Vilnius University Press

Contents lists available at Vilnius University Press

Taikomoji kalbotyra, 20: 126–140
https://www.journals.vu.lt/taikomojikalbotyra

eISSN 2029-8935
DOI: https://doi.org/10.15388/Taikalbot.2023.20.9

Automatic part-of-speech tagging of the Tartu Corpus of Estonian Learner English with CLAWS7: impact of learner errors

Liina Tammekänd
University of Tartu, Institute of Foreign Languages and Cultures
liina.tammekand@ut.ee

Reeli Torn-Leesik[1]
University of Tartu, Institute of Foreign Languages and Cultures
reeli.torn-leesik@ut.ee

Abstract. The present paper, which is a continuation of Tammekänd and Torn-Leesik's (2022) study, aims to examine how learner errors affect the CLAWS7 tagger's automated assignment of part-of-speech (POS) tags to a sample of 24,812 words of the Tartu Corpus of Estonian Learner English (TCELE). Learner errors causing tagging errors in the sample were identified, based on which a working error taxonomy was created. The POS-tagged and error-tagged samples were collated and compared to map correlations between learner and tagging errors. Error groups that correlated with significantly increased rates of tagging errors were identified. Possible reasons were suggested to account for the impact of learner errors on the tagger's performance. The CLAWS7 tagger misanalysed only 2.8% of forms representing learners' language errors but assigned wrong tags to every fifth spelling error (22%).

Keywords: *learner English, automatic POS-tagging, learner errors, TCELE, CLAWS7*

Automatinis kalbos dalių žymėjimas (POS) Tartu estų anglų kalbos mokinių tekstyne: mokinių klaidų poveikis CLAWS7 įrankio tikslumui

Santrauka. Pagrindinis šio darbo, kuris yra Tammekändos ir Torn-Leesikos (2022) tyrimo tęsinys, tikslas buvo ištirti mokinių klaidų įtaką automatizuotam kalbos dalių (POS) žymų priskyrimui naudojant CLAWS7 įrankį. Tyrimas paremtas 24 812 žodžių imtimi; duomenys surinkti iš Tartu estų anglų kalbos mokinių tekstyno (TCELE). Duomenims anotuoti buvo naudojamas CLAWS7 žymų rinkinys, kurį sudaro 137 žodžių žymės. Siekiant įvertinti mokinių klaidų įtaką anotavimo tikslumui, detalesnei analizei buvo atrinkti atvejai, kai mokinio klaidos sutapo su automatinio žymėjimo klaidomis šio tyrimo imtyje. Remiantis tyrime nustatytomis klaidomis, sukurta klaidų taksonomija ir klaidų tipų, turinčių įtakos anotavimo rezultatams, klasifikavimo ir analizės sistema. Siekiant ištirti ryšį tarp besimokančiojo klaidų ir automatinio anotavimo klaidų, automatinio žymėjimo klaidos sistemingai lygintos su mokinių klaidomis, o tai padėjo nustatyti sąsajas ir dėsningumus šiuose dviejuose duomenų rinkiniuose. Kitaip tariant, tyrime buvo siekiama nustatyti klaidų

[1] The authors are listed alphabetically.

tipus, kurie gali turėti ženklios įtakos automatinio anotavimo klaidoms. Nurodytos galimos priežastys, paaiškinančios tyrime pastebėtą mokinių kalbos klaidų poveikį anotavimo įrankio tikslumui. Taip pat darbe bandyta paaiškinti pagrindinius veiksnius, galimai lėmusius automatinio kalbos dalių žymėjimo klaidas, susijusias su tekste esančiomis mokinių klaidomis.

Tyrimo rezultatai rodo, kad mokinių kalbos klaidų nulemtos CLAWS7 įrankio klaidos sudaro vos 2,8 % visų atvejų. Tačiau rašybos klaidos daro kur kas didesnį poveikį anotavimo tikslumui – įrankis priskyrė neteisingas žymas net 22 % tokių klaidų. Taigi atliktas darbas atskleidė, jog apskritai mokinių klaidos neturi didelės įtakos CLAWS7 įrankio tikslumui, tačiau rašybos klaidos – turi.

Raktažodžiai: *mokinių anglų kalba, automatinis kalbos dalių žymėjimas (POS), mokinių kalbos klaidos, Tartu estų anglų kalbos mokinių tekstynas (TCELE), CLAWS7*

1. Introduction

Contemporary linguistic studies on L1 often use large collections of data or corpora to test their hypotheses. The same applies to studies on learner language. Learner language, also called interlanguage (Selinker 1972; Selinker, Rutherford 1992; Corder 1981), is a foreign language that the learner is learning and that is not an official language spoken in their home country (Granger 2008: 260). It represents the linguistic system that the learner builds on the basis of learned-language input. Learner language is characterised by its dynamic nature and variation, which reflect the stages of the learner's progress towards achieving target language norms (Ellis 1994: 16).

