УДК 81.322.2 **DOI: 10.18413/2313-8912-2022-8-3-0-6**

Котюрова И. А.[1]
Щеголева Л. В.[2]

Анализ некорректной работы POS-разметчиков в корпусе немецких ученических текстов с лингвистическими ошибками

[1] Петрозаводский государственный университет
пр. Ленина, 33, Петрозаводск, 185910, Республика Карелия, Россия
E-mail: koturova@petrsu.ru

[2] Петрозаводский государственный университет
пр. Ленина, 33, Петрозаводск, 185910, Республика Карелия, Россия
E-mail: schegoleva@petrsu.ru

Статья поступила 13 мая 2022г.; принята 27 июля 2022г.;
опубликована 30 сентября 2022г.

Аннотация. Электронный корпус ученических текстов на немецком языке ПАКТ содержит разметку частей речи. Разметка выполняется автоматически с помощью RFTagger. Так как тексты корпуса написаны обучающимся, то они могут содержать разного рода ошибки: грамматические, орфографические, стилистические и другие. Предложения могут быть сформулированы некорректно, без учета правил языка и принятых норм. Это может влиять на работу программ, обрабатывающих тексты в автоматическом режиме, и в результате формировать неправильную разметку, которую необходимо верифицировать вручную. Целью исследования является анализ степени влияния разного рода ошибок в неаутентичных текстах на результаты работы автоматического частеречного разметчика. На основе экспертной разметки в текстах корпуса ПАКТ были выделены 11 типов ошибок, которые влияют на качество работы частеречного разметчика. Для каждого такого типа из корпуса были отобраны по десять предложений, содержащих ошибку. Полученный пул текстов был обработан частеречными разметчиками RFTagger и TreeTagger. Части речи, предложенные этими автоматическими таггерами, были сопоставлены с частями речи, определенными экспертами вручную. В результате сравнения удалось выявить следующие закономерности: частеречные разметчики ошибаются: в случае написания несклоняемой формы прилагательного вместо склоняемой; при раздельном написании одного слова; при отсутствии суффикса «-er» в притяжательных прилагательных, образованных от географических наименований; при написании существительных со строчной буквы; при написании глагола с прописной буквы. Для каждого случая в статье приведен анализ форм и причин неправильной частеречной разметки, а также различий в работе двух разметчиков. Учет выявленных закономерностей позволит более эффективно организовать верификацию автоматической частеречной разметки в ученических корпусах на немецком языке. Результаты исследования также будут полезны для разработчиков автоматических частеречных разметчиков.

Ключевые слова: Частеречная разметка; Ученический корпус; Немецкий язык; RFTagger; TreeTagger

Информация для цитирования: Котюрова И. А., Щеголева Л. В. Анализ некорректной работы POS-разметчиков в корпусе немецких ученических текстов с лингвистическими ошибками // Научный результат. Вопросы теоретической и прикладной лингвистики. 2022. Т. 8. № 3. С. 87-99. DOI: 10.18413/2313-8912-2022-8-3-0-6

UDC 81.322.2 **DOI: 10.18413/2313-8912-2022-8-3-0-6**

Irina A. Kotiurova[1]
Liudmila V. Shchegoleva[2]

Analysis of incorrect POS-tagging in student texts with linguistic errors in German

[1] Petrozavodsk State University
33 Lenin St., Petrozavodsk, 185910, Republic of Karelia, Russia
E-mail: koturova@petrsu.ru

[2] Petrozavodsk State University
33 Lenin St., Petrozavodsk, 185910, Republic of Karelia, Russia
E-mail: schegoleva@petrsu.ru

Received 13 May 2022; accepted 27 July 2022; published 30 September 2022

Abstract. The electronic learner corpus of student texts in German, the PACT, contains the parts-of-speech (POS) tagging. This markup is performed automatically using RFTagger. Since the texts of the corpus are written by students, they may contain various kinds of errors: grammatical, spelling, stylistic, and others. Sentences may be formulated incorrectly, without taking into account the rules of the language and accepted norms. This can affect the work of programs that process texts in automatic mode, and as a result, generate incorrect tagging that needs to be verified manually. The purpose of the study is to investigate the degree of influence of various kinds of errors in non-authentic texts on the results of the automatic part-of-speech tagging. Based on expert error markup in the corpus texts, 11 types of errors were identified that affect the part-of-speech tagger quality. For each type of error, ten sentences containing an error were selected from the corpus. The resulting pool of texts was processed by the part-of-speech taggers RFTagger and TreeTagger. The parts of speech that were suggested by these automatic taggers were compared with the parts of speech determined by experts manually. As a result of the comparison, the following patterns were revealed: part-of-speech taggers are mistaken when writing the non-declinable form of the adjective instead of the declinable; when writing one word separately; in the absence of the suffix "-er" in possessive adjectives formed from geographical names; when writing nouns with a lowercase letter; when writing a verb with a capital letter. For each case, the article provides an analysis of the forms and causes of incorrect POS-tagging, as well as differences in the work of the two part-of-speech taggers. Taking into account the revealed patterns will allow more efficient organization of the POS-tagging verification in the learner corpus in German. The results of the study will also be useful for developers of part-of-speech taggers.

