Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2025»

23–25 апреля 2025 г.

Clause Linkers for the Ruscon Database: Selecting Criteria and Statistical Evaluation

Svetlana Timoshenko
IITP, Russian Academy of Sciences
Moscow, Russia
timoshenko@iitp.ru

Natalia Serdobolskaya
Institute of Linguistics, Russian Academy of Sciences
Moscow, Russia
serdobolskaya@gmail.com

Irina Kobozeva
Lomonosov Moscow University / Institute of Linguistics, Russian Academy of Sciences
Moscow, Russia
kobozeva@list.ru

Abstract

The study is devoted to compiling a list of clause linkers, or connectives, for the database Ruscon (available at http://ruscon.belyaev.io/). It is aimed at differentiating between complex connectives (like *a ne to* 'else') and occasional combinations of connectives (*a vsledstvie etogo* 'and henceforth'). There is a large number of cases when phonetic, morphosyntactic and semantic properties of connectives do not permit to make this differentiation. In these cases we propose to use the MMI metric (Modified Mutual Information). It rests on the rule of probability multiplication, which allows to estimate the probability of independent events.

Keywords: connective, clause linker, Russian language, subordination, complex sentence, complex connective, multiword expression extraction, collocation, collocation extraction measure, mutual information

DOI: 10.28995/2075-7182-2025-23-XX-XX

Статистические метрики как критерии отбора составных коннекторов (на материале базы Рускон)

Тимошенко С. П.
ИППИ РАН
Москва
timoshenko@iitp.ru

Сердобольская Н. В.
Институт языкознания РАН
Москва
serdobolskaya@gmail.com

Кобозева И. М.
МГУ им. М. В. Ломоносова, Институт языкознания РАН
Москва
kobozeva@list.ru

Аннотация

Исследование направлено на определение словника составных коннекторов для базы данных Рускон (доступна на сайте http://ruscon.belyaev.io/). Цель работы — разграничить составные коннекторы (напр. *а не то*) и окказиональные сочетания (*а вследствие этого*). В тех случаях, когда фонетические, морфосинтаксические и семантические критерии не позволяют решить вопрос о включении составного коннектора в словник, мы предлагаем использовать метрику MMI (Modified Mutual Information). Она опирается на правило умножения вероятностей, позволяющее оценить вероятность независимых событий.

Ключевые слова: коннектор, союз, русский язык, сложное предложение, составной коннектор, извлечение словосочетаний, мера устойчивости словосочетаний, взаимная информация

1 Введение

Одним из базовых вопросов изучения полипредикации является выделение инвентаря союзов, или, шире – коннекторов – единиц, способных соединять клаузы в сложном предложении. В особенности, эта задача актуальна для языков с богатой литературной традицией, которые активно используют комплексы союзов и развивают единицы, промежуточные между самостоятельной клаузой и коннектором, такие как *в расчёте на то что*, *в надежде на то чтобы*. Зачастую бывает трудно разграничить самостоятельные коннекторы, которые должны быть включены в лексикографические источники, и окказиональные сочетания. Настоящая работа ставит цель выявить такие критерии для русских коннекторов на основе статистических метрик, применяемых для извлечения коллокаций.

Материалом исследования служат коннекторы базы данных Рускон (база данных, включающая синтаксическую и семантическую информацию по современным коннекторам русского языка, доступна на сайте http://ruscon.belyaev.io/). Инвентарь коннекторов создан на основе ряда источников, включая словари русского языка МАС (МАС, 1981 1984), БТС (БТС, 1998), (Морковкин (ред.), 2003), (Ефремова, 2004) и Академическую грамматику (Шведова and others, 1980). В отличие от каталога (Богданов and Рыжова, 1997), Рускон основывается на отредактированном словнике и, кроме непосредственно терминологической информации (например, частеречный статус коннектора), содержит данные о значении, употреблении и ряде синтаксических и семантических свойств каждого коннектора.

Часть этого исследования выполнена благодаря поддержке гранта РНФ номер 24-18-00988. Авторы выражают фонду глубокую признательность.