Whereas earlier research on learner language was often based on data drawn from highly controlled language tests conducted with a small number of learner groups (Granger, Meunier 2015), contemporary research employs learner corpora – electronic collections of language learners' texts (Granger 2008). Such corpora[2] are large in size, provide samples from many learners and, owing to their electronic form, allow instantaneous searches and can be used in different types of studies. The results of learner corpus research help shed light on the characteristics of learner language, contribute to second language acquisition theory in general and to pedagogical methods and tools that are helpful in meeting language learners' needs (Granger 2008).

Only a small number of learner language corpora have been compiled in Estonia to date. These include the two large corpora of Estonian as learner language (the Estonian Interlanguage Corpus of Tallinn University (EIC) and the learner language corpus of the University of Tartu) and a smaller one of learner Spanish (Tartu Learner Corpus of Spanish as a L3+), with Estonian learner English remaining a largely unexplored field. A study that deals with that field was published in 2022 by Tammekänd and Torn-Leesik, who tested the suitability of the automatic CLAWS7 (Constituent Likelihood Automatic Word-tagging System) tagger for tagging Estonian learner English by assessing the tagger's error rate.

The present paper continues the study conducted by Tammekänd and Torn-Leesik (2022). Its aim is to determine the types of learner errors that have a marked impact on the performance of the CLAWS tagging system when tagging Estonian learner English.

The paper is divided into two main parts. The first one provides an overview of automatic part-of-speech (POS) tagging and POS taggers, introduces Tammekänd and Torn-Leesik's (2022) study, explains the authors' use of the terms 'error' and 'mistake' and discusses existing research on learner

[2] e.g., The Longman Learners' Corpus of 10m words or The Cambridge Learner Corpus of 50m words.

errors that influence the automatic part-of-speech tagging process. The second part analyses learner errors in the Tartu Corpus of Estonian Learner English (TCELE) and their impact on the CLAWS7 tagger's performance when tagging Estonian learner English.

2. POS-tagging and learner errors

2.1 Automatic POS-tagging and POS-taggers

Corpus annotation involves adding interpretative, linguistic data to the corpus text (Leech 2013). Linguistic annotation such as POS-tagging, as well as syntactic, semantic and discourse annotation, allows information to be extracted that would otherwise be unobtainable from the corpus. For instance, finding reduced relative clauses in a large learner corpus without linguistic annotation would be very difficult as the construction is characterised by a null element, i.e., the absent relative pronoun (Kübler, Zinsmeister 2015: 21). In contrast, appropriate linguistic annotation enables researchers to retrieve a wide range of linguistic phenomena without much effort.

There are three commonly used types of POS-taggers. Rule-based POS-taggers employ hand-written disambiguation rules to assign POS-tags to words. Examples of these taggers include TAGGIT (Green, Rubin 1971), TOSCA (Oosdijk 1991), Constraint Grammars and EngCG (Voutilainen 1994, Karlsson et al. 1995). Stochastic taggers, such as the CLAWS tagger (Garside et al. 1987), rely on training from pre-tagged corpora to calculate the probability of a word having a specific tag in a given context. Hybrid taggers combine both manual disambiguation rules and probability calculations, with Brill (1992) being an example.

POS-tagging forms the basis for other types of corpus annotation such as parsing or semantic tagging. POS-tagging is mostly automatic, which means that a computer program (the tagger) assigns a part-of-speech tag to each word in the corpus without additional user input (Gries, Berez 2017; van Rooy 2015; Jurafsky, Martin 2008). POS-tagging takes place in three stages: first, the tagger divides the text into tokens; second, it finds possible tags for the words from the lexicon – or, if the word does not have a lexicon entry, the tagger attempts to guess which POS category it belongs to; finally, the tagger disambiguates the assigned POS-tags using contextual and statistical information (Voutilainen 1999, 2003).

For taggers of English, the final stage appears to be the most problematic (Voutilainen 2003) since many frequently used English words are ambiguous. For example, the tagger may have problems disambiguating prepositions, particles and adverbs. Also, participles and adjectives, as well as common nouns, proper nouns and adjectives when they appear as noun (phrase) modifiers may pose problems for the tagger (Jurafsky, Martin 2008). The accuracy of POS tagging depends on the morphological complexity of the corpus language, corpus size, the size of the tag set and the nature of the training corpus (Griez, Berez 2017).

