Comparative Analysis of Automatic POS Taggers Applied to German Learner Texts

Irina Kotiurova
Petrozavodsk State University
Petrozavodsk, Russia
koturova@petrsu.ru

Polina Trenina
Petrozavodsk State University
Petrozavodsk, Russia
trenina_p@mail.ru

***Abstract*—The process of assigning morpho-syntactic categories of each element in the sentence including punctuation marks in a text document according to the context is called Part of Speech (POS) tagging. The article presents the analysis of testing and comparison of five part-of-speech taggers: CoreNLP, spaCy, TextBlob, RFTagger and TreeTagger, based on the texts from the annotated learner corpus of Petrozavodsk State University (PACT, Petrozavodsk Annotated Corpus of Texts). All these tools are applicable to the German language; however, learner texts have their own characteristics, primarily associated with a large number of errors. The problem of scientific research is in finding out how traditional instruments will cope with the task of automatic annotation when they come across a large number of grammatical, lexical and spelling mistakes. The conclusions were drawn about the frequency of errors in the part-of-speech identification, the tagging quality, weaknesses and strengths of each tagger.**

I. Introduction

In today's world many things are automated including the processes of thinking. Linguistics is also moving along the path of processing and analyzing big data, leaving the search and research material collecting made manually in the past. Corpus linguistics, which appeared just a few decades ago, has become one of the most promising, rapidly developing areas of science, opening up enormous opportunities, such as authorship identification of anonymous texts [1], data-driven learning [2], automated generation of test questions [3] and other fields of Natural Language Processing (NLP) [4].

Smart services rapidly come to our lives for assisting people in various domains, including education. Advanced digital tools can be constructed using the progress in Artificial Intelligence (AI) and Internet of Things (IoT) [5].

The importance of Corpus technologies in linguistics increases fast, developing both extensively, in breadth, i.e. spreading to ever wider spheres, including pedagogy and prediction processes, and intensively, in depth, i.e. improving the quality and analysis capabilities of corpus data. The first corpus was Brown Corpus, created in 1961 in the USA [6]. The opened up prospects were appreciated in other countries that led to the creation of corpora in other languages showing high performance in the volume and language representation. The effectiveness of their use has encouraged the appearance of numerous linguistic corpora, both large and small.

When starting to create a linguistic corpus, developers have to solve many conceptually important issues; one of them is the question of choosing a part of speech tagger (POS-tagger). The task of POS annotation includes giving a tag with the corresponding name of the part of speech to each token in the text [7]. At present there are various taggers available showing themselves successfully to a different extent depending on the language and types of text [8], [9], [10], [11].

One of the types of linguistic corpora is the learner corpus, its teaching and research possibilities are difficult to overestimate [12]. At Petrozavodsk State University the work on creating a corpus of learner texts (PACT - Petrozavodsk annotated learner corpus) in German is in progress. The server part of the corpus was developed with the aid of the language (ver. 3.8), storage of data on the users and works is ensured through the database SQLite 3. The interaction of the customer and server part is ensured by the Ajax и CGI technologies. For the corpus PACT it was necessary to choose one of five well-known taggers: TreeTagger, RFTagger, spaCy, CoreNLP, TextBlob.

These 5 taggers were selected since they are the most well-known taggers for the German language, and their libraries are accessible for common use. All these tools are applicable to the German language; however, learner texts have their own characteristics, primarily associated with a large number of errors. For instance, the noun spelling with a lowercase letter instead of a capital letter as well as mistakes in the grammatical forms may be critical for some POS taggers [13].

Thus, the problem of scientific research is in finding out how traditional instruments will cope with the task of automatic annotation when they come across a large number of grammatical, lexical and spelling mistakes. Solution of this problem is urgent for the creators of German language learner text corpora, since the quality of tagging is directly connected with any serious application in different fields of Natural Language Processing (NLP). Good quality tagging of corpus texts in particular is necessary for creating digital educational tools on the basis of this corpus, such as error correction, Test Maker, automated essay evaluation etc., including the use of Artificial Intelligence [14], [15], [16].

Similar tasks are investigated by other researchers including Bertus van Rooy and Lande Schäfer [13], Tomoya Mizumoto and Ryo Nagata [17], Markus Dickinson and Marwa Ragheb [18] and many others [19], [20], [21]. Thus, R.Nagata and A.Kawai [20] show, that "the discrepancy

between a POS tagger and its target text often results in POS-tagging errors, which in turn leads to performance degradation in related tasks“. The authors describe that in the article “A POS Tagging Model Designed for English Learners” use of Stanford CoreNLP 3.8.0 to tag the learner of English with the purpose to show, that „it is only natural that a POS tagger for canonical English should make errors as in this example because they do not simply assume erroneous or unnatural inputs” [21]. The researchers discuss potential causes of POS-tagging errors in English learners and describe how deep neural models are particularly suitable for the adaptation of POS tagging to the English learner to “reduce their influence and thus contribute to achieving better performance in the related tasks”. These works, just as the works of other authors, even though they concern our topic, could not help us in selecting the most suitable POS-tagger for the German learner corpus, since these research works dealt only with English.

The rest of the paper is organized as follows. Section II describes the methodology, which we used for choosing the automatic POS-tagger for our learner corpus in German. Section III shows experimental facts on the tagging by each of the five tools and make specific findings about the quality of these taggers applied to German learner texts. Section IV summarizes the results of our experimental study. The Conclusion briefly emphasizes the main result, its significance for linguists, as well as plans for further development of this area of research.