2 Внутриязыковые критерии выделения составных коннекторов

Традиционно союзы делятся на простые и составные, или неоднословные – «нецельноформленные соединения двух или более элементов, каждый из которых одновременно существует в языке и как отдельное слово» (Шведова and others, 1980, с. 716), например, *благодаря тому что*, *даром что*, *не то что*, *оттого что*. При выделении класса составных союзов возникает проблема разграничения единых союзов (*потому что*) и окказиональных сочетаний (*а из-за этого*) – условно говоря, единиц, которые содержатся в памяти как единое целое или порождаются в ходе речепроизводства. В частности, сочетания *так как*, *так что*, *отчего и*, *не то* вводятся в БТС в зоне производных слов под ярлыком «в значении союза»; в МАС *отчего и* отсутствует, а единицы *так как*, *так что* и *не то* признаются сложным союзом; в словаре (Морковкин (ред.), 2003) *так как* и *отчего и* являются отдельными входами, в отличие от *так что* и *не то*; в (Ефремова, 2004) все эти единицы, кроме *отчего и*, вводятся в качестве отдельных входов. Сводный каталог коннекторов и информация о их фиксации в словарях содержится в (Богданов and Рыжова, 1997). Как можно видеть, источники серьезно расходятся в мнениях по поводу трактовки конкретных сочетаний – как в плане их лексико-грамматического статуса, так и, шире, в плане включения их в базовый инвентарь. В Академической грамматике вводится термин «союзные соединения», включающие сочетания с конкретизаторами (слова, которые уточняют значение союза), ср. *потому – и потому – а потому*; *зато – но зато*; *иначе – или иначе* (Шведова and others, 1980, с. 714). Кроме того, вводится термин «аналог союза» для единиц «с квалифицирующими лексическими значениями, которые активно вовлекаются в сферу союзных средств»: *вдобавок*, *кстати*, *лишь*, *потом* и др. Тем самым, в обиход вводится значительное количество коннекторов, которые ранее не рассматривались в словарях, включая также конструкции с вариативным лексическим наполнением – *конечно… но*, *если… то*, *для того чтобы*; *может* (повтор)…. *а может*, *возможно… а может быть* («союзные соединения»).

Настоящая работа ставит целью составить словник коннекторов для базы данных Рускон. База состоит из ряда модулей. В одном из них представлена справочная информация о лексико-грамматических и семантических свойствах коннекторов согласно авторитетным источникам, указанным в разделе 1. Тем самым в отношении коннекторов мы следуем идее введения многокомпонентных единиц в инвентарь союзов, предложенной в монографии (Копотев and Стексова, 2016, с. 41), содержащей ценную информацию по рассматриваемой проблеме. База содержит разметку коннекторов русского языка с точки зрения их синтаксических и семантических свойств, с примерами употребления и ссылкой на их источники. Поскольку речь идет о коннекторах (единицах, способных связывать части сложного предложения), а не о союзах, авторы исходят из функционала данных единиц, а не их частеречного статуса. Такой подход позволяет рассматривать интегрально все релевантные единицы, включая союзы, союзные соединения и аналоги союзов. Тем самым, мы в целом следуем подходу (Шведова and others, 1980). Однако мы предлагаем корректировку списка единиц, представленных в данном источнике, поскольку, как показано выше, многие единицы потенциально могут быть расширены за счет «конкретизаторов», а выделенные конструкции допускают значительную вариативность частей (*если… то для того чтобы – если… то потому что –*

если… то оттого что). Такая вариативность учитывается в подходе О. Ю. Иньковой, где базовой единицей является речевая реализация коннектора, т.е. «та форма, в которой коннектор встречается в данном конкретном высказывании» (Инькова and Кружков, 2018, с. 171). Таким образом, любое сочетание включается в базовый инвентарь и описывается как обладающее собственными уникальными свойствами. Проблема, однако, состоит в том, что число речевых реализаций потенциально бесконечно (Инькова, 2019, с. 24), и в результате, невозможно дать подробное описание синтаксиса и семантики каждой единицы. Между тем, база данных Рускон включает более 30 параметров описания для каждого коннектора, что предполагает ограничение словника. В свете этого, необходимы объективные эксплицитные критерии выделения коннекторов, позволяющие принимать последовательные решения.

Критерии выделения составных коннекторов в типологических работах базируются на эвристике, что коннектор должен вести себя как единая словоформа (а не сочетание слов) с точки зрения различных уровней языка – фонетического, морфологического, синтаксического и семантического (Kortmann, 1997, p. 71-77). Данные критерии входят в более общий круг свойств грамматикализации – явления, при котором лексические единицы переходят в разряд грамматических, или шире, функциональных элементов (таких как союзы) (Hopper and Traugott, 2003; Kouteva et al., 2019). Приведем примеры применения критериев.