The tagging of learner language may pose additional problems to taggers as learner language features structures and words that the tagger may not have encountered in the training corpus, which usually is a collection of native language texts (van Rooy 2015). Nagata et al. (2018) highlight three main issues a POS tagger is likely to run up against when tagging learner language. First, learner language may include unknown forms resulting from spelling or grammar mistakes that make the underlying word impossible for the tagger to recognise. Second, learner language may have different POS distributions compared to the training corpus. For instance, in newspaper texts, which are commonly used in training corpora, the word *concentrate* is usually a noun (e.g., *orange juice concentrate*), but in academic learner English, it is often a verb (e.g., *concentrate on sth*) (Chodorow and Leacock 2002). Third, learner language has characteristic POS-sequences (Nagata et al. 2018). Aarts and Granger (1998) observed 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 may have a negative impact on automatic POS-tagging.

2.2 Tagging Estonian learner English with CLAWS7

As there are no separate automatic POS-taggers specifically designed for learner English, researchers have no other option but to utilise POS-taggers trained on native English data for tagging learner English. When selecting a POS-tagger for a learner English corpus, the first step is to evaluate the performance of the chosen tagger. In Tammekänd and Torn-Leesik (2022), the authors chose to test the suitability of the CLAWS7 automatic POS-tagging system for tagging the Estonian learner English corpus TCELE (for a more detailed description of TCELE, see Section 3). CLAWS7 was chosen because of its availability as a freely accessible tool and its convenient online user interface.

In Tammekänd and Torn-Leesik (2022), manually and automatically tagged samples of TCELE were compared, the tagger error rate was calculated, and possible reasons for tagger errors were investigated. The analysis showed that the CLAWS7 tagger had problems assigning correct tags to determiners, adverbs, general adverbs, and singular common nouns. The tagger successfully assigned general noun and verb tags but experienced problems when attempting to analyse words at a more granular level. Also, the tagger had problems differentiating between nouns and adverbs, as well as between conjunctions and adverbs. The analysis of the results also highlighted a shortcoming of the C7 tag set. For instance, the set does not have a separate tag for *this/that* in the (relative) pronominal function, which, in turn, makes it problematic for studying relative clause constructions in Estonian learner English.

The results of the study showed that the CLAWS7 tagger exhibited an error rate of 4.01%, consistent with previous findings in automatic POS-tagging of learner English (van Rooy 2015, van Rooy and Schafer 2002, de Haan 2000). Of the tagger's errors, 0.56% were attributed to learner errors. In the current study, the authors aimed to investigate the specific types of learner errors that pose the greatest challenges for the CLAWS automatic tagging system when tagging Estonian learner English.

2.3 Errors and mistakes

Lennon (1991) points out that providing an unambiguous definition of 'error' is a challenging task, as can be seen from the range of formulations offered by different scholars. The definition of 'error' that is probably the broadest – and colourfully captures the phenomenon in an astutely brief turn of the phrase – is suggested by James (2013: 1), who considers it an 'unsuccessful bit of language'. For him, 'error' is a relative term as it only becomes such in relation to other forms or to the rules that it violates. A similar but more prosaic formulation is proposed by Ellis (1994: 51) – 'deviation from the norms of the target language'.

An example of a more specific definition is the one advanced by Corder (1967, 1981), who differentiates between learner errors and learner mistakes. The former reflect a failure of competence, while the latter are a failure of performance. Errors demonstrate a systematic lack of knowledge, which means that the learner is not aware of the error and is thus unable to correct it even if the error is pointed out to them (see also Hymes 1972). Mistakes, on the other hand, do not reflect a deficit of knowledge – rather, they are caused by some other circumstance and can be self-corrected by the learner (Pfingsthorn 2011). Ellis (1994: 47) compares learner mistakes with native speakers' 'slips of the tongue'.

Corder (1971, 1981) also distinguishes between overt and covert errors. Overt errors are digressions from form and are easy to identify. Covert errors occur when the utterance is "superficially 'well-formed'" (Corder 1981: 21) but does not have the meaning the learner intended to convey. Covert errors are difficult to identify because identification involves a subjective evaluation on the part of the researcher, who makes stylistic rather than grammatical judgements (Ellis, Barkhuizen 2005). In other words, overt errors are related to grammaticality, while covert errors are associated with acceptability (Ellis, Barkhuizen 2005).

The present paper follows Ellis' definition, approaching errors as deviations from the target norm. Since the authors are not interested in differentiating between student errors and mistakes and focus on the tagger's performance instead, they have chosen a framework that allows a straightforward assessment of how deviations 'from the norms of the target language' affect the tagger.

2.4 Earlier research on learner errors influencing the automatic POS-tagging process

Researchers (de Haan 2002; van Rooy, Schäfer 2002; Mizumoto, Nagata 2017; Nagata et al. 2018) divide errors in texts produced by language learners into two broad categories: spelling errors and language errors. Spelling errors include typing, spacing and capitalisation errors. While typing errors are obvious keyboard mistakes, spacing errors either merge the words that need to stand separately (e.g., *bankcard*) or split the words that need to appear as a single form (e.g., *can not*). Capitalisation errors occur when a word that should be capitalised is not capitalised and vice versa.