II. Methodology

It was necessary to conduct a study, where TreeTagger, RFTagger, spaCy (version 2.3.0), CoreNLP and TextBlob were applied every time to 35 student texts selected for this experiment using the continuous sampling method. In these texts, about 13% of tokens had error tags, at the same time, the distribution of errors by class indicates the overwhelming predominance of grammatical errors. Fig. 1 shows the numbers of enlarged error classes (detailed error classification in the PACT corpus has 90 categories):

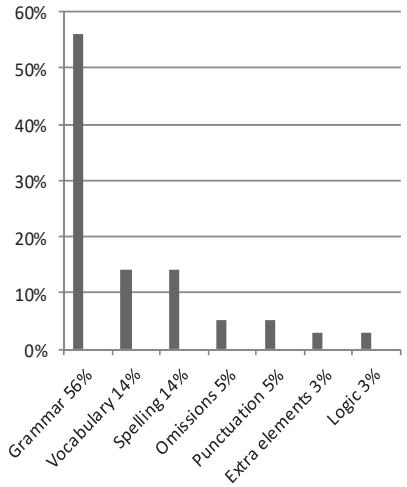

Fig. 1. Evaluation of error distribution

Thus, five POS-taggers were applied to 35 student texts with a total volume of about 10,000 tokens in order to compare the quality of the tagging. Such a volume of identical material was sufficient to compare the efficiency of taggers and choose the one which was applied later to the entire corpus of learners' texts.

The total tokens number is named approximately not by chance. These taggers carrying out the part-of-speech identification of lexical units have a tokenization function as well, which precedes the definition of the part of speech of each selected token. At this stage significant differences are already observed.

So, spaCy divided the studied volume of student texts by 10277 tokens, RFTagger - by 10148 tokens, TextBlob - by 10272, TreeTagger - by 10277, and CoreNLP - by 10373 (Fig.2).

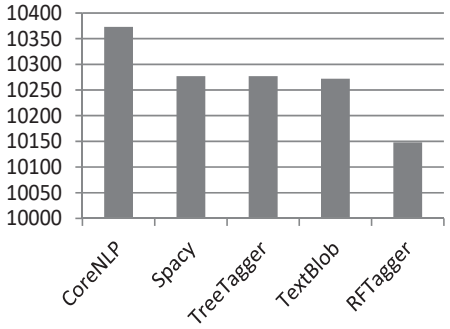

Fig. 2. Number of tokens in different taggers

The difference in the number of tokens in the same array of texts after passing through different taggers is primarily due to the fact that CoreNLP assigns mistakenly part of speech labels to punctuation marks, such as whether they are verbs or nouns. Besides, some taggers sometimes do not recognize tokens as parts of speech at all and do not give them any tag. A deeper analysis of the reasons for this discrepancy is a subject for an independent study.

One of the research problems is that the selected taggers have different classifications of parts of speech and morphological categories; therefore the number of part-of-speech tags as well as their designations differs in the analyzed instruments. The TextBlob uses the Penn TreeBank Tagset [22], the CoreNLP and spaCy use the Universal POS tag set [23], the TreeTagger uses STTS ("Stuttgart-Tübingen Tagset") [24] and RFTagger uses German RFTagger-POS-Tagset [25]. The two latter tag sets were designed specifically for the German language, so it is supposed to have performed better.

III. Analysis

Thus, having passed the same 35 texts through five different taggers, we got the following picture of the mistakes they made:

TextBlob

The first tagger in our research is ***TextBlob***. Out of 10272 tokens in TextBlob, 1201 part-of-speech tags are wrong. Error rate is 11.69%.

TextBlob uses a set of tags designed specifically for the English language. Its basic set of tags contains tags that are not needed for tagging German texts (for example, VBP - verb, 3sg pres, VBZ - verb, non-3sg pres, VBG - verb, gerund). On the other hand, it lacks tags that are important for German (for example, to annotate possessive pronouns or separable verb prefixes) (Table I) [26].

TABLE I. THE PENN TREEBANK POS TAGSET

CC	Coordinating conj.	TO	infinitival *to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP$	Possessive pronoun	)	Right bracket character
RB	Adverb	″	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol	”	Right close double quote

The most frequent error made by TextBlob is in the wrong tagging of nouns and proper names. A lot of errors are associated with the assignment of the NNP - Proper Noun tag to a wide range of words and even punctuation marks, for example, such as *Lieblingsfilm, ahnte, bewusst, Spielzeug, Schlösser, offener* and ". In most cases, for the nouns with the tag (NN) the proper name tag is given (NNP) and vice versa. It may happen because in the German language nouns are written with a capital letter according to the punctuation rules, as well as proper names. In some cases the tag (NNP) is given to other parts of speech, such as verbs and adjectives (Table II) Here and below, bold cell borders mark different types of mismatches between POS-tagging and the real situation.

The next error is also connected with nouns (NN), this time not proper names, but common nouns. TextBlob makes critically many mistakes related to non-distinction between singular and plural nouns. It is important to note that only TextBlob generally has different tags for singular and plural nouns (Table III).