Фонетический критерий. Союзы *потому что* и *как будто* фонетически ведут себя как единая словоформа: в повседневной речи *потому что* произносится как [птуштъ, туштъ], *как будто* произносится как [кабут:ъ], союз *так как* — [так:ък], и все три союза имеют одно ударение и произносятся без паузы между частями, т.е. образуют одно фонетическое слово. Безусловно, далеко не все союзы характеризуются единым ударением и отсутствием паузы, поэтому данный критерий работает не как необходимый, но как достаточный для отнесения спорной единицы к союзам. В других случаях просодическое оформление может быть использовано в качестве эвристики за или против того или иного решения. [1]

Семантический критерий. Семантически у *так как* значение причины трудно вывести из семантики составляющих единиц, *так* и *как*; у союза *и то* значение уступки (*всю ночь работал, и то не успел*) не выводится из значений *и* и *то*. Таким образом, для данных единиц работает критерий семантической «неразложимости», или семантической некомпозициональности, который предполагает, что значение целого не может быть выведено из суммарного значения частей.

Данный критерий, однако, затруднительно применить к сочетаниям многофункциональных союзов и частиц *а*, *да*, *и*, *но*, *же*, *даже* и (реже) *-то* с имеющими ограниченный набор значений «конкретизаторами» (согласно (Шведова and others, 1980)): *а* / *и* / *но в результате*, *если и* / *же*, и т.п.). Зачастую значение всего комплекса является композициональным, однако происходит десемантизация союза/частицы, например, *и* в составе *и однако* может иметь только значение логической конъюнкции (но не сопоставления или временного следования). При этом *однако* выступает в своем обычном противительном значении «ненормального следствия». Такое ограничение на значение сочинительного союза может быть проинтерпретировано как десемантизация в рамках составного коннектора или же как ограничение на лексическую сочетаемость, стандартным образом возникающее при контактном употреблении лексем. От выбора интерпретации зависит решение о признании *и однако* единым коннектором.

Морфологический критерий предполагает, что коннектор (практически) не имеет способности к словоизменению и выступает в одной фиксированной форме (fossilization в теории грамматикализации); соответственно, его части, даже если они граматикализовались из знаменательной части речи, утрачивают способность к словоизменению. Например, существительное фиксируется в определенной падежной форме и не допускает изменения по числу в составе коннекторов *в тех целях чтобы – *в той цели чтобы*, но *с той целью чтобы – *с теми целями чтобы*; наречия часто закрепляются в форме сравнительной степени (при глаголах речи): *точнее говоря – *точно говоря*, *вернее сказать – *верно сказать*, *проще сказать – *просто сказать* и т.д. Заметим, одна-

[1] Интересную проблему представляют собой коннекторы, допускающие разрыв, напр. *потому... что, затем... чтобы* и др. Вопрос об их самостоятельности требует отдельного рассмотрения. При этом, неразрывный вариант таких единиц представляет собой одно фонетическое слово и, соответственно, включается в базу.

ко, что коннектор может сохранять способность различать некоторые формы, например *коротко говоря – короче говоря* (но не **коротко сказав*), *по правде говоря – по правде сказать* (но не **по правде сказав*) и т.п.

Синтаксический критерий. Единицы, входящие в состав союзов, не могут быть модифицированы, ср. *потому что – *по тому самому что*; союзы не разрываются[2]: *так как он пришел – *так он как пришел*, *так же как он пришел* (такое сочетание возможно, однако не имеет причинной семантики). Части коннектора не могут быть опущены, ср. *а то*; точнее, опущение частей меняет семантику отношения между клаузами, и следовательно единица *а то* трактуется как единый коннектор. Семантика повторяющихся союзов *или... или, либо... либо* не вполне идентична значению одиночного *или* и *либо*, т.к. обычно вводит непересекающиеся альтернативы; соответственно, такие сочетания признаются самостоятельными коннекторами. Особую проблема опущения представляют собой т.н. рамочные коннекторы, т.е. коннекторы, смысловая часть которых находится в зависимой предикации, а главная предикация содержит сочинительный союз, частицу или местоимение, отсылающее к ситуации в зависимой клаузе: *не только... но и; ещё; а также; не то чтобы... а; но; просто* (обычно список таких средств является открытым). Такие коннекторы ограниченно допускают опущение второй части, чаще всего при маркированном просодическом оформлении. В базе они включены в виде входов с вариативным квази-коррелятом. Трактовка рамочных коннекторов представляет собой особую проблему, которая находится за рамками настоящего исследования; ниже мы рассматриваем только случаи, когда многокомпонентной является основная часть коннектора (в составе зависимой клаузы).

