Università
Ca'Foscari
Venezia

Corso di Laurea Magistrale

in

Lingue e letterature europee, americane e postcoloniali

Tesi di Laurea

Linguistica dei corpora e corpora linguistici: analisi teorica e indagine pratica per lo studio della lingua russa contemporanea di Internet

Relatrice
Prof.ssa Luisa Ruvoletto

Correlatore
Prof. Pavel Duryagin

Laureanda
Federica Longo
n. matr. 882140

Anno Accademico 2020 / 2021

Indice

Корпусная лингвистика и лингвистические корпусы: теоретический анализ и практическое исследование для изучения современного русского языка в Интернете

Интернет представляет собой необходимое средство, чтобы установить контакт с остальным миром. Во время пандемии коронавируса наша повседневность глубоко изменилась, и многие из этих изменений стали возможны благодаря Интернету. Например, смарт-воркинг (от англ. *smart working*) позволил миллиардам людей выполнять свои рабочие задачи из дома. Интернет позволил остаться на связи с остальным миром через социальные сети, заниматься спортом и делать покупки онлайн. Только посредством Интернета можно скачать Грин Пасс (от англ. *Green Pass*).

Это лишь некоторые из наиболее очевидных примеров, показывающих, как Интернет влияет на нашу жизнь, но есть и другие, менее очевидные аспекты, такие как языковые изменения.

Эти изменения вызваны не только контактами между различными странами, но и самим Интернетом, который вынуждает пользователей общаться в уменьшенных времени и пространстве. Использование Интернета привело к рождению нового, сетевого языка, так называемого «нетспика» (от англ. *Netspeak*), который имеет характеристики как письменного, так и разговорного языка.

Данная работа представляет собой теоретическое исследование истории и методов корпусной лингвистики и в то же время практическое исследование современного русского языка Интернета.

Первая глава посвящена корпусной лингвистике как разделу традиционной лингвистики. В частности, в главе приводятся разные определения корпусной лингвистики и анализируются два основных подхода дисциплины: *corpus-based* ‘основанный на корпусе’ и *corpus-driven* ‘управляемый корпусом’. *Corpus-based* подход использует корпусы, чтобы опровергать или подтверждать догадки или ожидания лингвиста; напротив, при *corpus-driven* подходе языковед анализирует данные корпуса и на них обосновывает лингвистические правила и приходит к выводу без предвзятостей или первоначальных ожиданий.

Затем было рассмотрено историческое развитие корпусной лингвистики в англоязычных странах, в Италии и в России.

Корпусную лингвистику часто считают молодым разделом лингвистики, получившим развитие в 80-х годах, но в действительности её истоки гораздо старше.

Ещё в первой половине XV века такие учёные, как Томас Гибсон и Джон Мербеке создали языковые аналитические индексы, так называемые *concordances* священных текстов, таких как Ветхий и Новый Завет . Во второй половине XVIII века были созданы первые текстовые базы данных, касающиеся изучения языка и речи; примером этого является база данных Чарльза К. Фрайса, который проанализировал около 70 британских и американских комедий для изучения английских глаголов *shall* и *will*.

В 1964 году возник первый современный электронный корпус — *Brown Corpus of American Written English*, а впоследствии появились *Lancaster-Oslo-Bergen Corpus* в 70-х годах и *COBUILD English Language Dictionary* в 80-х годах. Другими словами, корпусная лингвистика в англоязычных странах проложила дорогу остальному миру.

Напротив, трудно описать историю этой дисциплины в Италии, особенно в период между 50-ми и 80-ми годами. Согласно Ф. Сабатини, почётному президенту «Академии делла Круска», истоки корпусной лингвистики в Италии восходят к временам Данте Алигьери. Данное исследование показало, что многие учёные в англоязычных и русскоязычных странах относят истоки дисциплины ко времени задолго до Данте, то есть, к периоду, когда жил Святой Антоний Падуанский (1195-1231), автор *Concordantiae Morales*, первого аналитического индекса Вульгаты, латинского перевода Библии Святого Джироламо V века.

Наконец Роберто Буза и Томас Уотсон-старший, основатель *IBM* (*International Business Machines Corporation*) создали первый электронный индекс, так называемый Индекс Томистикус, в котором были собраны все латинские тексты Святого Фомы Аквинского.

Первый итальянский речевой корпус – *Corpus Stammerjohann* – появился в 1965 году. Впоследствии, в 70-х годах, были созданы *Corpus LIF* (‘Корпус частной лексики современного итальянского языка’) и *Corpus LABLITA* (‘Корпус

лингвистической лаборатории факультета итальянистики Флорентийского университета').

В 90-е годы произошло полное созревание дисциплины, и благодаря информатизации 2000-х корпусная лингвистика двигалась вперёд семимильными шагами.

До сих пор в Италии нет Итальянской ассоциации корпусной лингвистики, и это показывает, что в стране такой подход к дисциплине ещё не приобрел полной автономии по сравнению с традиционной лингвистикой.

Что касается России, то в этой стране корпусная лингвистика является относительно недавней методологией исследования, получившей распространение в 2000-е гг.

Уппсальский корпус представляет собой первый русскоязычный корпус, созданный в 80-е и 90-е годы в Швеции на кафедре славянского факультета Уппсальского университета. Несмотря на то, что это первый корпус русского языка, следует отметить, что это не российский проект, а шведский. «Компьютерный корпус текстов русских газет конца XX века» является первым русскоязычным корпусом, созданным в России Лабораторией общей и компьютерной лексикологии и лексикографии МГУ им. М.В. Ломоносова, в период с 2000 по 2003 год.