Also quite common are mistakes related to identification of adjectives. Sometimes the tagger defines an adjective as a noun (NN) or pronoun (IN). Such parts of speech as nouns (NN) and verbs (VERB) get a tag of an adjective (JJ) with no clear dependence (Table IV).

TABLE II. ERRONEOUS TAGGING OF PROPER NAMES BY THE TEXTBLOB TAGGER

Word	Recognized as	Correct tag
Landon	**NN**	**NNP**
Carter	NN	NNP
Jamie	NN	NNP
Sullivan	NN	NNP
Landon	NN	NNP
Jamie	NN	NNP
Jamie	NN	NNP
bewusst	**NNP**	**JJ**
ahnte	NNP	**VBD**
träumte	NNP	VBD
einfach	NNP	JJ

TABLE III. ERRONEOUS TAGGING OF COMMON NOUNS BY THE TEXTBLOB TAGGER

Word	Recognized as	Correct tag
Jahrhunderte	**NN**	**NNS**
Arbeiten	NN	NNS
Jahren	NN	NNS
Erfolge	NN	NNS
Wissenschaften	NN	NNS
Haupteigenschaften	NN	NNS
Experimente	NN	NNS
Kontinenten	NN	NNS
Teile	NN	NNS
Jahre	NN	NNS
Tabellen	NN	NNS
Beschreibungen	NN	NNS
Sprachen	NN	NNS
Länder	NN	NNS

TABLE IV. ERRONEOUS TAGGING OF ADJECTIVES BY THE TAGGER TEXTBLOB

Word	Recognized as	Correct tag
rentabel	**NN**	**JJ**
negative	NN	JJ
instabilen	NN	JJ
schnelle	NN	JJ
natürliche	IN	JJ
Phantasie	JJ	NN
Entdeckung	JJ	NN
Phantasie	JJ	NN
wünschen	JJ	VB

In the German language, an important function is performed by a separable prefix, which usually finds itself at the end of the sentence. TextBlob does not have a special tag for tagging the separable prefix, so it regularly tags them incorrectly: most often as a particle (Table V).

TABLE V. ERRONEOUS TAGGING OF THE SEPARABEL PREFIX BY THE TEXTBLOB TAGGER

Word	Recognized as	Correct tag
durch	**RP**	**??? (no suitable tag)**
aus	RP	???
statt	RP	???

As it was mentioned above, incorrect classification of tags often leads to occurrence of errors. Thus, the tag (DT = Determiner) is used for different parts of speech, such as the article, demonstrative and quantitative pronouns, etc. Such despecification is certainly a serious drawback of the TextBlob tagger, which does not allow for using it for corpus of learner texts, on the basis of which further research is supposed to be conducted and tools created for teaching German as a foreign language (Table VI).

TABLE VI. TEXTBLOB TAGGING OF ARTICLES AND SOME TYPES OF PRONOUNS BY THE COMMON TAG

Word	Recognized as	Correct tag
einige	**DT**	**??? (no suitable tag)**
vielen	DT	???
Alle	DT	???
viele	DT	???
diese	DT	???
dieser	DT	???

SpaCy

The next tagger in our research is spaCy. This tagger uses a universal set of tags (Universal POS tags) (Table VII) [27].

TABLE VII. UNIVERSAL POS TAGS

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary
- CCONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PART: particle
- PRON: pronoun
- PROPN: proper noun
- PUNCT: punctuation
- SCONJ: subordinating conjunction
- SYM: symbol
- VERB: verb
- X: other

SpaCy split the same 35 texts into 10277 tokens, i.e. much the same as TextBlob, but made 1022 errors in a part-of-speech tagging. Error rate is 9,94 %.

The most errors in spaCy are associated with the definition of the lemmas *werden*, *sein* and *haben* (in various grammatical forms). The tagger defines full verbs (VERB) *werden, sein, haben* as auxiliary, giving them the tag (AUX). It is important to note that the developers themselves, referring to AUXILIARY Tense auxiliaries (***has*** *(done),* ***is*** *(doing),* ***will*** *(do)),* Passive auxiliaries (***was*** *(done),* ***got*** *(done)),* Modal auxiliaries (***should*** *(do),* ***must*** *(do))* and Verbal copulas (*He* ***is*** *a teacher),* pay attention to the fact, «that not all languages have grammaticalized auxiliaries, and even where they exist the dividing line between full verbs and auxiliaries can be expected to vary between languages». And further on: «*Modal verbs* may count as auxiliaries in some languages (English). In other languages their behavior is not too different from the main verbs and they are thus tagged VERB» [26]. It is worth noting that this repeated mistake in tagging is not related to students' mistakes in sentences, as it occurs in both places.

It is worth mentioning that this mistake is very common in all the researched taggers except TextBlob. It does not have this mistake since the set Penn TreeBank has no special tag for the auxiliary verb (Table VIII).

TABLE VIII. ERROUNEOUS TAGGING OF FULL VERBS AS AUXILIARY BY THE SPACY TAGGER

Word	Recognized as	Correct tag
werden	**AUX**	**VERB**
werden	AUX	VERB
sein	AUX	VERB
war	AUX	VERB
hat	**VERB**	**AUX**
haben	AUX	VERB

In the German language the principal parts of speech have special distinctive features, which allow for defining and differentiating parts of speech. For instance, nouns in German are spelled with a capital letter; verbs in German have a definite system of endings and strict place in the sentence; in the German language there are certain types of pronouns which change according to the concrete scheme. However, the spaCy tagger repeatedly and erroneously recognized all the aforementioned parts of speech without any particular order (Table IX).