Language errors, on the other hand, involve the language learner's morphological, syntactic and lexical errors. Such errors are more diverse across different studies because their nature depends, among other things, on the learner's L1. For instance, in addition to spelling errors, De Haan (2000) discusses word transfer, verb morphology and hypercorrection errors as well as errors specifically related to L1 spelling, morphology, lexis and pronunciation. For Van Rooy and Schäfer (2002), the category of language errors includes errors of articles, prepositions, agreement, lexical choice, clause patterns, pronouns, infinitives and errors of omission. In addition to these, Abdul Aziz and Mohd Don (2019) have pointed out word order, word form and overgeneralisation errors as those potentially specific to the learner's L1.

POS-tagging errors may occur when tagging unknown as well as known words (Mizumoto, Nagata 2017, Nagata et al. 2018). The former result from spelling, spacing or capitalisation errors, while the latter represent language errors. Nagata et al. (2018) also note that foreign words and words not present in the tagger's lexicon can be considered instances of unknown words and cause tagging problems. Researchers (van Rooy, Schäfter 2002; Mizumoto, Nagata 2017; Nagata et al. 2018) agree that spelling errors greatly affect the accuracy of correct POS-tagging. When evaluating the performance of three taggers when automatically POS-tagging the Tswana Learner English Corpus, Van Rooy and Schäfer (2002) found that correcting spelling errors significantly improved the taggers' performance. However, the language errors in their spelling-corrected corpus still influenced the correct assignment of POS-tags. Van Rooy and Schäfer (2002) noted that not all language errors lead to tagging errors. For instance, the use of a wrong article or preposition still received a correct tag respectively, while verb conjugation errors posed serious problems to the taggers in their study.

3. Material and methods

The present study is based on the Tartu Corpus of Estonian Learner English (TCELE) – a written learner English corpus compiled at the Department of English Studies of the University of Tartu (Estonia). The corpus consists of short essays written as part of the entrance examination for the BA programme in English language and literature. The essays are modelled on a short journalistic text and as a rule run to 250–300 words. Writing the essay is timed, and its assumed CEFR level is B2. The candidates whose score falls below a certain threshold in the first part of the examination (a test of the examinee's general lexico-grammatical competence) are not admitted to the second part, which tests reading and writing skills. This means that only the most linguistically competent candidates progress to the essay stage.

The main goal of this research, which builds upon Tammekänd and Torn-Leesik's (2022) study, was to examine how learner errors influence the automatic POS-tagging assignment produced by the CLAWS7 tagger. As mentioned in Section 2.3, Ellis's (1994) definition of error as 'deviation from the norms of the target language' was chosen as the working definition for the purposes of this paper. British English and American English varieties were taken as the 'norm' or 'target language' in the sense of Ellis's error definition as the former is generally taught at Estonian schools, while the latter is prevalent in mass and social media. Thus, it can be assumed that the Estonian English learner has the most contact with these varieties. The focus was on overt learner errors as a possible influence on the tagger's performance.

Having the above in mind, the following analytical steps were taken:

1. A TCELE sample of 24,812 words (92 essays) was POS-tagged using the CLAWS7 tag set.
2. Learner errors in the sample were manually identified by the authors of the paper.
3. Based on previous research (de Haan 2002; van Rooy, Schäfer 2002; Mizumoto, Nagata 2017; Nagata et al. 2018), errors were classified into two main groups, and a working error taxonomy was created (see Section 3.1).
4. POS-tagged and error-tagged samples were collated and compared to map correlations between learner errors and tagging errors.
5. Learner error taxons that correlated with a notable increase in the tagger's error rate were identified.
6. Possible reasons were suggested to explain the impact of learner errors on tagging errors.

3.1 Learner error taxonomy

Following the approach taken in several studies (de Haan 2002; van Rooy, Schäfer 2002; Mizumoto, Nagata 2017; Nagata et al. 2018), errors in the TCELE sample were divided into two broad categories: spelling errors and language errors. Both categories can be subdivided further.

Spelling errors can be divided into subcategories reflecting erroneous use of the hyphen, typing slips, omissions or insertions of spaces between words, and capitalisation errors (see Table 1). In the sample used in this study, as in the previous studies (de Haan 2002; van Rooy, Schäfer 2002; Mizumoto, Nagata 2017; Nagata et al. 2018), typing errors resulted in nonwords (words not present in the language (e.g., *litertaure*) as well as real words that are listed in the lexicon but do not fit the context or sentence they appear in (e.g., *it* instead of *in*). Capitalisation errors occur when a word that should be capitalised is not capitalised and vice versa. A separate subcategory was created for instances where two or more different spelling errors occurred in a single word (e.g., *id* for *I'd*).

