Ülikooli Kirjastus.

Leech, Geoffrey 2013. Introducing corpus annotation. – Roger Garside, Geoffrey Leech, Tony McEnery (Eds.), Corpus Annotation. Linguistic Information from Computer Text Corpora. London: Routledge, 1–18.

Nagata, Ryo; Mizumoto, Tomoya; Kikuchi, Yuta; Kawasaki, Yoshifumi; Funakoshi, Kotaro 2018. A POS tagging model designed for learner English. – Wei Xu, Alan Ritter, Tim Baldwin, Afshin Rahimi (Eds.), Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text. Brussels: Association for Computational Linguistics, 39–48. https://doi.org/10.18653/v1/W18-6106

Oostdijk, Nelleke 1991. Corpus Linguistics and the Automatic Analysis of English. Amsterdam: Rodopi.

Selinker, Larry 1972. Interlanguage. – International Review of Applied Linguistics, 10 (3), 209–231. https://doi.org/10.1515/iral.1972.10.1-4.209

Sõrmus, Kadri; Lepajõe, Kersti 2014. Eesti keele kui emakeele õppija tekstikorpus EMMA [‘The Estonian native-speaking students’ text corpus EMMA’]. – Philologia Estonica Tallinnensis, 16, 205–227.

UCREL CLAWS7 Tagset. http://ucrel.lancs.ac.uk/claws7tags.html (30.10.2021).

UCREL Team 1996. A Post-Editor’s Guide to Claws7 Tagging. http://www.natcorp.ox.ac.uk/docs/claws7.html (30.10.2021).

van Rooy, Bertus 2015. Annotating learner corpora. – Sylviane Granger, Gaëtanelle Gilquin, Fanny Meunier (Eds.), The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press, 79–106. https://doi.org/10.1017/CBO9781139649414.005

van Rooy, Bertus; Schäfer, Lande 2002. The effect of learner errors on POS tag errors during automatic POS tagging. – Southern African Linguistics and Applied Language Studies, 20 (4), 325–335. https://doi.org/10.2989/16073610209486319

van Rooy, Bertus; Schäfer, Lande 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. – Dawn Archer, Paul Rayson, Andrew Wilson, Tony McEnery (Eds.), Proceedings of the Corpus Linguistics 2003 Conference. Lancaster University Centre for Computer Corpus Research on Language Technical Papers 16, 835–844.

Voutilainen, Atro 2003. Part-of-speech-tagging. – Ruslan Mitkov (Ed.), The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, 219–232. https://doi.org/10.1093/oxfordhb/9780199276349.013.0011

Appendix 1. CLAWS7 Tagset

APPGE	possessive pronoun, pre-nominal (e.g. *my*, *your*, *our*)
AT	article (e.g. *the*, *no*)
AT1	singular article (e.g. *a*, *an*, *every*)
BCL	before-clause marker (e.g. *in order (that)*, *in order (to)*)
CC	coordinating conjunction (e.g. *and*, *or*)
CCB	adversative coordinating conjunction (*but*)
CS	subordinating conjunction (e.g. *if*, *because*, *unless*, *so*, *for*)
CSA	*as* (as conjunction)
CSN	*than* (as conjunction)
CST	*that* (as conjunction)
CSW	*whether* (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. *such*, *former*, *same*)
DA1	singular after-determiner (e.g. *little*, *much*)
DA2	plural after-determiner (e.g. *few*, *several*, *many*)
DAR	comparative after-determiner (e.g. *more*, *less*, *fewer*)
DAT	superlative after-determiner (e.g. *most*, *least*, *fewest*)
DB	before determiner or pre-determiner capable of pronominal function (*all*, *half*)
DB2	plural before-determiner (*both*)
DD	determiner (capable of pronominal function) (e.g *any*, *some*)
DD1	singular determiner (e.g. *this*, *that*, *another*)
DD2	plural determiner (*these*, *those*)
DDQ	wh-determiner (*which*, *what*)
DDQGE	wh-determiner, genitive (*whose*)
DDQV	wh-ever determiner, (*whichever*, *whatever*)
EX	existential *there*
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker (*'* or *'s*)
IF	*for* (as preposition)
II	general preposition
IO	*of* (as preposition)
IW	*with*, *without* (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. *older*, *better*, *stronger*)
JJT	general superlative adjective (e.g. *oldest*, *best*, *strongest*)
JK	catenative adjective (*able* in *be able to*, *willing* in *be willing to*)
MC	cardinal number,neutral for number (*two*, *three*)
MC1	singular cardinal number (*one*)
MC2	plural cardinal number (e.g. *sixes*, *sevens*)
MCGE	genitive cardinal number, neutral for number (*two's*, *100's*)
MCMC	hyphenated number (*40–50*, *1770–1827*)
MD	ordinal number (e.g. *first*, *second*, *next*, *last*)
MF	fraction, neutral for number (e.g. *quarters*, *two-thirds*)
ND1	singular noun of direction (e.g. *north*, *southeast*)
NN	common noun, neutral for number (e.g. *sheep*, *cod*, *headquarters*)
NN1	singular common noun (e.g. *book*, *girl*)
NN2	plural common noun (e.g. *books*, *girls*)