TABLE IX. ERRONEOUS TAGGING OF NOUNS, VERBS AND PRONOUNS BY THE SPACY TAGGER

Word	Recognized as	Correct tag
Song	**PROPN**	**NOUN**
Lebens	PROPN	NOUN
Globalisierungsprozess	PROPN	NOUN
können	PROPN	VERB
glaube	**NOUN**	**VERB**
sein	**AUX**	**PRON**

SpaCy often gives verb tags (VERB) to adjectives with the ending –en. This mistake, unlike the previous ones, has a logical explanation, since this ending (-en) also characterizes some verb forms in the German language (Table X).

TABLE X. ERRONEOUS TAGGING OF ADJECTIVES WITH –EN ENDING BY THE SPACY TAGGER

Word	Recognized as	Correct tag
sozialen	**VERB**	**ADJ**
natürlichen	VERB	ADJ
engen	VERB	ADJ
vielfältigen	VERB	ADJ

Besides, the spaCy tagger could not recognize the biggest number of tokens compared to other taggers. Most frequently, these linguistic units were left without any tag (Table XI).

TABLE XI. LACK OF TAGGING IN CERTAIN TOKENS IN SPACY

Word	Recognized as	Correct tag
mittel-(amerikanischen)	**not recognized**	**ADJ**
bat	not recognized	**VERB**
mir	not recognized	**PRON**
pro	not recognized	**ADP**

Mistakes also occurred during recognition of the part of speech – the numeral (NUM). SpaCy marked the numeral by the X tag (other) or gave it a false tag. SpaCy does not mark numerals, it does not matter whether we are talking about cardinal or ordinal numerals, which are followed by a dot in a written German text (for example, 300 is a cardinal number, but 300. is an ordinal number) (Table XII).

TABLE XII. MISTAKES OF SPACY IN TAGGING NUMERALS

Word	Recognized as	Correct tag
2000	**X**	**NUM**
dritten	**NOUN**	NUM
1993	**X**	**X**
1799		NUM
7.	**not recognized**	**NUM (ordinal)**
31.	not recognized	NUM (ordinal)

SpaCy regularly makes mistakes when identifying complex nouns, ordinal numbers and the particle *zu* (Table XIII).

TABLE XIII. MISTAKES OF SPACY IN TAGGING COMPLEX NOUNS, ORDINAL NUMBERS AND THE PARTICLE *ZU*

Word	Recognized as	Correct tag
Erwachsenenzustand	**PROPN**	**NOUN**
Lebensstandard	**ADV**	NOUN
Verbrauchtsstandarts	ADV	NOUN
zu	**ADP**	**PART**
zu	ADP	PART
7.	**not recognized**	**NUM**
31.	not recognized	NUM

About a third of all errors in spaCy part-of-speech tagging are associated with incorrect identification of adjectives and verbs as nouns (Table XIV).

TABLE XIV. MISTAKES OF SPACY IN TAGGING ADJECTIVES AND VERBS

Word	Recognized as	Correct tag
religiöse	**NOUN**	**ADJ**
langen	NOUN	ADJ
zuziehen	NOUN	**VERB**
plagte	NOUN	VERB
störte	NOUN	VERB

CoreNLP

CoreNLP out of 10373 tokens mistakenly tagged 610. Percentage of error - 5.88%

Mistakes made by the tagger CoreNLP are very similar

to those of SpaCy. Possibly this is due to the fact that both the taggers use the same universal set of tags (Universal POS tags), which has a rather simplified morphological classification, no tags for some parts of speech, and also has some generalized tags for some forms differentiated by other taggers (e.g. there is no distinction between verb or pronoun types).

Just like the previous tagger, CoreNLP makes many irregular mistakes when defining auxiliary and semantic verbs, nouns, adjectives and pronouns (see the clarifying comments on the definition of auxiliary verbs in the errors section spaCy) (Table XV).

TABLE XV. MISTAKES OF CORENLP IN TAGGING VERBS, NOUNS, ADJECTIVES AND PRONOUNS

Word	Recognized as	Correct tag
ist	**AUX**	**VERB**
wird	AUX	VERB
kann	AUX	VERB
hat	**VERB**	**AUX**
Genie	**PROPN**	**NOUN**
Riesige	PROPN	**ADJ**
eine	**PRON**	**DET**

The CoreNLP tagger does not have a special tag for modal verbs, therefore they are regularly identified as auxiliary (AUX). Such identification, in our opinion, is contentious. Even though modal verbs belong to the category of synsemantic, they comprise its separate subtype “modifying, specifying (modal) verbs, which is opposed to another kind of verbs called auxiliary” (Table XVI).

TABLE XVI. CONTENTIOUS TAGGING OF MODAL VERBS IN CORENLP

Word	Recognized as	Correct tag
mag	**NOUN**	**VERB**
könnte	**AUX**	VERB
möchte	AUX	VERB
moechte	AUX	VERB

Like spaCy, this tagger very often tags the full-valued verbs *werden*, *sein* and *haben* (in different grammatical forms) as auxiliary verbs (Table XVII). The table shows only the result of manual analysis of tagged sentences where functioning of these verbs as full-valued is determined by the context.

