Сопоставительный анализ наборов тегов для разметки корпусов устной и письменной речи как отражение особенностей функционального стиля

Comparative Analysis of the POS-tagsets in the Written and Speech Corpora as a Reflection of the Functional Style Features

Котюрова И.А.
канд. филол. наук, доцент, зав. кафедрой немецкого и французского языков
ФГБОУ ВО «Петрозаводский государственный университет»
e-mail: koturova@petrsu.ru

Kotiurova I.A.
Candidate of Philological Sciences, Head of the Department of German and French Languages, Petrozavodsk State University.
e-mail: koturova@petrsu.ru

Аннотация
В статье проводится сравнение набора тегов для частеречной разметки лингвистических немецкоязычных корпусов письменной и устной речи. Такой сопоставительный анализ перечня тегов STTS и STTS 2.0, используемых частеречными разметчиками немецких корпусов, позволяет наглядно показать контраст между устной и письменной речью в новом ракурсе: не с позиции языкового стиля, а с точки зрения корпусной разметки.
Ключевые слова: корпусная лингвистика, лингвистический корпус, частеречный разметчик, устная речь, немецкий язык.

Abstract
The article compares Part-of-Speech Tagsets for written and speech corpora in German. Such a comparative analysis of STTS (Stuttgart-Tübingen-TagSet) and STTS 2.0 (Stuttgart-Tübingen-TagSet 2.0) used by part-of-speech tagging of German linguistic corpora allows to visually show the contrast between spoken and written language from a new perspective: not from the position of language style, but from the point of view of corpus tagging.
Keywords: corpus linguistics, linguistic corpora, part-of-speech tagging, **s**poken language, German.

Корпусные технологии, которые в последние годы все более активно входят в инструментарий ученых-лингвистов, привлекают, прежде всего, простотой анализа больших объемов материала и верифицируемостью результатов исследований. Это основные, но не единственные достоинства лингвистических корпусов как материала для научной работы. Однако в представленном ниже исследовании объектом служит не сам корпус, а созданный для работы с ним инструмент разметки.

Корпусом в языкознании принято считать репрезентативное собрание текстов в машиночитаемом формате, имеющее разметку по различным лингвистическим и металингвистическим параметрам [1]. Первыми появились корпуса письменной речи, для которых, в первую очередь, создавались инструменты для ручной, а затем и автоматической частеречной разметки [3]. Ее принято называть POS-разметка, или POS-таггинг, от англ. Part-of-Speech-Tagging. Инструменты разметки, называемые POS-таггеры (Part-of-Speech-Tagger), используют различные наборы тегов, т.е. меток с определением той или иной части речи, присваиваемых таггером каждому элементу текста. Существуют тагсеты (наборы тегов), разработанные как для отдельных языков, так и универсальные, как, например, Universal POS tagset [10]. Конечно, наборы тегов, созданные под конкретный язык, являются более подробными и точными. Одним из наиболее популярных наборов тегов для разметчиков частей речи, применимых к немецкому языку, был и остается набор STTS (Stuttgart-Tübingen-TagSet) [9].

Когда стали появляться корпуса узкой направленности, такие, как исторические корпуса [5-8], содержащие тексты на немецком языке разных периодов, начиная с 9 в., корпуса с транскрипциями современной устной речи [2] и корпуса интернет-коммуникации [4], то оказалось, что традиционный набор тегов, разработанный лингвистами из Штуттгарта и Тюбингена, во многих случаях имеет явные недостатки, что привело к созданию нового варианта тагсета — STTS 2.0., разработанного специально для разметки транскрипций немецкоязычных корпусов **устной** речи. Этот тагсет основан на тех же принципах, что и его предшественник, но некоторые теги, включенные в STTS, скорректированы в плане дефиниции или вовсе удалены, а некоторые добавлены. Изменения существующих тегов незначительны и касаются уточнений некоторых типов частиц, неопределенных местоимений и наречий. Однако включение в новый набор дополнительных 14 тегов, которые учитывают особенности устной речи и ее транскрипции, весьма ярко отражают характеристики этого функционального стиля, отличающие его от стиля письменной речи. Рассмотрим, какие это пункты.