NNA	following noun of title (e.g. *M.A.*)
NNB	preceding noun of title (e.g. *Mr.*, *Prof.*)
NNL1	singular locative noun (e.g. *Island*, *Street*)
NNL2	plural locative noun (e.g. *Islands*, *Streets*)
NNO	numeral noun, neutral for number (e.g. *dozen*, *hundred*)
NNO2	numeral noun, plural (e.g. *hundreds*, *thousands*)
NNT1	temporal noun, singular (e.g. *day*, *week*, *year*)
NNT2	temporal noun, plural (e.g. *days*, *weeks*, *years*)
NNU	unit of measurement, neutral for number (e.g. *in*, *cc*)
NNU1	singular unit of measurement (e.g. *inch*, *centimetre*)
NNU2	plural unit of measurement (e.g. *ins.*, *feet*)
NP	proper noun, neutral for number (e.g. *IBM*, *Andes*)
NP1	singular proper noun (e.g. *London*, *Jane*, *Frederick*)
NP2	plural proper noun (e.g. *Browns*, *Reagans*, *Koreas*)
NPD1	singular weekday noun (e.g. *Sunday*)
NPD2	plural weekday noun (e.g. *Sundays*)
NPM1	singular month noun (e.g. *October*)
NPM2	plural month noun (e.g. *Octobers*)
PN	indefinite pronoun, neutral for number (*none*)
PN1	indefinite pronoun, singular (e.g. *anyone*, *everything*, *nobody*, *one*)
PNQO	objective wh-pronoun (*whom*)
PNQS	subjective wh-pronoun (*who*)
PNQV	wh-ever pronoun (*whoever*)
PNX1	reflexive indefinite pronoun (*oneself*)
PPGE	nominal possessive personal pronoun (e.g. *mine*, *yours*)
PPH1	3rd person sing. neuter personal pronoun (*it*)
PPHO1	3rd person sing. objective personal pronoun (*him*, *her*)
PPHO2	3rd person plural objective personal pronoun (*them*)
PPHS1	3rd person sing. subjective personal pronoun (*he*, *she*)
PPHS2	3rd person plural subjective personal pronoun (*they*)
PPIO1	1st person sing. objective personal pronoun (*me*)
PPIO2	1st person plural objective personal pronoun (*us*)
PPIS1	1st person sing. subjective personal pronoun (*I*)
PPIS2	1st person plural subjective personal pronoun (*we*)
PPX1	singular reflexive personal pronoun (e.g. *yourself*, *itself*)
PPX2	plural reflexive personal pronoun (e.g. *yourselves*, *themselves*)
PPY	2nd person personal pronoun (*you*)
RA	adverb, after nominal head (e.g. *else*, *galore*)
REX	adverb introducing appositional constructions (*namely*, *e.g.*)
RG	degree adverb (*very*, *so*, *too*)
RGQ	wh- degree adverb (*how*)
RGQV	wh-ever degree adverb (*however*)
RGR	comparative degree adverb (*more*, *less*)
RGT	superlative degree adverb (*most*, *least*)
RL	locative adverb (e.g. *alongside*, *forward*)
RP	prepositional adverb, particle (e.g. *about*, *in*)
RPK	prepositional adverb, catenative (*about* in *be about to*)
RR	general adverb
RRQ	wh- general adverb (*where*, *when*, *why*, *how*)
RRQV	wh-ever general adverb (*wherever*, *whenever*)
RRR	comparative general adverb (e.g. *better*, *longer*)
RRT	superlative general adverb (e.g. *best*, *longest*)