Table 1. Subcategories of spelling errors in the TCELE sample

Subcategory	Explanation	Examples
Omission of a hyphen		Whether it is an artistic work of fiction or a **real life** experience, … (correct: ‘a real-life experience’).
Extra hyphen		When studying literature **through-out** the years, … (correct: ‘throughout’)?
Nonword	Wrong spelling creates a word that does not exist	**Litertaure** has been around for hundreds of years (correct: ‘literature’).
Real word	Wrong spelling creates a word that exists but is wrong in the given context (homonyms)	It also played a huge role **it** their entertainment (correct: ‘in’)
Space merging	Two words written together	/… the negativity towards it stems from not getting to do it out of **freewill**, … (correct: ‘free will’).
Extra space		… becoming **book worms** (correct: ‘bookworms’**).**
Capitalisation	A word that needs to be capitalised is not and vice versa	... but **i** believe that the negativity towards_it … (correct: ‘I’)
Compound spelling error	Two or more different spelling errors in one word	And **id** say the general consecion on the role of literature has stayed the same. (correct: ‘I’d’)

Language errors include instances of morphological, syntactic, and lexical errors. The subcategories identified in the TCELE sample are provided in Table 2. Verb errors are subcategorised for category errors consisting in the wrong choice of tense, agreement pattern, mood or voice, as well as pattern errors involving the wrong choice of verb form in the subcategorisation frame (for instance, the infinitive is used instead of the participle; see the example in Table 2). Errors with nouns are divided into two subcategories: (1) instances where the student has problems with the number category of the noun (singular vs plural) or with the use of uncountable nouns, and (2) instances of wrong use of the genitive construction (since the number of such instances was relatively large, the authors decided to treat it as a separate subcategory). The data also allowed for subcategories focussing on adjectives, articles, quantifiers, pronouns, prepositions and conjunctions. The errors in these consist in the use of incorrect forms of the intended word (e.g., of the comparative degree of an adjective) or various omissions or insertions (e.g., of articles or prepositions).

Table 2. Subcategories of language errors in the TCELE sample

Subcategory	Explanation	Examples
Verb’s grammatical category	wrong tense, agreement, mood or voice	In the past literature **has been regarded** … (correct: ‘was regarded’)
Verb pattern errors	wrong verbal form (infinitive or participle) in the verb pattern	/.../ I definitely see the creative community **be** more active, … (correct: ‘being’).
Genitive construction errors	missing apostrophe in the s-genitive	Which in **todays** currency is about 40$, … (correct: ‘in today’s currency’).
Noun phrase errors	number, countable/ uncountable nouns	… the basic **knowledges** among us … (correct: ‘knowledge’)
Quantifier errors	wrong quantifier	There are **less and less** libraries … (correct: ‘fewer and fewer’)
Article errors	wrong, missing or superfluous article	I am of **an** opinion (correct: ‘of the opinion’)
Pronoun errors	wrong, missing or superfluous pronoun	Not too long ago, there was a time **where** most people couldn‘t even read. (correct: ‘when’)
Adjective and adverb errors	wrong comparative forms, wrong adjective/adverb forms	… publishing a book has never been **more easier**. (correct: ‘easier’).

Subcategory	Explanation	Examples
Preposition errors	wrong, missing or extra preposition	… to build a stronger foundation **to** the world we live in now (correct: 'for') In his text he argued **for** that literature is more important … (correct: 'argued that')
Conjunction errors	wrong or missing conjunction	… that is a huge reason for the change in attitude towards **literature, technology** (correct: 'literature and technology').
Sentence structure errors	word order errors, comma splices, fragments, faulty parallelism, missing subjects and objects	Though literature has moved on from being physical to being more online, people **tend to not have** as much interest in it, as it had a hundred years ago (correct: 'tend not to have').
Derivation errors	word formation errors	… was perceived as **merely** entertainment. (correct: mere)
Lexical choice errors	collocation and idiomaticity errors	… can be seen in the numbers of people who have a **literary degree**. (correct: 'literature degree')
Miscellaneous errors	instances that did not fit in any of the above subcategories	… 100 years ago **the of** literature was considered a universal language … (correct: 'ago literature')

The analysed data also include errors in sentence structure and lexical errors. The subcategory of lexical errors involves problems with word derivation and collocation patterns. Sentence structure errors, in turn, reflect the learner's problems with word order and clause combination. There were also instances of errors that did not fit in any of the above categories and were thus classed as miscellaneous.

The error taxonomy that emerged from the analysis serves as a tool for evaluating various aspects of the tagger's performance and is not treated as a basis for error annotation.

4. Results and discussion

As mentioned in Section 3, a TCELE sample of 24,812 words (92 essays) was POS-tagged using CLAWS7 tag set and then manually error-tagged by the authors. The total number of errors made by the learners in the sample was 678, of which 560 were language errors and 118 spelling errors (see Table 3).