Таблица тегов STTS 2.0 открывается нововведенным тегом AB (Abbruch auf Wortebene) – обрыв слова. Такую разметку получат начатые, но незаконченные слова в примерах *sie **gef**, sie hat **ge***. В тех случаях, когда оборванное слово может быть реконструировано, разметчик дополнительно дает и тег реконструированного слова. Однако во многих случаях даже по контексту нельзя определить начатое, но недоговоренное слово. В этом случае оно размечается тегом AB. Такие нереконструируемые обрывы начатых слов являются характерной чертой спонтанной устной речи.

Изменения коснулись междометий. То, что в первоначальном тагсете обозначалось общим тегом междометия ITJ – Interjektion, в варианте для устной речи разбивается на 4 отдельных тега. Таким образом, удаляется тег IJT, но появляются 4 новых: NGIRR, NGHES, NGAKW, NGONO. Они входят в группу так называемых «неграмматических элементов» − Nicht grammatische Elemente, поэтому все четыре начинаются с аббревиатуры NG.

Специальный тег NGIRR присваивается всем междометиям (Interjektion), сигналам восприятия (Rezeptionssignale) и элементам реакции (Responsive), например, таким, как ***mhm, ach, tja, hmhm.*** Кроме классических междометий („hm“, „ach“, „oh“) он включает образованные от лексем сигналы восприятия или реакции, такие как “gut”, “klar” или “oh Gott”. Отличить такие элементы от полнозначных слов помогает их одиночное положение в речи и то, что они не имеют синтаксических связей с другими лексемами. NGIRR получают соответственно слова-реакции „bitte“, „danke“, „nein“ и „ja“, если только последнее не выступает в функции модальной частицы. Например, тегом

NGIRR будут размечены все следующие элементы устной речи: halt moment moment moment; mm ja; ja klar.

Тег NGHES − Hesitationssignale − обозначает элементы, указывающие на сомнение. Это могут быть любые элементы звучащей речи, которые можно заменить на «эээ...» с интонацией раздумья, нерешительности: ***ähm, öhm, äh.***

Отметим, что, как и в случае с сигналами восприятия и элементами реакции, сигналы сомнения зачастую выражаются в устной речи не в виде полнозначных слов, а в виде отдельных звуков, иногда ни на что не похожих и не фиксируемых в словарях.

Маркировку NGAKW – Aktionswörter – слова-акторы, элементы, называющие действия или чувства во время коммуникации: “lach”, “freu”, „lol“ и т.п. Чаще всего это глагольные корни, инфлективы – seufz, lach и т.п. Такие называния действий или чувств пришли в устную речь из онлайн-коммуникации, где пишущий использует такую возможность выразить свои эмоции в неформальном письменном общении в соц. сетях.

Блок неграмматических элементов закрывает тег NGONO – Onomatopoeia – звукоподражательные слова, используемые, как и все остальные неграмматические элементы, вне синтаксических конструкций. Они отделяются от прочих междометий, поскольку имеют функцию имитации каких-либо звуков. Например, тегом NGONO будут размечены элементы peng, miau, bla bla bla и т.п.

Следующий в ряду добавленных тегов в наборе STTS 2.0 – тег порядковых числительных ORD. В первой версии тегов числительные не делились на категории количественных и порядковых. Для обозначения числительных служил один тег – CARD, который используется только для количественных числительных. В версии STTS 2.0 для разметки числительных используется и второй специальный тег − ORD. В разряд порядковых попадают числительные с аттрибутивной и замещающей функцией: der erste zylinder, zu zweit и т.д.

Тег ORD безусловно актуален в равной степени для текстов устной и письменной речи, однако включен в данное описание, поскольку входит в перечень нововведенных тегов набора STTS 2.0.

Три следующих новых тега касаются частиц (Partikel), поэтому они начинаются с букв PTK: PTKIFG, PTKMA и PTKMWL.