Keywords: POS-tagging; Learner corpus; German; RFTagger; TreeTagger
How to cite: Kotiurova, I. A. and Shchegoleva, L. V. (2022). Analysis of incorrect POS-tagging in student texts with linguistic errors in German, *Research Result.*

Theoretical and Applied Linguistics, 8 (3), 87-99. DOI: 10.18413/2313-8912-2022-8-3-0-6

Введение

Современные исследования языка опираются на электронные корпусы текстов. Чаще всего такие корпусы содержат специальную разметку, позволяющую проводить эффективный поиск и выполнять статистические исследования в рамках корпусной лингвистики. Одним из вариантов такой разметки может быть определение частей речи (POS-разметка). О роли частеречной разметки специалисты заговорили с самого начала исследований в области машинной обработки естественного языка (Heeman, 1998; Manning, 2003; Qian, 2012 и др.)

Частеречная разметка может быть выполнена вручную с верификацией несколькими экспертами. Однако это очень трудоемкий процесс. Часто для определения частей речи используют компьютерные программы, реализующие специализированные алгоритмы или модели языка (разметчики, или так называемые таггеры). Результаты работы этих программ могут отличаться и по перечню определяемых частей речи, и по качеству выполненной работы.

Так, Т. Хорсманн, Н. Эрбс и Т. Зеш в своей статье «Fast or Accurate? – A Comparative Evaluation of PoS Tagging Models» сравнивают работу 22-х разметчиков на англо- и немецкоязычных текстах и приходят к выводу, что для каждого корпуса должен быть подобран разметчик, подходящий для конкретных задач, поскольку в разных контекстуальных условиях более эффективно проявляют себя разные программы (Horsmann et al., 2015).

Это, безусловно, относится и к текстам так называемых ученических корпусов, составленных из текстов, написанных обучающимися. Ученические корпусы имеют очень важную особенность. Текст, написанный обучающимся, может содержать разного рода ошибки: грамматические, орфографические, стилистические и другие. Предложения могут быть сформулированы некорректно, без учета правил языка и принятых норм. Это составляет главную ценность корпуса, особенно когда корпус содержит тексты, написанные при обучении иностранному языку, поскольку именно анализ ошибок дает понимание того, как происходит усвоение иностранного языка и позволяет находить наиболее эффективные способы преподавания.

Использование автоматических частеречных разметчиков на таких текстах вызывает вопрос: не повлияют ли эти ошибки на результат работы таггеров?

Таким образом, настоящее исследование имеет своей целью проверить гипотезу о степени влияния разного рода ошибок в неаутентичных текстах на результаты работы автоматического частеречного разметчика. Результаты исследования позволят более эффективно использовать автоматические частеречные таггеры. При ручной корректировке результатов их работы можно будет обращать внимание на наиболее вероятные ошибки, что сократит время на обработку и верификацию разметки текстов корпуса. Возможно, эта информация заинтересует разработчиков таггеров для совершенствования работы алгоритмов или разработки специализированных алгоритмов в особых условиях обработки текстов с ошибками.

Обзор литературы

Тема качества работы частеречных разметчиков, применяемых к разного рода текстам, интересует исследователей с самого момента появления этих разметчиков. Актуальность вопроса не вызывает сомнений, поскольку от качества работы таггеров зависит работа многих приложений и автоматических

инструментов (например, перевода, редактирования, поиска и т.д.), базирующихся на частеречном анализе.

Кроме упомянутой выше статьи Т. Хорсманна, Н. Эрбса и Т. Зеша, которые дают обзор работы 22-х разметчиков, примененных к англо- и немецкоязычным текстам, можно найти немало работ, посвященных отдельным разметчикам или отдельным типам текстов, к которым они применяются.

Так, о сложностях частеречной разметки и способах улучшения ее качества в исторических текстах 15-18 вв. на немецком языке пишет М. Больманн (Bollmann, 2013).