Table 3. Categories and number of learner errors in the TCELE sample

Categories	No of errors
Language errors	560
Spelling errors	118
TOTAL	678

The analysis focuses on the co-incidence of learner (language and spelling errors) and tagging errors (see Table 4). In this study, tagging errors were deemed to have been caused by learner errors when the tagger assigned the wrong POS-tag to the learner's erroneous form. If the learner's form is a real English word (although incorrect in the context) and the tagger tagged it as such, this is not considered a tagging error. Consider example (1), where the learner has omitted the apostrophe in the genitive construction required by the context (*person's*) and written the plural form of the noun (*persons*) instead. Although the learner's error leads the tagger to choose the NN2 tag, its choice is not wrong as such – *persons* is a noun and the *s*-suffix signals that the tagger is dealing with a plural noun; thus, the tagger has correctly identified the form it was presented with.

(1) *I_PPIS1 think_VV0 that_DD1 literature_NN1 is_VBZ very_RG important_JJ to_II a_AT1* ***persons_NN2*** *life_NN1 ,_, because_CS literature_NN1 nurtures_NN2 and_CC helps_VVZ our_APPGE creativity_NN1 flow_VVI .*

A similar situation occurs when an apostrophe is inserted in the possessive determiner *its*, as illustrated in (2). Here the learner's mistaken presentation (*it's*) causes the tagger to classify the contracted form as a pronoun followed by the present tense of *be*. Incorrect as this may be in the context, the form that the tagger sees is a real English structure and the tagger recognises it as such.

(2) *Nowadays_RT the_AT study_NN1 of_IO literature_NN1 has_VHZ once_RR21 again_RR22 reclaimed_VVN* ***it_PPH1 'š_VBZ*** *rightful_JJ place_NN1 in_II both_DB2 academia_NN1 and_CC with_IW the_AT general_JJ public_NN1 ._.*

The percentage of language errors that defied the tagger's analysis and were attributed an incorrect POS tag was relatively low (38 of 560 errors, or 2.8%). As to spelling errors, every fifth such error (28 of 118, or 22%) resulted in a word that was wrongly tagged. This suggests that spelling errors – which, in the study, were associated with triple the rate of tagging errors caused by simple language errors on the part of the learner – appear to be considerably more problematic for the tagger. The results confirm those of Mizumoto and Nagata's study (2017), which claims that spelling errors pose a major difficulty in automatic POS-tagging.

Table 4. Learner errors correlated to tagging errors

	Total no of errors	**No of tagging errors correlated to learner errors**	**% of tagging errors correlated to learner errors**
Language errors			
Verb's grammatical category	75	0	0%
Verb pattern	17	0	0%
Genitive	25	0	0%
Noun phrase	4	0	0%
Quantifier	10	0	0%
Article	105	1	0.95%
Pronouns	12	0	0%
Adjective and adverb	7	1	14.3%
Preposition	80	0	0%
Conjunction	13	0	0%
Sentence structure	112	0	0%
Derivation	6	1	17%
Lexical choice	37	2	5.4%
Miscellaneous	57	11	19%
Language errors total	***560***	***16***	***2.8%***
Spelling errors			
Omitted hyphen	31	12	38.7%
Extra hyphen	3	0	0%
Nonword	42	5	12%
Real word	12	0	0%
Space merging	6	2	33.3%
Extra space	12	2	16.7%
Capitalisation	11	5	45.5%
Compound spelling errors	1	0	0%
Spelling errors total	***118***	***26***	***22%***

Examining language and spelling errors separately, it can be observed that within the category of language errors, the most frequent subcategories were sentence structure (112), the use of articles (105) and prepositions (80). Additionally, errors were noted in verb categories, including tense, mood, number and voice (75). Although the numbers of errors in these subcategories are relatively high, the resulting forms predominantly still received the correct POS-tag. For instance, in example (3), the learner uses the wrong participle form of the verb *lead* (the correct form would have been *led*), yet the tagger is able to assign it the correct tag (VVN), marking it as the past participle of the verb. In such cases the tagger appears to make its decision based on probabilities and the grammatical context.

(3) *This_DD1 has_VHZ* ***lead_VVN*** *to_II the_AT downfall_NN1 of_IO the_AT quality_NN1 of_IO literature_NN1 nowadays_RT.*

In the category of spelling errors, nonwords (42) and omitted hyphens (31) were the most frequent correlates of tagging errors. Although both numbers are relatively high among the relevant subcategories, the missing hyphen caused the tagger to return a markedly higher number of contextually incorrect tags. As illustrated in (4), the hyphen's omission in the word *real-life* causes the tagger to assign the word two separate tags, JJ (adjective) and NN1 (noun) instead of a single one (JJ).