Таким образом, части коннектора при грамматикализации обычно подвергаются изменениям, которые связаны с интеграцией частей коннектора в единое целое, неразложимое с точки зрения тех или иных уровней языка. Если же такие процессы не происходят, это обычно означает, что рассматриваемая единица является свободным сочетанием отдельных лексем.

В результате обработки данных источников база Рускон включала 700 составных коннекторов. Из них относительно 591 коннекторов решение о включении (или удалении из базы) было принято на основании внутриязыковых критериев; для остальных 109 данные критерии не позволяют принять объективное решение.

3 Статистический анализ коннекторов

3.1 Сравнение метрик, опирающихся на идею взаимной информации

Для решения вопроса о степени слитности 109 оставшихся кандидатов использовались статистические инструменты — меры оценки устойчивости словосочетаний. Эти меры позволяют ранжировать словосочетания и сравнивать их между собой. Они очень разнообразны: в диссертации (Evert, 2005) представлено около 30 метрик (этот обзор широко известен благодаря удачному воплощению в справочном сайте `collocations.de`), в статье (Su et al., 2024) сравнивается 16. Наиболее популярными метриками являются: хи-квадрат (Chi-square), t-тест (T-score), коэффициент Дайса (Dice's coefficient), отношение правдоподобия (Log-likelihood ratio), а также варианты оценки взаимной информации — точечная взаимная информация (Pointwise Mutual Information или PMI), кубическая взаимная информация (Cubic Mutual Information или MI^3) и другие. Все метрики рассчитываются на основе данных о частотности элементов в том или ином корпусе. В работе (Deng and Liu, 2022) сравнивается 7 различных метрик на материале разных типов устойчивых словосочетаний и разных корпусов. В среднем лучшие результаты показывают отношение правдоподобия и кубическая взаимная информация. Сравнению метрик на материале русского языка посвящена статья М. Хохловой (Khokhlova, 2018). Лингвистическим материалом в этой рабое выстпают двухсловные сочетания "прилагательное + существительное". Найденные в корпусе сочетания ранжируются с помощью 13 разных метрик, а потом верхние 100 позиций в каждом списке оцениваются экспертами и проверяются по словарям. В этом эксперименте лучшими метриками оказываются MI^3, t-тест, коэффициент Фишера и коэффициент Пуассона. Можно сказать, что MI^3 показывает хорошие результаты на материале разных языков.

[2]В настоящем исследовании мы не рассматриваем разрывные коннекторы, например *да и* (*да он и не хотел этого*), *если и, пока не* и проч.

Лингвистически содержательные интерпретации метрик расходятся: так, согласно результатам, изложенным в (Krenn, 2000), PMI и коэффициент Дайса лучше подходят для поиска высокочастотных коллокаций, а отношение правдоподобия — для низкочастотных. А в более современном учебнике (Копотев, 2014) утверждается противоположное: «MI лучше ищет довольно редкие коллокации узкой тематической области (например, термины, составные названия компаний или сочетания имени и фамилии), а t-score лучше справляется с высокочастотными общеязыковыми "эквивалентами слова" (сложные предлоги, вводные конструкции и т. п.)».

Особенность задачи выявления единых коннекторов состоит в том, что необходимо сравнивать между собой выражения разной длины, от двухсловных (*но и*, *а не*, *как будто*) до пятисловных (*в сравнении с тем как*, в *расчёте на то что*). Формула расчета отношения правдоподобия применима только к двухсловным словосочетаниям (Ramisch, 2014, p. 215). Выбор среди остальных представляет трудность, т. к. известные нам исследования, выполненные на материале русского языка, посвящены биграммам, как и упоминавшаяся выше работа (Khokhlova, 2018), их результаты не могут служить главным основанием при выборе метрики для нашей задачи. Мы начали экспериментировать с MI^3 как с одной из самых популярных и хорошо себя зарекомендовавших в разных задачах. Мы адаптировали ее для словосочетаний произвольной длины по той же логике, что и реализация PMI в программе mwetoolkit - см. формулу в работе (Ramisch et al., 2010). Результаты экспериментов с PMI и MI^3 показывают, что обе они не позволяют корректно сравнивать между собой словосочетания разной длины. Средние значения метрик для двухсловных словосочетаний достоверно ниже, чем для трехсловных, а для трехсловных - ниже, чем для четырехсловных, и т. д. Иными словами, любое более длинное словосочетание оказывается более устойчивым, чем более короткое. Поэтому мы предлагаем новую формулу, модификацию взаимной информации(далее - MMI, Modified Mutual Information) 1. Значения этой метрики для лингвистически однотипных двухсловных и трехсловных устойчивых выражениях оказываются близки (Тимошенко, 2024). В таблицах 1 и 2 приведены средние и медианные значения метрик для словосочетаний разной длины, полученные на материале 242 коннекторов. Программная реализация экспериментов и данные