Но и в современных текстах разных жанров и написанных в разных условиях выявляются особенности, требующие нюансированной настройки стандартных POS-разметчиков. Е. Бик (Bick, 2020) сравнивает качество частеречной разметки в аннотированном корпусе социальных медиа на немецком языке и делает вывод, что в твитах со стандартной орфографией это качество вдвое выше, чем в твитах с нарушениями орфографии и пунктуации, причем этот эффект был более заметен для морфологии, чем для синтаксиса.

К. Сугисаки, Н. Видмер и Г. Гаузендорф (Sugisaki et al., 2018) описывают собственный разработанный частеречный разметчик для своего корпуса немецкоязычных, написанных от руки открыток, поскольку тексты в нем часто содержат орфографические и грамматические отклонения от нормы классического письменного, например, публицистического текста.

А. Диаз-Негрилло, Д. Меурерс, С. Валера и Г. Вунш (Dıaz-Negrillo et. al., 2010) исследуют корпус ученических текстов NOCE, состоящий из текстов на английском языке, изучаемом испанскими учащимися, и характеризуют области, в которых свойства изучаемого языка систематически отличаются от тех, которые предполагаются схемами аннотации POS, разработанными для английского как родного языка.

Л. Кайпер, А. Горбах и С. Татер (Keiper et al., 2016) пишут о том, что в корпусе ученических текстов зачастую встречаются слова, которых нет в стандартном языке и которые поэтому часто не могут быть размечены или размечаются неверно. Авторы предлагают свой метод автоматического повышения точности таггеров на изучаемом языке с помощью использования структуры типичного задания для изучающего язык и автоматического создания POS-информации для слов, не входящих в стандартный тезаурус (так называемые out-of-vocabulary (OOV) words).

Свои решения проблемы выявления ошибок в автоматической частеречной разметке текстов на немецком языке предлагают И. Ребайн, Й. Руппенхофер (Rehbein et al., 2017), Х. Лофтссон (Loftsson, 2009), а также Д. Длигач и М. Палмер (Dligach et al., 2011), при этом они сходятся в том, что чтобы исправить неточность, сначала требуется ее обнаружить, и наиболее высокий показатель качества обнаружения ошибки в частеречной разметке дает учет и сравнение результатов работы сразу нескольких источников. Учет закономерностей, выявленных в нашем исследовании, возможно, также позволит улучшить работу детекторов ошибок частеречных разметчиков.

Материалы и методы

Исследование проводилось на текстах Петрозаводского аннотированного корпуса текстов ПАКТ[1], содержащего тексты эссе на немецком языке, написанные русскоговорящими студентами, изучающими немецкий язык в течение 1-4 лет. Все тексты корпуса содержат специализированную разметку ошибок, выполненную и

[1] ПАКТ: Петрозаводский аннотированный корпус текстов. URL: http://lingo.smartsensing.petrsu.ru/ (дата обращения: 15.04.2022)

верифицированную вручную, а также частеречную разметку, выполненную с помощью программы RFTagger.

На основе этих разметок был проведен анализ часто встречающихся типов ошибок, которые влияют на качество работы автоматического разметчика. В результате были выделены 11 типов ошибок:

1. Написание существительных со строчной буквы.
2. Написание глагола с прописной буквы.
3. Написание несклоняемой формы прилагательного вместо склоняемой.
4. Раздельное написание одного слова.
5. Опечатка, меняющая слово на несуществующее, или опечатка, меняющее одно слово на другое слово.
6. Отсутствие пробелов между токенами через дефис.
7. Отсутствие форматирования (оформление списка в строку и без разделительных знаков).
8. Отсутствие суффикса «-er» в притяжательных прилагательных, образованных от географических наименований.
9. Написание имени существительного, являющегося частью разорванного слова, со строчной буквы и без последующего дефиса.
10. Пробел внутри числительного.
11. Неверный порядок слов в предложении.

Затем для каждого типа ошибки были отобраны из корпуса по 10 предложений, содержащих такую же языковую ошибку, с различными вариантами окружения некорректного слова. Для каждого предложения вручную была выполнена частеречная разметка, которая считалась эталонной и с которой сверялось автоматическое определение тега. Таким образом был создан пул из 110 предложений, на котором проверялась выдвинутая гипотеза и исследовались закономерности влияния разных типов ошибок на работу частеречных разметчиков.

Для исследования качества работы разметчиков были отобраны два – RFTagger и TreeTagger, поскольку они показали наилучшие результаты частеречной разметки для немецкоязычных студенческих текстов (Котюрова, 2021).