(4) *Whether_CSW it_PPH1 is_VBZ an_AT1 artistic_JJ work_NN1 of_IO fiction_NN1 or_CC a_AT1* ***real_JJ life_NN1*** *experience_NN1.*

In the case of nonwords, only 5 out of 42 instances led to a tagging error (12%). For instance, in example (5), the spelling mistake results in the nonword *litertaure*; however, it is likely that the similarity to the real word *literature* and the probable nominal slot in the sentence helps the tagger to assign the contextually correct tag to the learner's form.

(5) ***Litertaure_NP1*** *has_VHZ been_VBN around_RP for_IF hundreds_NNO2 of_IO years_NNT2 ._.*

Although the number of space merger errors is small (6), every third one (33.3%) correlates with a tagging error. Example (6) illustrates one instance of the resulting misclassification. The learner's presentation of the words *at least* as a single form leads the tagger to analyse these – incorrectly – as a unit, which it then classifies as a noun.

(6) *Especially_RR books_VVZ that_CST are_VBR* ***atleast_NN1*** *one_MC1 hundred_NNO years_NNT2 old_JJ*

The learners in the sample made 11 capitalisation errors, 5 of which affected the tagger's recognition of the resulting form. For instance, in example (7) the learner has written the 1st person pronoun *I* as a lowercase letter, which results in the tagger classifying it as a singular cardinal number (MC1). Unlike in example (3) above, it seems that here the tagger does not base its decision on the grammatical context, instead relying on the spelling of the form.

(7) *of_IO hatred_NN1 for_IF having_VHG to_TO study_VVI the_AT artform_NN1 ,_, but_CCB i_MC1 believe_VV0 that_CST the_AT negativity_NN1 towards_II it_PPH1*

As noted in Section 2.2, Tammekänd and Torn-Leesik's (2022) study showed that the CLAWS7 tagger's low error rate (4.01%) makes it a suitable tool for tagging Estonian learner English. However, the results of the present study show that learners' spelling errors are likely to have a marked impact on the tagger's performance. When evaluating the performance of three different taggers (TOSCA_ICLE, Brill tagger, CLAWS) on the Tswana Learner English Corpus, Van Rooy and Schäfer (2002) found that editing out spelling errors improved the taggers' performance. Thus, editing the TCELE sample would probably further reduce the tagger's overall error rate.

5. Concluding remarks

The aim of this study was to identify learner errors that are the likely cause of tagging errors during automatic POS-tagging of Estonian learner English. For that, a 24,812-word sample of the Tartu Corpus of Estonian Learner English (TCELE) was, first, automatically POS-tagged using the automatic CLAWS7 POS-tagging system. Then, the learners' errors were identified by the authors. Similarly to the studies reported by de Haan (2002), van Rooy and Schäfer (2002), Mizumoto and Nagata (2017), and Nagata et al. (2018), the data of this study allowed learner errors to be classified into two major groups – language errors and spelling errors. Both were then subcategorised (see Section 3.1). The POS-tagged and error-tagged samples were collated and compared to identify the error taxons that increased the likelihood of tagging errors.

The total number of learner errors in the sample was 678, of which 560 were language errors and 118 spelling errors. Only 16 (2.8%) of the 560 language errors appearing in learners' texts were misanalysed by the tagger. In contrast, the tagger was misled by 26 (22%) of the 118 spelling errors. The study highlighted that while the CLAWS7 tagger has shown low error rates in tagging Estonian learner English, learners' spelling errors impact the tagger's performance.

References

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

Abdul Aziz, R., Mohd Don, Z. 2019. *Tagging L2 Writing: Learner Errors and the Performance of an Automated Part-of-Speech Tagger*. GEMA Online Journal of Language Studies 19(3), 140–155. https://doi.org/10.17576/gema-2019-1903-09

Brill, E. 1992. A simple rule-based part of speech tagger. Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155. doi:10.3115/974499.974526

Chodorow M., Leacock C. 2002. Techniques for detecting syntactic errors in text. *IEICE Technical Report* (TL2002-39), 37–41.

Corder, P. 1967. The Significance of Learner Errors. *International Review of Applied Linguistics* 5, 161–170. https://dx.doi.org/10.1515/iral.1967.5.1-4.161

Corder, P. 1971. Idiosyncratic dialects and error analysis. *IRAL: International Review of Applied Linguistics in Language Teaching*, 9(2), 147–160. https://doi.org/10.1515/iral.1971.9.2.147

Corder, P. 1981. *Error Analysis and Interlanguage*. Oxford: Oxford University Press. https://doi.org/10.3138/cmlr.40.4.649

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

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

Ellis, R., Barkhuizen, G. 2005. *Analysinig Learner Language*. Oxford University Press: Oxford. https://doi.org/10.1093/ijl/eck003

Garside, R., Leech, G., Sampson, G. 1987. The Computational Analysis of English: A Corpus-based Approach. Harlow: Longman.