RT	quasi-nominal adverb of time (e.g. *now*, *tomorrow*)
TO	infinitive marker (*to*)
UH	interjection (e.g. *oh*, *yes*, *um*)
VB0	*be*, base form (finite i.e. imperative, subjunctive)
VBDR	*were*
VBDZ	*was*
VBG	*being*
VBI	*be*, infinitive (*to be or not...*, *it will be...*)
VBM	*am*
VBN	*been*
VBR	*are*
VBZ	*is*
VD0	*do*, base form (finite)
VDD	*did*
VDG	*doing*
VDI	*do*, infinitive (*I may do...*, *to do...*)
VDN	*done*
VDZ	*does*
VH0	*have*, base form (finite)
VHD	*had* (past tense)
VHG	*having*
VHI	*have*, infinitive
VHN	*had* (past participle)
VHZ	*has*
VM	modal auxiliary (*can*, *will*, *would*, etc.)
VMK	modal catenative (*ought*, *used*)
VV0	base form of lexical verb (e.g. *give*, *work*)
VVD	past tense of lexical verb (e.g. *gave*, *worked*)
VVG	*-ing* participle of lexical verb (e.g. *giving*, *working*)
VVGK	*-ing* participle catenative (*going* in *be going to*)
VVI	infinitive (e.g. *to give...*, *it will work...*)
VVN	past participle of lexical verb (e.g. *given*, *worked*)
VVNK	past participle catenative (e.g. *bound* in *be bound to*)
VVZ	*-s* form of lexical verb (e.g. *gives*, *works*)
XX	*not*, *n't*
ZZ1	singular letter of the alphabet (e.g. *A*, *B*)
ZZ2	plural letter of the alphabet (e.g. *A's*, *B's*)

SÕNALIIKIDE MÄRGENDAMINE TARTU INGLISE ÕPPIJAKEELE KORPUSES CLAWS7 MÄRGENDAJAGA

Liina Tammekänd, Reeli Torn-Leesik
Tartu Ülikool

Uurimuse eesmärk oli tuvastada, kas CLAWS7 automaatset sõnaliigi märgendajat saab kasutada Tartu inglise õppijakeele korpuse (TCELE) märgendamiseks. TCELE-st juhuslikkuse alusel valitud käsitsi ja automaatselt märgendatud tekstilõike võrreldi omavahel, arvutati automaatse märgendaja veamäär ning analüüsiti märgendamisel tekkinud vigade võimalikke põhjuseid. Automaatse märgendaja veamääraks oli 4,01%. Märgendajal tekkisid ühestusraskused määratlejate ja adverbide, adverbide ja ainsuses olevate noomenite ning adjektiivide ja adverbide märgendamisel. Samuti oli märgendajal raskusi sobiva täpsema märgendi määramisel noomeni ja verbi kategooriates. Nimetatud raskusi mainiti ka CLAWS7 järeltoimetamise juhendis. Lisaks tekkisid märgendajal õppijavigadega seotud raskused. CLAWS7 oluline nõrkus on veel märgendite puudumine relatiivpronoomeni ning samuti sõnade *this* ja *that* pronoomenkasutuse jaoks. Vaatamata nimetatud puudustele saab CLAWS7 märgendajat kasutada eestlaste inglise õppijakeele märgendamiseks.

Võtmesõnad: inglise õppijakeel, TCELE, sõnaliikide märgendamine, automaatse märgendaja vead, korpuslingvistika

Liina Tammekänd (University of Tartu). Her research interests include Estonian learner English, EFL writing and oral narratives.
J. Liivi 4-312, 50409 Tartu, Estonia
liina.tammekand@ut.ee

Reeli Torn-Leesik (University of Tartu). Her research interests include morphosyntax (esp. voice constructions), first and second language acquisition and learner language research.
J. Liivi 4-312, 50409 Tartu, Estonia
reeli.torn-leesik@ut.ee