	2 слова	3 слова	4 слова	5 слов
PMI	7.79	22.11	40.28	61.55
MI^3	22.05	29.9	38.68	53.68
MMI	18.5	19.47	19.95	19.06

Таблица 1: Средние значения метрик для словосочетаний разной длины

	2 слова	3 слова	4 слова	5 слов
PMI	7.61	21.91	40.14	61.12
MI^3	22.92	30.88	39.24	53.19
MMI	18.3	18.87	20.05	19.41

Таблица 2: Медианные значения метрик для словосочетаний разной длины

доступны по адресу: `https://github.com/therolinguist/mmi_and_other_measures`

3.2 Метрика MMI

$$MMI(collocation) = ln(frec_{collocation} \div \frac{w_1 \times ... \times w_i}{freq^i_{pattern}}) \quad (1)$$

где

$freq_{collocation}$ - наблюдаемая абсолютная частота словосочетания в корпусе;

$freq_{pattern}$ - наблюдаемая абсолютная частота морфологического паттерна, который ему соответствует;

$w_1 \ldots w_i$ - наблюдаемые абсолютные частоты словоформ, входящих в состав словосочетания, на "своей" позиции в паттерне.
Пример расчета метрики см. в Приложении А. В качестве источника количественных данных использовался основной корпус НКРЯ (НКРЯ, 2024).

Морфологический паттерн — это последовательность частеречных характеристик входящих в него слов. Для выражения *в сравнении с тем как* он выглядит так: "1-е слово — предлог, 2-е — существительное, 3-е — предлог, 4-е — местоименное наречие, 5-е слово — союз" ((PR)(S)(PR)(SPRO)(CONJ) на языке запросов Национального корпуса русского языка [(НКРЯ, 2024)). Однако если мы посмотрим на морфологическую разметку вхождений этого выражения в основном корпусе НКРЯ, то обнаружим, что из 12 вхождений только у 4 характеристики будут соответствовать этому паттерну. В 8 оставшихся случаях *как* будет считаться не союзом, а местоименным наречием. (Здесь и далее числа отражают состояние Основного корпуса на июль 2024 года. С тех пор корпус пополнился новыми текстами, поэтому многие числа изменились). В большей части основного корпуса разрешение морфологической омонимии выполнено автоматически с помощью программы rubic (Lyashevskaya et al., 2023). Тем не менее, мы не считаем подобные "биения" системы ошибками: как говорилось выше, коннекторы находятся в процессе грамматикализации, а этот процесс происходит постепенно. Непоследовательности в ответах системы, выбирающей предпочтительный морфологический разбор, могут отражать переходное состояние отдельных элементов выражения. Эта гипотеза представляется захватывающе интересной: дальнейшие исследования сулят статистические открытия в области грамматикализации. Чтобы компенсировать разночтения разметки, мы использовали в паттернах дизъюнкции: итоговый паттерн *в сравнении с тем как* допускает в качестве 5-го слова не только союз и местоименное наречие, но и частицу, вводное слово и наречие — (PR)(S)(PR)(SPRO)(ADV|PARENTH|ADVPRO|CONJ|PART). В общей сложности мы использовали 86 паттернов.

Чтобы на основе статистического показателя понять, какие спорные случаи следует отнести к единым коннекторам, необходимо выбрать пороговое значение. Для этого мы рассчитали значение метрики для 242 выражений, ранее отнесенных к классу единых коннекторов на основе типологических критериев (см. выше). В качестве порогового значения мы использовали нижнюю границу стандартного отклонения. Для двухсловных выражений она составляет 16.39, для трехсловных — 16.66, для четырехсловных — 17.9 и для пятисловных — 17.03. Было рассмотрено 109 спорных случаев. Для них также были рассчитаны значения метрики MMI. У 84 единиц они оказались больше порогового значения.