PTKIFG служит для маркировки частиц, указывающих на интенсивность, фокус или степень меры: Intenitäts-, Fokus- und Gradpartikeln. В качестве примера частиц-интенсификаторов можно привести ganz besonders или extrem; в качестве частиц, акцентирующих внимание – echt (echt leicht, echt groß и т.п.), в качестве частиц, указывающих на степень меры – ziemlich, sehr, relativ. Такие языковые единицы имеют функцию модификаторов следующих за ними выражений и в отличие от наречий могут менять положение в предложении только вместе с соотнесенной с ним фразой и перемещаться, занимая место только перед финитным глаголом. Например, в предложении Ich finde es ziemlich dunkel частица ziemlich можно поменять место в предложении только неразрывно со словом dunkel и только заняв место перед финитным глаголом: Ziemlich dunkel finde ich es.

Наречия же могут переставляться в предложении более свободно: Das hat mich extrem belastet можно преобразовать в разговорные варианты Extrem hat das mich belastet или Extrem belastet hat‘s mich. Поэтому в этом случае extrem получит разметку наречия.

PTKMA − Modal- und Abtönungspartikel – тег для модальных частиц, указывающих на тон высказывания. Большинство таких частиц имеют омонимы в других частях речи – ja, halt, schon. Характерным признаком, позволяющим однозначно отличать их от омонимичных частей речи, является принципиальная невозможность перестановки их в предложении и их положение в середине предложения: Das ist ja schön!, Schauen

wir mal das genau an. Das ist aber schlau!

PTKMWL –Mehrwortlexem- Partikeln – так обозначаются частицы, неразрывно связанные с другими лексемами, которые образуют одну многосоставную лексему, не поддающуюся отнесению к какой-либо определенной части речи: [x] noch, immer [+прилагательное в сравнительной степени], schon [+прилагательное в сравнительной степени]. Разбиение этого единого целого на две отдельные лексемы и удаление одной из них привело бы к изменению значения, а, следовательно, недопустимо. Разработчики STTS 2.0 предлагают условно делить такую сложную языковую единицу на «головную лексему», которая может быть представлена различными частями речи и частицу многосоставной лексемы с тегом PTKMWL. Например: Ich wiederhole das immer wieder. Immer – PTKMWL, wieder – ADV.

Новый тег добавился в разряде неопределенных местоимений. В целом, разработчики делят все неопределенные местоимения на определяющие (например, **Man** muss noch andere fragen) и замещающие (**Ich** möchte etwas lesen). Кроме того, разработчики тагсета указывают в теге неопределенного местоимения, может ли оно сопровождаться детерминативом без изменения своей формы или нет. Например, в предложении Andere Kinder bekommen ein Eis неопределенное местоимение andere получит тег PIAT – Attribuierendes Indefinitpronomen, поскольку форма andere (3. Pers Pl. Nom.) в случае добавления определенного артикля поменяет форму на anderen – Die anderen Kinder bekommen ein Eis. А в предложении Anderen Kindern gebe ich ein Eis местоимение anderen получит тег PIDAT – Attribuierendes Indefinitpronomen mit Determinierer, поскольку форма anderen (3. Pers Pl. Dat.) может стоять и с определенным артиклем: Den anderen Kindern gebe ich ein Eis. Такая интерпретация неопределенных местоимений является новой, хотя теги PIAT и PIDAT были и в первой версии набора STTS. В доработанную версию STTS 2.0 добавился тег PIDS – Substituierendes Indefinitpronomen mit Determinierer – замещающее неопределенное местоимение с детерминативом, например, ein bisschen, die beiden. Местоимение jeder в предложении Das hat jeder verstanden также получит тег PIDS.

Нельзя утверждать, что изменения в интерпретации неопределенных местоимений связаны с особенностями устной речи, для транскрипции корпуса которой разрабатывалась версия тагсета STTS 2.0. Но такое утверждение будет верным в отношении следующей новой группы тегов − Satzexterne Elemente – элементов вне предложения. Эта нововведенная группа состоит из двух классов: SEDM и SEQU.