TABLE XVII. MISTAKES OF CORENLP IN TAGGING FULL-VALUED VERBS *WERDEN*, *SEIN* AND *HABEN*

Word	Recognized as	Correct tag
ist	**AUX**	**VERB**
wird	AUX	VERB
habe	AUX	VERB

Similar to spaCy and TextBlob described above, the CoreNLP does not have a special tag for tagging separable prefixes, marking them as ADP (Adposition) or ADV (Adverb) (Table XVIII).

TABLE XVIII. MISTAKES OF CORENLP IN TAGGING SEPARABLE PREFIXES.

Word	Recognized as	Correct tag
auf (aufwenden)	**ADP**	**separable prefix**
statt(findet)	ADP	separable prefix
ab(decken)	ADP	separable prefix
ein(kaufen)	**ADV**	**separable prefix**
teil(nehmen)	ADV	separable prefix

CoreNLP confuses some punctuation marks, tagging, for example, quotation marks with different parts of speech: PROPN, NUM, ADP, NOUN, etc. or does not recognize them at all (Table XIX).

TABLE XIX. MISTAKES OF CORENLP IN TAGGING SIGNS OF PUNCTUATION

Word	Recognized as	Correct tag
"	**PROPN**	**PUNCT**
"	PROPN	PUNCT
"	**NUM**	**PUNCT**
"	**not recognized**	**PUNCT**

TreeTagger

Another tagger in our research is – ***TreeTagger***. 523 out of 10277 tokens in TreeTagger are tagged wrong. Error rate - 5.08%

TreeTagger uses the following classification of parts of speech: STTS ("Stuttgart-Tübingen-TagSet") (Table XX) [24].

The set of tags in STTS consisting approximately of 50 tags was developed especially for the German language. The number of mistakes made by the tagger, which uses this more exact set of tags, hypothetically must be significantly lower. It was proved by the experiment: the percentage of mistakes is almost twice lower compared to the three instruments described above.

TABLE XX. STTS = STUTTGART-TÜBINGEN-TAGSET

Tag	Description	Tag	Description
ADJA	attributive adjective (including participles used adjectivally)	PPOSAT	possessive determiner
ADJD	predicate adjective; adjective used adverbially	PRELAT	relative depending on a noun
ADV	adverb (never used as attributive adjective)	PRELS	relative pronoun (i.e. forms of der or welcher)
APPR	preposition left hand part of double preposition	PTKA	particle with adjective or adverb
APPRART	preposition with fused article	PTKANT	answer particle
APPO	postposition	PTKNEG	negative particle
APZR	right hand part of double preposition	PTKREL	indeclinable relative particle
ART	article (definite or indefinite)	PTKVZ	separable prefix
CARD	cardinal number (words or figures); also declined	PTKZU	infinitive particle
FM	foreign words (actual part of speech in original language may be appended, e.g. FMADV/ FM-NN)	PWS	interrogative pronoun
ITJ	interjection	PWAT	interrogative determiner
KON	co-ordinating conjunction	PWAV	interrogative adverb
KOKOM	comparative conjunction or particle	PWAVREL	interrogative adverb used as relative
KOUI	preposition used to introduce infinitive clause	PWREL	interrogative pronoun used as relative
KOUS	subordinating conjunction	TRUNC	truncated form of compound
NA	adjective used as noun	VAFIN	finite auxiliary verb
NE	names and other proper nouns	VAIMP	imperative of auxiliary
NN	noun (but not adjectives used as nouns)	VAINF	infinitive of auxiliary
PAV [PROAV]	pronominal adverb	VAPP	past participle of auxiliary
PAVREL	pronominal adverb used as relative	VMFIN	finite modal verb
PDAT	demonstrative determiner	VMINF	infinitive of modal
PDS	demonstrative pronoun	VMPP	past participle of auxiliary
PIAT	indefinite determiner (whether occurring on its own or in conjunction with another determiner)	VVFIN	finite full verb
PIS	indefinite pronoun	VVIMP	imperative of full verb
PPER	personal pronoun	VVINF	infinitive of full verb
PRF	reflexive pronoun	VVIZU	infinitive with incorporated zu

Here are some of the mistakes revealed during the tagging analysis with TreeTagger.

The biggest number of tagging mistakes by TreeTagger is related to the wrong interpretation of verbs Firstly, many errors in the performance of the TreeTagger tagging are associated with the definition of the lemmas *werden, sein* and *haben* in all cases as auxiliary (VAFIN), even in those cases when verbs perform as semantic (VVFIN) (Table XXI). Only TextBlob does not have this error, since there is no special tag for an auxiliary verb.

TABLE XXI. MISTAKES OF TREETAGGER IN TAGGING FULL-VALUED VERBS

Word	Recognized as	Correct tag
hatte	**VAFIN**	**VVFIN**
war	VAFIN	VVFIN
wird	VAFIN	VVFIN

Secondly, the tagger often made mistakes in differentiating homonymic grammatical verb forms: infinitive and 1st and 3rd person plural (Table XXII).