В нашей выборке нередки случаи, когда единицы большей длины включают в себя единицы меньшей. Например, трехсловные единицы *а тем более*, *и тем более*, *тем более что* включают в себя слова, совпадающие с самостоятельным коннектором *тем более*. Результаты поискового запроса к корпусу "словоформа **тем**, на расстоянии 1 — словоформа **более**" включают вхождения всех трех более длинных единиц. Поэтому корректный показатель абсолютной частотности вхождений *тем более* получается путем вычитания количества вхождений более длинных единиц из количества вхождений *тем более*. Такая поправка была сделана во всех случаях, когда более длинные единицы включают в себя элементы, соответствующие более коротким.

Предлагаемый подход к оценке устойчивости словосочетаний имеет свои ограничения. Строго говоря, он оценивает взаимную обусловленность словоформ в цепочке; семантика же словосочетаний никак не контролируется. Поэтому в случае омонимичных конструкций результат оказывается смазан. Рассмотрим коннектор *а также*:

(2) *Кроме того, что кандидаты являются тезками, у них есть еще две общие черты: оба они пришли в университет в 1966 году,* ***а также*** *оба являются докторами физико-математических наук.* [НКРЯ: Анастасия Гулина. Ректорство на день рождения // «Богатей» (Саратов), 2003.10.23]

Чтобы оценить частотность именно коннектора, нужно отобрать только те примеры, где справа от *а также* как минимум находится простое предложение, а в идеале — присоединяется. Иными

словами, нужна информация о синтаксической связи между вершиной клаузы и *а также*. Несмотря на то, что в основном корпусе в 2024 году появилась возможность накладывать синтаксические ограничения на запрашиваемые словосочетания, в данном случае она не может помочь нам отобрать вхождения потенциальных коннекторов. Причин две. Во-первых, сочинительные связи устанавливаются автоматически хуже других за счет своей вариативности. Во-вторых, даже если бы точность их определения была близка к 100%, принятый в основном корпусе формализм не позволяет различить сочинение предложений, как в примере (2) и сочинение сказуемых, как в примере 3:

(3) *Тем не менее, это очень страшная штука, которую хотелось бы, чтоб мы сами,* ***а также*** *наши политики стремились не допускать в будущем во благо следующего поколения.*[НКРЯ: Форум: 17 мгновений весны (2005-2010)]

В обоих случаях союз *а* будет зависеть от глагола (*назначил* и *зарабатывали* соответственно), а сами глаголы будут размечены как подчиненная клауза. Различить такие случаи может только человек. Принимая во внимание огромное количество примеров, можно выполнить экспертную оценку для ограниченной случайной выборки примеров, например, для 1000, а потом экстраполировать результат на абсолютные цифры. Т. е. если в корпусе нашлось 59995 вхождений *а также*, и в случайно отобранном их подмножестве окажется лишь 5% примеров, в которых действительно представлен коннектор, то можно рассчитывать MMI, принимая абсолютную частотность коннектора равной 3000. В данном исследовании мы таких поправок не делали.

4 Заключение

В ходе разработки базы данных русских коннекторов Рускон возникает необходимость отбора словника составных коннекторов как единиц лексикографического описания. На основании внутриязыковых критериев (фонетические, морфосинтаксические свойства коннектора и семантическая композициональность) не всегда удается разграничить устойчивые единицы (напр., *а и то*), которые являются единицами языка, и окказиональные сочетания, которые порождаются говорящим в процессе реализации высказывания (напр., *а вследствие этого*).

Мы предприняли попытку статистически оценить устойчивость составных коннекторов. Для этого мы использовали метрику MMI (модифицированную взаимную информацию). Ее главное достоинство состоит в том, что она позволяет сравнивать между собой словосочетания разной длины. Значения метрики были посчитаны для 351 выражения. В качестве источника количественных данных использовался основной корпус НКРЯ. 242 выражения из 351 были классифицированы как коннекторы на основе внутриязыковых критериев. Среди них были двухсловные, трехсловные, четырехсловные и пятисловные единицы. Для каждой длины мы рассчитали стандартное отклонение значений и приняли его нижнюю границу в качестве порога для классификации спорных случаев. На основе этого порогового значения мы классифицировали 109 кандидатов в коннекторы. В итоге 84 из них были включены в базу.