SEDM – Diskursmarker − идентифицирует маркеры дискурса, которые занимают в предложении крайнее место слева и имеют так называемую проецирующую функцию внутри своей семантики, т.е. направлены на следующее за ними высказывание. Например, weil sie wollen ja auch wieder ein königreich werden. Маркеры дискурса придают следующей за ними фразе определенное указание того, как ее понимать. В руководстве по STTS 2.0 приводятся следующие примеры маркеров дискурса:

trotzdem essen gehen macht find ich mmer noch mehr spaß;

also da sprechen alle noch mehr oder weniger platt;

obwohl affengehege kommen doch immer ganz gut an.

Мы видим, что тег SEDM получают токены also, weil, obwohl, trotzdem, т.е. такие единицы языка, которые имеют омонимы среди NGIRR (междометий, респонсивов и сигналов восприятия), союзов придаточных предложений и наречий. В случаях с омонимией с NGIRR отличительным критерием является то, можно ли поставить точку после данного элемента. Если это принципиально возможно, то речь идет о NGIRR, поскольку маркеры дискурса без продолжающей фразы полностью потеряли бы свой смысл. Сравним, например, also в функции сигнала реакции и как маркера дискурса:

– **also** / NGIRR (как реакция на реплику wir machen nun so);

− **also** / SEDM da sprechen alle noch mehr oder weniger platt.

Отграничить маркер дискурса от совпадающего с ним по звучанию и написанию союза, вводящего придаточное предложение, помогает глагол: если он стоит сразу после подлежащего, то мы имеем дело с дискурсным маркером, так как в придаточном предложении глагол будет занимать крайнее правое место. Для сравнения:

obwohl /KON affengehege doch immer ganz gut ankommen

obwohl / SEDM affengehege kommen doch immer ganz gut an

Отличить маркер дискурса от наречия также помогает положение глагола: если дискурсный маркер не влияет на порядок слов, то в случае с наречием наблюдается обратный порядок слов, т.е. сказуемое, выраженное глаголом, стоит перед подлежащим. Сравните использование маркеров trotzdem и also в предложениях выше со следующими двумя примерами:

aber trotzdem (ADV) änderte sich nichts bei ihm

also (ADV) is es meiner für zehn.

Тег SEQU, также присваиваемый элементам вне предложения, является аббревиатурой для Satzexterne Elemente, Question-Tag и обозначает единицы языка, имеющие функцию запроса обратной связи собеседника: ne, wa, gell и др.

Как и маркеры дискурса такие теги-запросы обратной связи напрямую связаны с другими синтагмами, которые могут стоять как перед запросом обратной связи, так и до нее. Например: „Ne? Das können wir so machen" или „Das können wir so machen, ne?". Зачастую функция запроса обратной связи сводится к функции обратить внимание на высказывание, а не получить ответ. Для устной речи характерно, что такой запрос обратной связи с эмфатической функцией может исходить даже не от того, кто произносит ключевую фразу, а от его собеседника, который как бы надстраивает этот сигнал на фразу. Например, один из собеседников говорит: „Das können wir so machen", а второй добавляет: «Ne?». В этом случае ne нельзя трактовать как отрицательную частицу, поскольку ее функция здесь заключается не в отрицании сказанного, а в эмфатическом подчеркивании услышанного. Таким образом, ne в данном случае будет элементом запроса обратной связи.

Исключениями являются „ja" и „oder". Как поясняют разработчики в руководстве к STTS 2.0. [11], „Ja", стоящее до или после высказывания, всегда получает тег NGIRR (междометье, сигнал обратной связи), поскольку его функция запроса обратной связи может быть доказана исключительно интонационно. А „oder", стоящее справа от высказывания, является эллиптическим выражением „oder nicht?" и размечается как союз KON:

das wär ja nicht schlecht oder (KON)

also vor dem nägelbild ja (NGIRR)

Новый тег SPELL вводится для токенов озвучивания букв. Тег SPELL получают отдельно называемые буквы, только если они не являются частью аббревиатур. Например, в предложении "das eff steht für fähigkeiten" „eff" получит тег SPELL.