TABLE XXII. MISTAKES OF TREETAGGER IN TAGGING HOMONYMIC GRAMMATICAL VERB FORMS

Word	Recognized as	Correct tag
bewundern	**VVINF**	**VVFIN**
gehen	VVINF	VVFIN
kennen	VVINF	VVFIN
erzahlen	VVFIN	VVINF
kriegen	VVFIN	VVINF

Another frequent error of TreeTagger is related to the failure to recognize proper names (NE). Most often the tagger defines a proper name as a common one (NN) (Table XXIII).

TABLE XXIII. MISTAKES OF TREETAGGER IN TAGGING PROPER NAMES

Word	Recognized as	Correct tag
Karelien	**NN**	**NE**
Towns	NN	NE
Prinz	NN	NE
Jaden	NN	NE

Besides, TreeTagger does not recognize marks of punctuation requiring the tag $(, i.e. tags inside the sentence except the comma. For instance, inverted commas inside a sentence are often marked as adjectives (ADJD/ADJA) by the tagger (Table XXIV).

TABLE XXIV. MISTAKES OF TREETAGGER IN TAGGING PUNCTUATION MARKS INSIDE A SENTENCE

Word	Recognized as	Correct tag
"	**ADJD**	**$(**
"	ADJD	$(
"	ADJA	$(

Some errors in tagging parts of speech are related to student mistakes, first of all syntactical. Thus, incorrect setting of the separable prefix not at the end of the sentence, but in the middle, leads to its interpretation as an adverb (Fig. 3).

Fig. 3. Mistake of Treetagger in indentifying the separable prefix „ab"

RFTagger

394 POS-tags out of 10148 tokens were recognized as errors. Percentage of error is 3,88%.

The developer of the RFTagger, just like the TreeTagger, is the university of Munich [28]. To teach the model of RFTagger, the data of Tiger treebank from Stuttgart university were used, which jointly with Tubingen university developed a set of tags STTS for TreeTagger (the set of tags see above). Since the developers and their developments overlap in many ways, the trends in errors are largely similar. Nevertheless, a separate set of tags was developed for RFTagger (Table XXV).

TABLE XXV. COMMON TAGS FROM GERMAN RFTAGGER PART-OF-SPEECH TAGSET

Tag	Description	Tag	Description
ADJA	attributive adjectives	Name	proper noun
ADJD	adjective with predicative or adverbial usage	PART	particle
ADV	adverbs	PRO	pronoun
APPO	postposition	PROADV	pronomial adverb
APPR	preposition	SYM	symbol
APPRART	preposition with incorporated article	PUN	punctuation
APZR	circumposition (right part)	TRUNC	truncated word form
ART	article	VFIN	finite verb
CARD	cardinal number	AUX	auxiliary
CONJ	circumposition (right part)	VIMP	imperative verb
FM	foreign word	VINF	infinitival verb
ITJ	interjection	VPP	participle verb
N	noun	VPP	participle verb

In general, it can be argued that RFTagger, like TreeTagger, determines parts of speech much more accurately in comparison with the first three tools, but it detects in many ways the same regular types of errors as the TreeTagger.

Thus, the RFTagger does not distinguish infinitive or finite forms of 1st and 3rd person plural verbs either. Errors can also be detected in both directions: finite forms may get the tag of the infinitive, or infinitive may be defined as a finite form of the verb by the tagger (Table XXVI).

TABLE XXVI. MISTAKES OF THE RFTAGGER IN TAGGING HOMONYMIC GRAMMAR FORMS OF VERBS

Word	Recognized as	Correct tag
kommunizieren	**VINF**	**VFIN**
streiten	VINF	VFIN
besuchen	**VFIN**	**VINF**
glauben	VINF	VFIN
bemerken	VFIN	VINF

Like other taggers, it does not recognize proper names. Just like other taggers, RFTagger often makes mistakes in identifying proper names: RFTagger ascribes them different parts of speech or marks them as FM – fremdes Material = an element of a foreign language (Table XXVII).

TABLE. XXVII. MISTAKES OF RFTAGGER IN TAGGING PROPER NAMES

Word	Recognized as	Correct tag
Jamie	**FM**	**N**
Landon	FM	N
Lila	**ADJA**	**N**
Olivie	ADJA	N

Characteristically, tokens *was* (question word) and *war* (past form of the verb *sein*) are recognized by the RFTagger tool as foreign elements (FM), since the tool evidently identifies them as the English language tokens *was* and *war*.

In the following example the wrong marking is related to the student's mistake in spelling the noun *Reden* with the lower case letter, which leads firstly to determining this word as a verb, and secondly – to the wrong interpretation of the following verb schlagen as infinitive, and not as an homonymic form of 3rd person plural, since the finite verb in such a syntactic position is hardly possible (Fig.4).

Fig. 4. Mistake of RFTagger in identifying the noun *reden* and verb *schlagen*

However, the percentage of errors in tagging using RFTagger is lower than in other cases.

IV. EXPERIMENTS SUMMARY

Thus, as you see in the diagram below (Fig.5), the highest percentage of errors in the tagging was shown by TextBlob, and the lowest - by RFTagger.