Проверка качества работы разметчика выполнялась методом сравнения результата разметки с эталонной ручной разметкой.

Результаты исследования

Пул из 110 предложений с ошибками был обработан частеречными разметчиками RFTagger и TreeTagger. Части речи, которые были предложены этими автоматическими таггерами, были сопоставлены с частями речи, определенными экспертами вручную. Результаты сопоставления представлены в таблице 1.

Анализ результатов сопоставления показал, что:

Оба разметчика всегда (100% случаев) неверно размечают часть речи в словах со следующими языковыми ошибками:

– написание несклоняемой формы прилагательного вместо склоняемой;

– раздельное написание одного слова;

– отсутствие пробелов между токенами через дефис;

– отсутствие суффикса «-er» в притяжательных прилагательных, образованных от географических наименований;

– пробел внутри числительного;

– написание имени существительного, являющегося частью разорванного слова, со строчной буквы и без последующего дефиса (в этом типе ошибки RFTagger ошибся в 10 случаях из 10, а TreeTagger – в 9 случаях из 10).

RFTagger всегда ошибается в разметке существительных, написанных со строчной буквы, в то время как TreeTagger в каждом втором случае выдает правильный тег имени существительного.

Таблица 1. Сопоставление работы автоматических разметчиков с эталонной (ручной) разметкой частей речи
Table 1. Comparing the automatic part-of-speech tagging by RFTagger and TreeTagger with the reference (manual) part-of-speech tagging

№	Тип ошибки	Количество ошибок RFTagger	Количество ошибок TreeTagger
1	Написание существительных со строчной буквы	10	5
2	Написание глагола с прописной буквы	0	6
3	Написание несклоняемой формы прилагательного вместо склоняемой	10	10
4	Раздельное написание одного слова	10	10
5	Опечатка	7	7
6	Отсутствие пробелов между токенами через дефис	10	10
7	Отсутствие форматирования (оформление списка в строчку без разделительных знаков)	3	3
8	Отсутствие суффикса «-ег» в притяжательных прилагательных, образованных от географических наименований	10	10
9	Написание имени существительного-части разорванного слова со строчной буквы и без последующего дефиса	10	9
10	Пробел внутри числительного	10	10
11	Неверный порядок слов в предложении	4	1
	Итого:	84	81

RFTagger всегда верно распознает глагол в тех случаях, когда он написан с прописной буквы, в то время как TreeTagger в 6 случаях из 10 выдает неверную часть речи.

Оба разметчика примерно одинаково (не)справляются с задачей разметки текстов с языковыми ошибками, хотя RFTagger допустил чуть больше ошибок: на 84 ошибок RFTagger приходится 81 ошибка в TreeTagger.

Существенная разница в разметке двух таггеров наблюдается только в тех случаях, когда речь идет о замене строчной буквы на прописную и обратно в именах существительных и глаголах.

Обсуждение результатов

Таким образом, мы видим, что только для шести типов языковых ошибок частеречные разметчики постоянно ошибаются. Для остальных типов ошибок разметчики с разной степенью регулярности могут определять часть речи с допущенной языковой ошибкой то правильно, то неправильно.

Рассмотрим подробнее, какие ошибки допускают разметчики.

Анализ того, какую именно часть речи (тег) присваивает слову разметчик в тех случаях, где его ошибка накладывается на языковую неточность в тексте, позволяет поделить все такие примеры на те, которые являются ошибочными только в рамках предложения, но верными в

рамках отдельно взятого токена, и те, которые являются ошибочными как в рамках предложения, так и в рамках токена.

Например, если опечатка меняет слово так, что оно становится другим словом с принадлежностью к иной части речи, чем должно было быть (например, «Rind» (существительное) вместо «Sind» (глагол), и разметчик указывает часть речи «нового» слова (т.е. Rind – NN – имя существительное), то можно говорить о том, что ошибка в разметке касается токена только в рамках предложения, но не внутри токена. Если же «Rind», написанное вместо «Sind», размечено как ADJD – несклоняемое прилагательное, то можно говорить о том, что ошибка разметчика касается обоих уровней – и уровня токена, и уровня предложения.

Все представленные далее примеры являются ошибочными как минимум на уровне предложения, однако, на уровне токена разметчики ошибались не всегда. В некоторых случаях правильность работы разметчика на уровне токена напрямую зависит от типа ошибки.