Garside, R. 1996. The robust tagging of unrestricted text: the BNC experience. *Using corpora for language research: Studies in the Honour of Geoffrey Leech*. J. Thomas and M. Short (eds.), 167–180. London: Longman.

Granger, S. 2008. Learner corpora. *Corpus Linguistics. An International Handbook. Vol 1.* A. Lüdeling, M. Kytö (eds.). Berlin–New York: Walter de Gruyter, 259–275. https://doi.org/10.1002/9781405198431.wbeal0669.pub2

Granger, S., Gilquin, G., Meunier, F, 2015. Introduction. Learner corpus research: Past, present and future. *The Cambridge Handbook of Learner Corpus Research.* S. Granger, G. Gilquin, F. Meunier (eds.). Cambridge: Cambridge University Press, 1−5. https://doi.org/10.1017/CBO9781139649414

Greene, B. B., Rubin, G. M. 1971. Automatic Grammatical Tagging of English. Department of Linguistics, Brown University.

Gries, S. T., Berez, A. L. 2017. Linguistic annotation in/for corpus linguistics. *Handbook of Linguistic Annotation.* N. Ide, J. Pustejovsky (eds.). Dordrecht: Springer, 379–410. https://doi.org/10.1007/978-94-024-0881-2_15

Hymes, D. H. 1972. On Communicative Competence. *Sociolinguistics: Selected Readings.* J. B. Pride, J. Holmes (eds.). Harmondsworth: Penguin, 269–293.

James, C. 2013 *Errors in Language Learning and Use. Exploring Error Analysis*. Routledge: London and New York. https://doi.org/10.4324/9781315842912

Jurafsky, D., Martin, J. H. 2008. *Speech and Language Processing*. 2nd ed. Upper Saddle River: Prentice Hall.

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A (eds.). 1995. *Constraint Grammar: A Language-Independent Framework for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin / New York

Kübler, S., Zinsmeister, H. 2015. *Corpus linguistics and linguistically annotated corpora*. Bloomsbury Academic.

Leech, G. 2013. Introducing corpus annotation. *Corpus Annotation. Linguistic Information from Computer Text Corpora.* R. Garside, G. Leech, T. McEnery (eds.), London: Routledge, 1–18.

Lennon, P. 1991. Error: Some Problems of Definition, Identification, and Distinction. *Applied Linguistics* 12 (2), 180–196. https://doi.org/10.1093/applin/12.2.180

Mizumoto, T., Nagata, R. 2017. Analyzing the Impact of Spelling Errors on POS-Tagging and Chunking in Learner English. *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications*, 54–58. https://aclanthology.org/W17-5909 (31.01.2023)

Nagata, R., Mizumoto, T., Kikuchi, Y., Kawasaki, Y., Funakoshi, K. 2018. A POS tagging model designed for learner English. *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text.* W. Xu, A. Ritter, T. Baldwin, A. Rahimi (eds.). Brussels: Association for Computational Linguistics, 39–48. https://doi.org/10.18653/v1/W18-6106

Oostdijk, N. 1991. Corpus linguistics and the automatic analysis of English. Amsterdam: Rodopi.

Pfingsthorn, J. 2011. *Variability in Learner Errors as a Reflection of the CLT Paradigm Shift*. Peter Lang Edition: Frankfurt am Main. https://doi.org/10.3726/978-3-653-02772-3

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

Selinker, L., Rutherford, W. E. 1992. *Rediscovering Interlanguage*. Routledge: London. https://doi.org/10.4324/9781315845685

Tammekänd, Liina; Reeli Torn-Leesik. 2022. POS-tagging Tartu Corpus of Estonian Learner English with CLAWS7. *Estonian Papers in Applied Linguistics* 18, 263–278. https://doi:10.5128/ERYa18.15

UCREL Team. 1996. *A Post-Editor's Guide to Claws7 Tagging*. http://www.natcorp.ox.ac.uk/docs/claws7.html (15.01.2023).

UCREL. 2014. *CLAWS part-of-speech tagger for English*. available at https://ucrel.lancs.ac.uk/claws/

van Rooy, Bertus 2015. Annotating learner corpora. *The Cambridge Handbook of Learner Corpus Research.* S. Granger, G. Gilquin, F. Meunier (eds.). Cambridge: Cambridge University Press, 79–106. https://doi.org/10.1017/CBO9781139649414.005.

van Rooy, B., Schäfer, L. 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.

Voutilainen, A. 1994. *Designing a Parsing Grammar*. University of Helsinki, Department of General Linguistics, Helsinki

Voutilainen, A. 1999. A Short History of Tagging. *Syntactic Wordclass Tagging. Text, Speech and Language Technology, vol 9*. H. van Halteren (ed.). Springer, Dordrecht. 9–21. https://doi.org/10.1007/978-94-015-9273-4_2.

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

Appendix:

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	prep. adverb, particle (e.g., about, in)
RPK	prep. adv., 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)