Наконец, последний из тегов, введенных для разметки транскрипций устных текстов, − это тег UI – Uninterpretierbare Äußerung – неинтерпретируемое высказывание. Данная метка проявляет на формальном уровне еще одну характеристику устной речи. При транскрипции записанных на диктофон записей живой устной речи постоянно возникают ситуации, когда невозможно разобрать, что именно произносит говорящий. В таких случаях, когда слышно, что человек что-то сказал, но невозможно понять, что именно, в транскрипцию ставится условное обозначение „+++", а в частеречной разметке появляется тег UI.

Таким образом, для транскрипции устных текстов оказалось необходимо ввести 14 новых тегов:

1. AB − Abbruch auf Wortebene – обрыв слова.
2. NGIRR – Interjektionen, Rezeptionssignale, Responsive - междометия, сигналы восприятия и реакции.
3. NGHES − Hesitationssignale − элементы, указывающие на сомнение.
4. NGAKW – Aktionswörter – слова-акторы.
5. NGONO – Onomatopoeia – звукоподражательные слова вне синтаксических конструкций.
6. ORD – Ordinalzahlen − порядковые числительные.
7. PTKIFG − Intenitäts-, Fokus- und Gradpartikeln − частицы, указывающие на интенсивность, фокус или степень меры.
8. PTKMA − Modal- und Abtönungspartikel – модальные частицы, указывающие на тон высказывания.
9. PTKMWL – Mehrwortlexem-Partikeln – частицы – часть многосоставной лексемы, не поддающиеся отнесению к какой-либо определенной части речи.
10. PIDS – Substituierendes Indefinitpronomen mit Determinierer – замещающее неопределенное местоимение с детерминативом.
11. SEDM – Diskursmarker − маркеры дискурса.
12. SEQU – Rückversicherungssignal / Question-Tag − единицы языка, имеющие функцию запроса обратной связи собеседника.
13. SPELL – Buchstabiertes − отдельно называемые буквы.
14. UI – Uninterpretierbare Äußerung – неинтерпретируемое высказывание.

Конечно, более точная разметка числительных, частиц и неопределенных местоимений актуальна в равной степени, как для устной, так и для письменной речи. Однако можно утверждать, что многие из приведенного выше перечня теги характерны в большей степени или вовсе исключительно для устных высказываний. Во многих случаях категории POS (Part-of-Speech) в STTS 2.0 не входят ни в один традиционно принятый перечень частей речи немецкого языка, поскольку ориентируются не на грамматику, а имеют целью разметить все данные транскрипции, и чтобы при этом теги были взаимоисключающими. Таким образом, сопоставительный анализ перечня тегов обоих частеречных разметчиков позволяет наглядно показать контраст между устной и письменной речью в новом ракурсе: не с позиции языковых функциональных стилей, а с точки зрения корпусной разметки.

Литература

1. *Котов М.В.* Введение в корпусную лингвистику. Учебное пособие для студентов филологических и лингвистических специальностей университетов. – Прага: Animedia Company, 2014. – 195 с.
2. Archiv für Gesprochenes Deutsch: http://agd.ids-mannheim.de/index.shtml
3. Beale, Andrew David. 1985 Grammatical Analysis by Computer of the Lancaster-Oslo/Bergen (LOB) Corpus of British English Texts. Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics. University of Chicago Press, Chicago, Illinois: 293-298.
4. Dortmunder Chatkorpus: https://www.uni-due.de/germanistik/chatkorpus/
5. Kali-Korpus: http://www.kali.uni-hannover.de/index.php?mmc=2&smc=0
6. Kieler Runenprojekt: http://www.runenprojekt.uni-kiel.de/beschreibung/1/default_eng.htm
7. Mittelhochdeutsche Begriffsdatenbank: http://mhdbdb.sbg.ac.at/index.en.html
8. Referenzkorpus Altdeutsch: http://www.deutschdiachrondigital.de/home/projekt/?lang=de

9. Stuttgart-Tübingen Tagset STTS. URL: https://www.sketchengine.eu/german-stts-part-of-speech-tagset/
10. Universal POS tagset, Web: https://www.sketchengine.eu/tagsets/universal-pos-tags/
11. Westpfahl S., Schmidt T., Jonietz J., Borlinghaus A. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). 2017. 53S. URL: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/6063