В Таблице 2 представлена информация о регулярности допущения разметчиком ошибок на уровне отдельно взятого токена. Знак «–» в таблице означает отсутствие ошибки, то есть указывает, что разметчики показывали правильный тег того токена, который есть, а не того, который должен был бы быть в правильном языковом варианте. Знак «+/–» означает, что разметчики чаще ошибаются, чем не ошибаются. Знак «+» означает, что разметчики ошибаются всегда. В данной таблице нет разделения на RFTagger и TreeTagger, приведены общие результаты.

Таблица 2. Регулярность ошибок на уровне токена
Table 2. Error regularity at the token level

№	Тип ошибки	Наличие и регулярность ошибки на уровне токена
1	Написание существительных со строчной буквы	+/–
2	Написание глагола с прописной буквы	–
3	Написание несклоняемой формы прилагательного вместо склоняемой	–
4	Раздельное написание одного слова	–
5	Опечатка	+/–
6	Отсутствие пробелов между токенами через дефис	+
7	Отсутствие форматирования (оформление списка в строчку без разделительных знаков)	+/–
8	Отсутствие суффикса «-ег» в притяжательных прилагательных, образованных от географических наименований	–
9	Написание имени существительного-части разорванного слова со строчной буквы и без последующего дефиса	+/–
10	Пробел внутри числительного	–
11	Неверный порядок слов в предложении	+/–

Из Таблицы 2 видно, что всегда правильными на уровне токена оказались теги в следующих типах ошибок:

– написание глагола с прописной буквы;

– написание несклоняемой формы прилагательного вместо склоняемой;

– раздельное написание одного слова;

– отсутствие суффикса «-er» в притяжательных прилагательных, образованных от географических наименований;

– пробел внутри числительного.

Остановимся подробнее на каждом типе ошибки и обратим внимание на закономерности.

1. Написание существительных со строчной буквы.

Существительные, написанные со строчной буквы, размечаются или как глаголы (1), или как прилагательные (2):

(1) […] lösen Sie Fragen mit dem studentenamt, […]

studentenamt – VFIN (финитный глагол),

(2) Zu Beginn des Semesters gibt praktisch jeder Lehrer eine elektronische Version seiner vorlesungen aus.

vorlesungen – ADJD (краткая форма прилагательного)

Можно проследить следующую закономерность: если слова имеют окончания «-en» или «-t», т.е. личные окончания глаголов, то они размечаются как глаголы; если они имеют окончания «-en», «-er», «-e», т.е. окончания прилагательных, то таким токенам присваивается тег прилагательного.

В тех случаях, когда такие формы становятся неотличимы от слов, относящихся к другим частям речи, таггеры размечают их без ошибки внутри токена (3).

(3) Faszinierende reden schlagen ihm aus dem Munde.

reden – VINF (инфинитив глагола)

Вместо токена Reden оказался токен reden, который оба исследуемых разметчика идентифицировали как инфинитивную форму глагола reden – «говорить», а не как существительное Reden – «речи».

2. Написание глагола с прописной буквы.

Здесь ошибки наблюдаются только в TreeTagger, который в некоторых случаях указывает для глагола, написанного с прописной буквы, тег NN – имя существительное. Но во всех этих случаях также можно утверждать, что на уровне токена ошибок нет, поскольку в них глаголы имеют окончание «-en», а это значит, что в написании с прописной буквы такой глагол полностью совпадает с существительным, образованным от соответствующего глагола (субстантивированным глаголом) (4).

(4) […] DHB-Pokal Weiterspielen.

Weiterspielen – NN (имя существительное)

3. Написание несклоняемой формы прилагательного вместо склоняемой.

Оба разметчика во всех примерах с данным типом ошибки определяют часть речи ADJD – краткая форма прилагательного, что является правильной разметкой для токена, который они размечают, т.е. они размечают, что есть, а не то, что должно было бы быть – склоняемая форма прилательного (ADJA) (5).

(5) Bald trabten die klug Tiere den Weg allein.

klug – ADJD

4. Раздельное написание одного слова.

В тех случаях, когда по какой-то причине в тексте слово оказывается разбито на две части, оба разметчика предсказуемо сопоставляют отдельный тег для каждой из этих частей. При этом действуют те же закономерности и те же

ошибки, что и для любых других токенов: например, определение части речи по окончанию (например, существительные Wohlbefinden и Handballvereine в 6 и 7).

(6) […] materielles Wohl befinden und […]

Wohl – N (имя существительное), befinden – VINF (инфинитив глагола)

(7) Im Finale spielen die 2 besten Handball vereine im DHB-Pokal.