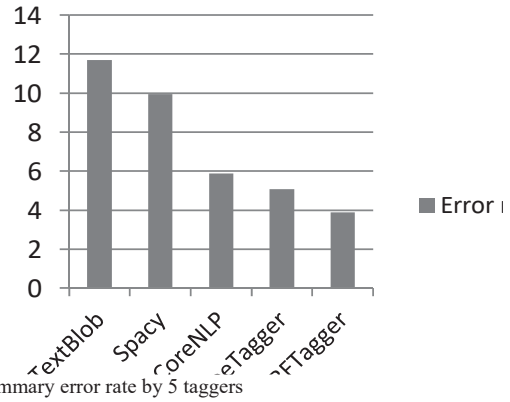

Fig. 5. Summary error rate by 5 taggers

On the whole, 5 tools - TreeTagger, RFTagger, spaCy, CoreNLP, TextBlob - have shown themselves in different ways in tagging the same texts. Summary statistics can be seen in the table below.

TABLE XXVIII. SUMMARY STATISTICS OF TAGGING WITH 5 TAGGERS

	Number of tokens	Number of mistakes	Error rate
CoreNLP	10373	610	5,88 %
Spacy	10277	1022	9.94 %
TreeTagger	10277	523	5,08 %
TextBlob	10272	1201	11,69 %
RFTagger	10148	394	3.88 %

Thus, it should be noted that different taggers revealed different tendencies in the inaccuracy of certain elements. Among the most significant are the following:

- Many errors in all taggers are associated with the replacement of a common noun with a proper name and vice versa.

- All taggers make many mistakes of various kinds when identifying proper names. For example, the name *Will* was defined by most taggers as a (modal) verb, and the element *von* in the proper noun *Alexander von Humbold* was defined as a preposition.

- In all taggers, the largest number of errors is associated with the incorrect verbs interpretation.

- The efficiency of the tagger depends on the set of tags used by it: taggers using tag sets developed especially for the German language determine parts of speech more precisely even in sentences containing errors.

- In some cases, where the level of tag generalization is higher, fewer mistakes in tagging are observed (for instance, when tagging verbs *werden, sein, haben* and pronouns of different types), but such generalization may have negative impact on usage for the purpose of recognition of errors and other tasks for artificial intelligence, where distinction between auxiliary and full verbs is of principal importance, or the articles and demonstrative pronouns for that matter.

Summing up the study, it can be argued that the focus of the STTS tag set on the German language distinguishes the TreeTagger and RFTagger among their counterparts in a positive way, which is statistically confirmed. In most cases, spelling mistakes and misprints of students do not affect the correctness of the identification of parts of speech by the taggers.

Spelling mistakes and typos of students in the middle of a word in most cases do not affect the correctness of the identification of parts of speech by the tagger. However, the replacement of lowercase and uppercase letters at the beginning of a word, as well as errors in the syntactic design of a sentence (word order) are often critical for the automated definition of parts of speech. The most successful for POS-tagging of German-language student texts is RFTagger, since the percentage of errors in its tagging is the lowest compared to other tools - 3.88%.

Errors in many cases can be brought into accordance and, with the help of additional edits in the work of the tagger, they can be minimized. So, it is necessary to finalize the definition of digits with a dot as an ordinal numeral without breaking it from the remainder of the sentence. It also requires automated differentiation of the verbs *werden, sein* and *haben* in the function of an auxiliary or full-valued verb, which in most cases is formally expressed in the presence or absence of a second verb in the sentence. It is necessary to work on improving the automated process of distinguishing proper names and common names. It is obvious that the process of automated partial markup should not only be base on the formal features of certain parts of speech and the rules of syntactic compatibility, but also should take into account the interaction of both.

V. CONCLUSION

The obtained result is important first of all for the creators of the German language learner text corpora. It is still difficult to say whether the research result can be extended to similar learner texts containing German texts of speakers of not Russian, but primarily other languages by type, since the nature of the most common mistakes made by speakers of different languages is obviously different. In this aspect, the presented study can serve as part of a comparative study. In any case, for the learner's corpora of German texts written by students-native speakers of Russian, the most suitable today are RFTagger and TreeTagger.

At the next stage it is necessary to investigate in detail, which student errors in particular and in what manner impact the work of these selected taggers. At present, a comparison is made between the corpora sentences containing errors and the corrected sentences, as well as analysis of interconnections of error types with the work of RFTagger and TreeTagger. Thus, 11 types of errors were identified, which with varying degrees

of regularity are the cause of incorrect part of speech tagging, and where, when correcting students' errors, part of speech tagging turns out to be correct. However, this means a separate detailed research requiring a separate detailed description.

Our next goal is to improve the work of the RFTagger on the basis of the conducted analysis in the issue of distinguishing homonymous finite and infinitive forms and in identifying proper names. Reducing the percentage of errors in recognizing parts of speech in the corpus will allow creating high-quality digital educational tools that are applicable for German-language educational texts. In particular, we are speaking about such tools as automated error correction, preparation of individualized training exercises, tests, etc. Since all these tools require precise definition of a part of speech, it is necessary to use the most suitable tagger for these purposes in the corpus of student texts.

ACKNOWLEDGMENT

This research is implemented in Petrozavodsk State University (PetrSU) with financial support by the Innovation Promotion Fund of Russia within Agreement no. No 4GUCodIIC12-D7/72694 on the topic “OPEN LIBRARY FOR AUTOMATED ASSESSMENT OF EDUCATIONAL TEXTS IN GERMAN”.

The authors are grateful to anonymous reviewers of the FRUCT Association and experts of Artificial Intelligence Center of PetrSU for their valuable comments to improve the quality of this paper.