Главный недостаток предложенного способа классификации состоит в том, что оценивается устойчивость последовательности словоформ, а значение этой последовательности никак не проверяется. Между тем, в языке иногда сосуществуют коннекторы и омонимичные им свободные словосочетания (ср. *а также*). Корректный расчет значений статистических метрик в этих случаях требует предварительной ручной классификации примеров.

References

Yaochen Deng and Dilin Liu. 2022. A multi-dimensional comparison of the effectiveness and efficiency of association measures in collocation extraction. *International Journal of Corpus Linguistics*, 27(2):191–219.

Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, Stuttgart University.

P. J. Hopper and E. C. Traugott. 2003. *Grammaticalization*. Cambridge University Press.

Maria Khokhlova. 2018. Similarity between the association measures: a case study of noun phrases. // *RASLAN*, P 21–27.

B. Kortmann. 1997. *Adverbial Subordination: A Typology and History ofAadverbial Subordinators Based on European Lnguages*. Mouton de Gruyter, Berlin; New York.

T. Kouteva, B. Heine, B. Hong, H. Long, H. Narrog, and S. Rhee. 2019. *World lexicon of grammaticalization*. Cambridge University Press.

Brigitte Krenn. 2000. Empirical implications on lexical association measures. // *Proceedings of The Ninth EURALEX International Congress*, P 359–371.

O.N. Lyashevskaya, I.A. Afanasev, S.A. Rebrikov, Y.A. Shishkina, E.A. Suleymanova, I.V. Trofimov, and N.A. Vlasova. 2023. Disambiguation in context in the russian national corpus: 20 yeas later. // *Proceedings of the International Conference "Dialogue*, volume 2023.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: A framework for multiword expression identification. // *LREC*, volume 10, P 662–669. Valletta.

C. Ramisch. 2014. *Multiword Expressions Acquisition: A Generic and Open Framework*. Theory and Applications of Natural Language Processing. Springer International Publishing.

Qi Su, Chen Gu, and Pengyuan Liu. 2024. Association measures for collocation extraction: Automatic evaluation on a large-scale corpus. *International Journal of Corpus Linguistics*, 29(1):59–86.

БТС. 1998. *Большой толковый словарь русского языка под ред. С. А. Кузнецова*. Норинт, СПб.

С. И. Богданов and Ю. В. Рыжова. 1997. *Русская служебная лексика. Сводные таблицы*. Издательство Санкт-Петербургского Государственного Университета.

Т.Ф. Ефремова. 2004. *Толковый словарь служебных частей речи русского языка*. Астрель, АСТ, Москва.

О.Ю. Инькова and М.Г. Кружков. 2018. Метод описания структуры неоднословных коннекторов в надкорпусных базах данных. *Системы и средства информатики*, 28(4):168–181.

О.Ю. Инькова. 2019. Структура коннекторов: лингвистические методы описания. // О.Ю. Инькова, *Структура коннекторов и методы ее описания*, Р 5–47. ТОРУС ПРЕСС, Москва.

М.В. Копотев and Т.И. Стексова. 2016. *Исключение как правило: переходные единицы в грамматике и словаре*. Языки славянских культур: Рукописные памятники Древней Руси, Москва.

М. В. Копотев. 2014. *Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов. Корпусная лингвистика. Филология*. Animedia Company.

МАС. 1981-1984. *Словарь русского языка: в 4-х т*. Русский язык, Москва.

В.В. Морковкин (ред.). 2003. *Объяснительный словарь русского языка : структурные слова: предлоги, союзы, частицы, междометия, вводные слова, местоимения, числительные, связочные глаголы*. Астрель, АСТ.

НКРЯ. 2024. Национальный корпус русского языка. https://ruscorpora.ru.

С. П. Тимошенко. 2024. Оценка словосочетаний разной длины с помощью точечной взаимной информации. *Труды конференции ИТиС 2024*.

Н.Ю. Шведова et al. 1980. *Русская грамматика. В двух томах.* Наука, Москва.

Translation of references in Russian

References

Bogdanov S.I., Ryzhova YU.V. (1997), Russian functional words. Summary tables [Russkaya sluzhebnaya leksika. Svodnye tablicy], Saint-Petersburg: SPbGU.