Handball – N (имя существительное), vereine – VFIN (финитный глагол),

5. Опечатка

В данном типе ошибок и их частеречной разметке можно наблюдать следующую тенденцию: правильно на уровне предложения размечены те слова, у которых опечатка находится в корне, т.е. морфологические признаки части речи не затронуты (например, прилагательное wichtigen в (8)).

(8) Schleswig-Holstein: THW Kiel gewinnt wigtigen Wettbewerb.

wigtigen – ADJA (склоняемая форма прилагательного)

Там же, где опечатка затрагивала крайние справа буквы (это может быть как конечный элемент окончания, так и суффикса или даже корня), разметчики регулярно ошибались на уровне предложения (например, указательное местоимение diesem в (9) и числительное drei в (10)).

(9) In diesew Obstkisten waren Bananen.

diesew – N (имя существительное)

(10) Drie Kameraden. Erich Maria Remarque

Drie – ADJA (склоняемая форма прилагательного)

В опечатках, меняющих слово на несуществующее, на уровне тега действуют те же закономерности и те же ошибки, что и для любых других токенов: например, определение части речи по окончанию или значимость написания существительных с прописной буквы.

6. Отсутствие пробелов между токенами через дефис.

Предсказуемым образом оба разметчика идентифицируют два токена, написанные через дефис без пробела, как одну единицу (например, глагол и предлог reisen – ohne в примере (11)).

(11) Das Kind […] wird durch die ganze Welt reisen-ohne Paß.

reisen-ohne – ADJA

Поскольку дефисом разделяться могут слова, относящиеся вообще к любым частям речи, то единственной тенденцией, просматриваемой в работе разметчиков, является ориентирование таггера на конечные буквы такого токена-монстра: возможные окончания прилагательного, глагола, суффиксы существительных и т.д. При этом вследствие очень широкой омонимии в морфологических элементах в немецком языке, этот принцип является крайне ненадежным и постоянно приводит к ошибкам в автоматизированной частеречной разметке.

7. Отсутствие форматирования (оформление списка в строчку без разделительных знаков).

Удаление форматирования списка нередко случается при копировании и вставке текстов из различных источников в MS Word или непосредственно в корпус ПАКТ, откуда брался материал для исследования, поэтому данная ошибка встречается в ученических текстах. Оба разметчика нерегулярно допускают там ошибки в определении части речи перечисляемых в списке пунктов, при этом никакой логики или тенденции установить невозможно. Так, например, субстантивированный глагол размечается то как существительное, то как глагол, при этом по-разному в RFTagger и TreeTagger (например, перечисление существительных Lesen, Schreiben, Audieren, Freies (прил.) Sprechen, не оформленное списком или запятыми, в примере (12)).

(12) RFTagger:

Das besteht aus mehreren Modulen: Lesen Schreiben Audieren Freies Sprechen.

Lesen – N, Schreiben – VFIN, Audieren – N, Freies – ADJA, Sprechen – N.

TreeTagger:

Das besteht aus mehreren Modulen: Lesen Schreiben Audieren Freies Sprechen.

Lesen – VVFIN (финитный глагол), Schreiben – NN (имя существительное), Audieren – NN, Freies – NN, Sprechen – NN.

Как и в предыдущих случаях, ориентирование разметчиков на омонимичные окончания «-e» и «-en», встречающиеся и в существительных, и в глаголах, и в прилагательных, приводят к многочисленным ошибкам (например, существительное Gurken в примере (13).

(13) Da kann man verschiedene Gemüsearten anbauen: Gurken Rote Rüben Kürbis usw.

Gurken – VFIN, Rote – ADJA, Rüben – N, Kürbis - N.

8. Отсутствие суффикса «-er» в притяжательных прилагательных, образованных от географических наименований.

В данном типе ошибок оба разметчика почти во всех случаях показывают верную разметку на уровне токена, присваивая ему тег имени собственного. Однако в том случае, когда имя прилагательное образовано не от известного (например, Frankfurt, Sankt-Petersburg), а от малоизвестного города, то разметчики указывают тег имени нарицательного (сравните, например, определение тегов для имен собственных Berlin и Petrosawodsk в (14).

(14) Eine der größten Berlin Universitäten ist die Humboldt-Universität.

Berlin – NE (имя собственное)

Ich studiere an der Petrosawodsk Universität.

Petrosawodsk – NN (имя существительное, нарицательное)

Притяжательные прилагательные от географических наименований образуются добавлением к имени собственному суффикса «-er», поэтому в случае его отсутствия разметчик правильно размечает тот токен, который есть, а не тот, который должен был бы быть в правильном языковом варианте.