REFERENCES

[1] A.A. Rogov, R.V. Abramov, A.A. Lebedev, K.A. Kulakov, N.D. Moskin, „Text Attribution in Case of Sampling Imbalance by the Method of Constructing an Ensemble of Classifiers Based on Decision Trees” *in Proc. of the 22th International Conference DAMDID/RCDL'2020,* Oct. 2020, pp.319-328.

[2] A. O'Keeffe, “Data-driven learning – a call for a broader research gaze”, *Language Teaching*, vol. 54(2), pp. 259-272.

[3] O. Vinogradova, ”To automated generation of test questions on the basis of error annotations in EFL essays A time-saving tool?”, *Learner Corpora and Language Teaching*, vol. 92, 2019, pp. 29-48.

[4] Approaches to Natural Languages Processing Tasks – Data Driven Investor. Retrieved April 28, 2019, Web: https://medium.com/datadriveninvestor/approaches-to-natural-languages-processing-tasks-aa5ba37a7a6f

[5] D.Korzun, E.Balandina, A.Kashevnik, S.Balandin, & F.Viola. Ambient Intelligence Services in IoT Environments: Emerging Research and Opportunities. IGI Global. 2019, 199p. Web: https://www.igi-global.com/book/ambient-intelligence-services-iot-environments/218560

[6] O. Kholkovskaia, “Role of the Brown Corpus in the History of Corpus Linguistics”, Web: http://poseidon2.feld.cvut.cz/conf/poster/proceedings/Poster_2017/Section_HS/HS_018_Kholkovskaia.pdf

[7] A. Díaz-Negrillo et al, “Towards interlanguage POS annotation for effective learner corpora in SLA and FLT”, *Language Forum*, vol. 36, No 1-2, 2010, pp. 139-154

[8] S.Ghosh and B.K.Mishra, “Parts-of-Speech Tagging in NLP: Utility, Types, and Some Popular POS Taggers”, *Natural Language Processing in Artificial Intelligence*, 2020, pp.131-165

[9] M.Alabbas, R. Allan, “Improved POS-Tagging for Arabic by Combining Diverse Taggers”, *IFIP Advances in Information and Communication Technology*, 2012, pp. 107–116.

[10] I. Rehbein, “Fine-grained pos tagging of german tweets”, *Language Processing and Knowledge in the Web*, 2013, pp. 162–175.

[11] G. Wisniewski, F. Yvon, “How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project”, *in Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics,* Jun. 2019, pp. 218 – 227.

[12] I. Kotyurova, “Creation of Learner Corpora as a Developing Area of Corpus Linguistics”, *International Scientific Journal*, №5, 2020, pp. 100-109.

[13] B.Van Rooy, L. Schäfer. “The effect of learner errors on POS tag errors during automatic POS tagging”, *Southern African Linguistics and Applied Language Studies*, 2009, vol. 20, pp. 325-335.

[14] I.Torubarov, “Research for automated annotation of errors in REALEC, a corpus of English learner essays”, Web: https://github.com/isikus/realec-autoannotation

[15] K. Zupanc and Z.Bosnić. “Automated essay evaluation with semantic analysis”, *Knowledge-Based Systems,* vol. 120, 2017, pp. 118-132.

[16] O. Lyashevskaya, I.Panteleeva, O.Vinogradova, “Automated assessment of learner text complexity”, *Assessing Writing*, vol. 49, 2021, n. pag.

[17] T. Mizumoto, R. Nagata, “Analyzing the impact of spelling errors on POS-tagging and chunking in learner English”, *in Proc. of 4th Workshop on Natural Language Processing Techniques for Educational Applications*, 2017, pp. 54–58.

[18] M. Dickinson, M. Ragheb, „Dependency annotation for learner corpora”, *in Proc. of 8th Workshop on Treebanks and Linguistic Theories,* 2009, pp. 59–70.

[19] K. Sakaguchi, T. Mizumoto, M. Komachi, Y.Matsumoto, „Joint English spelling error correction and POS tagging for language learners writing”, *in Proc. of 24th International Conference on Computational Linguistics*, 2012, pp. 2357–2374.

[20] R. Nagata, A. Kawai, „Exploiting learners’ tendencies for detecting English determiner errors”, *in Lecture Notes in Computer Science*, volume 6882/2011, 2011, pp. 144–153.

[21] R. Nagata, T. Mizumoto, Y. Kikuchi, Y. Kawasaki, K. Funakoshi, „A POS Tagging Model Designed for Learner English”, *in Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, 2018, pp. 39–48.

[22] Penn TreeBank Tagset, Web: https://www.sketchengine.eu/penn-treebank-tagset/

[23] UD version 2. AUX: auxiliary. Web: https://universaldependencies.org/u/pos/AUX_.html

[24] Stuttgart-Tübingen Tagset STTS, Web: https://www.sketchengine.eu/german-stts-part-of-speech-tagset/

[25] German RFTagger-POS-Tagset, Web: https://www.sketchengine.eu/german-RFTTageer-part-of-speech-tagset/

[26] A.Taylor, V. Marcus, B. Santorini. “The Penn Treebank: An overview”, *Treebanks: Building and Using Parsed Corpora*, pp. 5–22.

[27] Universal POS tagset, Web: https://www.sketchengine.eu/tagsets/universal-pos-tags/

[28] Die Ludwig-Maximilians-Universität München, RFTagger, Web: https://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/