BTS: Kuznetsov S.A. (ed.) (1998), A big explanatory dictionary of Russian [Bol'shoi tolkovyi slovar' russkogo yazyka], Saint-Petersburg: Norint.

Efremova T.F. (2004), An explanatory dictionary of functional words of the Russian language [Tolkovyi slovar' sluzhebnykh chastei rechi russkogo yazyka], Moscow: Astrel', AST.
In'kova O.Yu., Kruzhkov M.G. (2018), A method of description of the structure of multiword connectives in supracorpora databases [Metod opisaniya struktury neodnoslovnykh konnektorov v nadkorpusnykh bazakh dannykh], Systems of means of informatics [Sistemy i sredstva informatiki]. Vol. 28 № 4, pp. 168-181.
In'kova O.Yu. (2019), A structure of connectives and methods of its description [Struktura konnektorov i metody ee opisaniya], Moscow: Torus Press.
Kopotev M.V., Steksova T.I. (2016), Exception is a rule. Transmitional items in grammar and dictionary [Iskljuchenie kak pravilo. Perehodnye edinicy v grammatike i slovare]. Moscow: Jazyki slav'anskoj kul'tury.
MAS: Evgen'eva A.P. (ed.) (1981–1984), A dictionary of Russian [Slovar' russkogo yazyka]: four volumes, AN SSSR, In-t rus. yaz.; 2-e izd., ispr. i dop., Moscow: Russkii yazyk.
Morkovkin V.V. (ed.) (2003), An explanatory dictionary of the Russian language: functional words: prepositions, conjunctions, particles, interjections, parentheticals, pronouns, cardinals, copulas: approx. 1200 words [Ob"yasnitel'nyi slovar' russkogo yazyka : strukturnye slova: predlogi, soyuzy, chastitsy, mezhdometiya, vvodnye slova, mestoimeniya, chislitel'nye, svyazochnye glagoly: okolo 1200 edinits], Gos. in-t rus. yaz. im. A. S. Pushkina, izd. 2-e, ispr., Moscow: Astrel', AST.
NKRYA. 2024. Russian National Corpus. [Nacional'nyi korpus russkogo yazyka]
Timoshenko S. P. (2024), Pointwise mutual information as an evaluation metric for collocations of different length [Ocenka slovosochetanij raznoj dliny s pomoshh'yu tochechnoj vzaimnoj informacii], Proceedings of the ITAS Conference 2004 [Trudy konferencii ITiS 2024.]
Shvedova N.Yu. (ed.) (1980; 2005), Russian grammar [Russkaya grammatika], Moscow: Nauka.

Приложение А. Пример расчета метрики MMI

Коннектор: *тем более* (классифицирован как единый на основе внутриязыковых критериев)
Морфологический паттерн:

(SPRO) (ADV | PARENTH | ADVPRO | CONJ | PART)

Абсолютная наблюдаемая частотность коннектора составляет 35323 (результат запроса "словоформа **тем**, на расстоянии 1 от нее словоформа **более**").

После вычитания количества вхождений более длинных единиц *а тем более* (2360 вхождения), *и тем более* (2572 вхождения), *тем более что* (15649 вхождения) абсолютная наблюдаемая частотность составляет 14742.

Абсолютная наблюдаемая частотность словоформы *тем* на первой позиции паттерна составляет 143588 (результат запроса "словоформа **тем** с грамматической характеристикой "местоименное существительное" , на расстоянии 1 от нее слово, которое является либо наречием, либо вводным словом, либо местоименным наречием, либо союзом, либо частицей").

Абсолютная наблюдаемая частотность словоформы *более* на второй позиции паттерна составляет 53077 (результат запроса "первое слово — местоименное существительное, на расстоянии 1 от него словоформа **более**, с одной грамматической характеристикой из набора: наречие, вводное слово, местоименное наречие, союз, частица").

Абсолютная наблюдаемая частотность морфологического паттерна "первое слово — местоимен-ое существительное, на расстоянии 1 от него слово, которое является либо наречием, либо вводным словом, либо местоименным наречием, либо союзом, либо частицей" составляет 7099859.

$$MMI(\textit{тем более}) = ln(frec_{collocation} \div \frac{w_1 \times w_2}{freq^2_{pattern}}) = ln(14742 \div \frac{143588 \times 53077}{7099859^2}) = 18.39 \quad (4)$$