9. Написание имени существительного – части разорванного слова со строчной буквы и без последующего дефиса.

В этом типе ошибки почти везде разметчик правильно размечает слова на уровне токенов, то есть то, что есть, а не то, что должно было бы быть (например, mittel – ADJD, klein – ADJD, elite – ADJD, unter – APPR). Ошибки на уровне токена появляются в тех случаях, где существительное написано с маленькой буквы (15):

(15) Es gelang 1804, weiß und Schwarzblech in gesteigerter Menge zu produzieren.

weiß – VVFIN (финитный глагол)

10. Пробел внутри числительного.

Иногда студенты делают пропуски между разрядами внутри одного числа. RFTagger в этом случае видит два числительных (16):

(16) Der Grab ist über 12 500 Jahre alt.

12 – CARD (числительное), 500-CARD

TreeTagger числительные не распознает вовсе, не присваивая им никакого тега.

11. Неверный порядок слов в предложении.

Анализ работы таггеров в предложениях с нарушением правил в порядке слов показывает, что данный тип ошибок влияет на ошибки в частеречной разметке лишь в некоторых случаях. Здесь можно отметить следующие тенденции. Большинство ошибок связаны с неверной разметкой причастия прошедшего времени (второе причастие). Если оно в нарушение правил порядка слов оказывается не в

конце рамочной конструкции, а в ее середине, то он получает тег несклоняемого прилагательного (17):

(17) Ich habe auch eingemacht unglaublich viele Gurken, die dann einfach so herumstanden.

eingemacht – ADJD

Заключение

Таким образом, можно утверждать, что гипотеза о влиянии языковых ошибок студентов на работу автоматического частеречного разметчика подтвердилась. Исследование показало, что можно говорить о взаимосвязи типа языковой ошибки и ошибки в частеречной разметке. Так, абсолютно критичными для обоих разметчиков являются следующие языковые ошибки:

1. Написание несклоняемой формы прилагательного вместо склоняемой;
2. Раздельное написание одного слова;
3. Отсутствие пробелов между токенами через дефис;
4. Отсутствие суффикса «-er» в притяжательных прилагательных, образованных от географических наименований;
5. Пробел внутри числительного.

Регулярно с разной степенью частотности таггеры ошибаются еще в трех случаях, а именно, когда в тексте встречается:

– написание существительных со строчной буквы;

– написание глагола с прописной буквы;

– написание имени существительного, являющегося частью разорванного слова, со строчной буквы и без последующего дефиса.

Во многих случаях на уровне отдельно взятого токена без учета контекста предложения разметчики указывают тег принадлежности к той или иной части речи правильно, размечая то, что есть, а не то, что должно было бы быть, если бы студент не допустил ошибки.

Так, абсолютно правильными на уровне токена оказались теги в следующих типах ошибок:

– написание глагола с прописной буквы;

– написание несклоняемой формы прилагательного вместо склоняемой;

– раздельное написание одного слова;

– отсутствие суффикса «-er» в притяжательных прилагательных, образованных от географических наименований;

– пробел внутри числительного.

В остальных типах ошибок тоже прослеживаются закономерности, большая часть которых связана с определением части речи на основе окончания.

Учет этих закономерностей позволит более эффективно организовать верификацию частеречной разметки в ученическом корпусе на немецком языке. Возможно, исследование будет полезно и для разработчиков частеречных разметчиков.

Список литературы

Котюрова И. А. Исследование инструментов частеречной разметки для создания корпуса ученических текстов // Педагогическая информатика. 2021. № 3. С. 81-89.

Bick E. An Annotated Social Media Corpus for German // Proceedings of the 12th international conference on language resources and evaluation. 2020. Pp. 6127-6135.

Bollmann M. POS tagging for historical texts with sparse training data // Proceedings of the 7th Linguistic Annotation, Sofia, Bulgaria. 2013. Pp. 11-18.

Dıaz-Negrillo A., Meurers D., Valera S., Wunsch H. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT // Language Forum. 2010. 36 (1-2). Pp. 139-154.

Dligach D., Palmer M. Reducing the need for double annotation // Proceedings of the 5th Linguistic Annotation Workshop, Portland, Oregon, USA. 2011. Pp. 65-73.

Heeman P. A. POS Tagging versus Classes in Language Modeling // Proceedings of the 6th Workshop on Very Large Corpora. 1998. URL:

https://aclanthology.org/W98-1121.pdf (Accessed: 22.04.2022).

Horsmann T., Erbs N., Zesch T. Fast or Accurate? – A Comparative Evaluation of PoS Tagging Models // Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology, Duisburg-Essen, Germany. 2015. Pp. 22–30.

Keiper L., Horbach A., Thater S. Improving POS tagging of german learner language in a reading comprehension scenario // Proceedings of the 10th International Conference on Language Resources and Evaluation, Portorož, Slovenia. 2016. Pp. 198-205.

Loftsson H. Correcting a POS-tagged corpus using three complementary methods // Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece. 2009. Pp. 523–531.

Manning C., Schütze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press, 2003. 620 p.

Qian X., Liu Y. Joint Chinese word segmentation, POS tagging and parsing // Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 2012. Pp. 501-511.

Rehbein I., Ruppenhofer J. Detecting annotation noise in automatically labelled data // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017. Pp. 1160-1170.

Sugisaki K., Wiedmer N., Hausendorf H. Building a Corpus from Handwritten Picture Postcards: Transcription, Annotation and Part-of-Speech Tagging // Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan. 2018, Pp. 255-259.

References

Kotyurova, I. A. (2021). Part-of-speech tagging tools applied to a learner corpus, *Pedagogical informatics*, 3, 81-89. *(In Russian)*

Bick, E. (2020). An Annotated Social Media Corpus for German, *Proceedings of the 12th international conference on language resources and evaluation*, Marseille, France, 6127-6135. *(In English)*

Bollmann, M. (2013). POS tagging for historical texts with sparse training data, *Proceedings of the 7th Linguistic Annotation*, Sofia, Bulgaria, 11-18. *(In English)*

Dıaz-Negrillo, A., Meurers, D., Valera, S. and Wunsch, H. (2010).Towards interlanguage POS annotation for effective learner corpora in SLA and FLT, *Language Forum*. 36 (1-2), 139-154. *(In English)*

Dligach, D. and Palmer, M. (2011). Reducing the need for double annotation, *Proceedings of the 5th Linguistic Annotation Workshop*, Portland, Oregon, USA, 65-73. *(In English)*

Heeman, P. A. (1998). POS Tagging versus Classes in Language Modeling, *Proceedings of the 6th Workshop on Very Large Corpora*, 179-187, available at: https://aclanthology.org/W98-1121.pdf (Accessed 22 April 2022). *(In English)*

Horsmann, T., Erbs, N. and Zesch, T. (2015). Fast or Accurate? – A Comparative Evaluation of PoS Tagging Models, *Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology*, Duisburg-Essen, Germany, 22–30. *(In English)*

Keiper, L., Horbach, A. and Thater, S. (2016). Improving POS tagging of German learner language in a reading comprehension scenario, *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia, 198-205. *(In English)*

Loftsson, H. (2009). Correcting a POS-tagged corpus using three complementary methods, *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens, Greece, 523–531. *(In English)*

Manning, C. and Schütze, H. (2003). *Foundations of statistical natural language processing*, Massachusetts Institute of Technology, MIT Press, Cambridge, MA, USA. *(In English)*

Qian, X. and Liu, Y. (2012). Joint chinese word segmentation, pos tagging and parsing, *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 501-511. *(In English)*

Rehbein, I. and Ruppenhofer, J. (2017). Detecting annotation noise in automatically labelled data, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1160-1170. *(In English)*

Sugisaki, K., Wiedmer, N. and Hausendorf, H. (2018). Building a Corpus from Handwritten Picture Postcards: Transcription, Annotation and Part-of-Speech Tagging, *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 255-259. *(In English)*

Конфликты интересов: у автора нет конфликта интересов для декларации.

Conflicts of Interest: the authors have no conflict of interest to declare.

Ирина Аврамовна Котюрова, доцент, кандидат филологических наук, зав. кафедрой немецкого и французского языков, Петрозаводский государственный университет, Петрозаводск, Россия.
Irina A. Kotiurova, Associate Professor, PhD in Philological Sciences, Head of the Department of German and French Languages, Petrozavodsk State University, Petrozavodsk, Russia.

Людмила Владимировна Щеголева, доцент, доктор технических наук, профессор кафедры прикладной математики и кибернетики, Петрозаводский государственный университет, Петрозаводск, Россия.
Liudmila V. Shchegoleva, Associate Professor, Doctor of Technical Sciences, Professor of the Department of Applied Mathematics and Cybernetics, Petrozavodsk State University, Petrozavodsk, Russia.