В 2003 году возник проект по созданию русскоязычного эквивалента Национального корпуса британского языка (*British National Corpus*), а уже в 2004 году появился Национальный корпус русского языка (НКРЯ), который на сегодняшний день является главным и крупнейшим русскоязычным корпусом, содержащим более 1 миллиарда слов.

Несмотря на то, что корпусная лингвистика поздно начала развитие в России, за 20 лет она утвердилась быстро и решительно.

Вторая глава данного исследования посвящена лингвистическим корпусам как средствам анализа. В ней приводятся некоторые определения «лингвистических корпусов». В этой главе также рассматриваются основные характеристики корпуса, такие как компьютеризация и аутентичность лингвистических данных, репрезентативность, сбалансированность,

определённый размер корпуса, а также повторяемость и воспроизводимость результатов поиска.

Далее в главе анализируются основные классификации, такие как неразмеченные корпусы и корпусы с морфологической, синтаксической, семантической разметкой, или с разметкой ошибок. Кроме того, существуют ещё и многоцелевые корпусы обычно национальные корпусы, такие как Национальный корпус русского языка, и специализированные корпусы, ориентированные на определённую разновидность языка.

Синхронические корпусы содержат тексты, относящиеся к определённому, обычно недолгому, отрезку времени, для анализа какого-либо периода развития языка. Напротив, диахронические корпусы позволяют исследовать лингвистическое развитие языка, языковую вариативность или употребление лексики на протяжении длительного времени.

Динамические, или мониторные, корпусы регулярно пополняются новыми данными, а статические – это корпусы конечного размера, не подлежащие обновлению.

Письменные корпусы собирают только письменные источники, такие как дневники, статьи, письма, учебники, журналы. Устные или речевые корпусы основываются на расшифрованных устных источниках, таких как телевизионные и радиопередачи, разговоры, беседы, монологи, доклады, лекции и т.д. На самом деле корпусы часто содержат текстовые данные как из письменных, так и из устных источников, и по этой причине их называют «смешанными корпусами».

Наконец, мультимедийные корпусы содержат видеоматериалы и звуковые записи коммуникативных действий, позволяющие исследовать язык жестов, вербальный и невербальный способ общения и человеческие эмоции.

Одноязычные корпусы образованы одноязычными текстовыми данными; напротив, двуязычные и многоязычные корпусы содержат множество текстов, написанных на двух и более языках.

Наконец, учебные корпусы обладают всеми характеристиками традиционных корпусов, но, в отличие от них, устные и письменные данные создаются не носителями языка, а изучающими язык. Большинство из них содержат различные уровни разметки, в частности, разметку ошибок.

Третья глава посвящена новаторскому подходу к корпусной лингвистике, так называемой Интернет-лингвистике, которая использует Всемирную паутину в качестве инструмента или источника данных для проведения лингвистических исследований. Интернет-лингвистика делится на два главных направления: Интернет как Корпус (от англ. *Web as Corpus*) и Интернет для Корпуса (от англ. *Web for Corpus*).

В соответствии с первым подходом Интернет можно считать лингвистическим корпусом, используемым для проведения лингвистических исследований. Согласно направлению «Интернет для Корпуса», он используется в качестве основного ресурса, посредством которого можно выбрать и скачать требующиеся тексты с любого сайта для того, чтобы создать традиционный лингвистический корпус.

Среди этих двух подходов более спорным считается «Интернет как Корпус». В частности, главными доводами, выдвигаемыми против него, являются следующие: во-первых, в Интернете нет надёжных и достоверных лингвистических данных; кроме того, невозможно установить и проверить действительную личность авторов текстовых данных с абсолютной уверенностью, поскольку Интернет-пользователи могут создать фиктивные профили и аккаунты.

Во-вторых, размер и конечность лингвистического корпуса представляются двумя фундаментальными характеристиками, влияющими не только на сбалансированность самого корпуса, но и на повторяемость и воспроизводимость результатов поиска. Однако невозможно определить размер Интернета, поскольку он постоянно меняется и потенциально бесконечен. Динамичность Сети, следовательно, влияет на сбалансированность корпуса, на повторяемость и воспроизводимость результатов поиска.

Далее в этой главе основное внимание уделяется описанию подхода «Интернет для Корпуса», посредством которого необработанные текстовые данные Интернета превращаются в эффективный исследовательский инструмент для выполнения сложных лингвистических исследований.

Глава заканчивается описанием процесса создания Интернет-корпуса. Во-первых, необходимо выбрать текстовые данные. Такое действие выполняется при помощи одного из двух возможных инструментов анализа: поисковой системы

или веб-сканирования (от англ. *web crawling*). После отбора данных необходимо провести процедуры по очистке, фильтрации и де-дупликации текстов: в первую очередь, надо устранить дубликаты текстов, чтобы убрать повторы, влияющие на окончательные результаты поиска, такие как простой запрос о частотности слов.

Затем необходимо очистить тексты от так называемых «бойлерплейтов», присутствующих на веб-страницах – т. е. всех нетекстовых и неинформативных элементов, таких как навигационные меню, дисклеймеры, рекламы, веб-спам, ссылки, информация об авторских правах, заголовки и т. д.

Существует ещё и другая процедура по очистке данных, касающаяся порнографического материала, содержащегося на веб-страницах.

После этих процедур текстовые данные пригодны и готовы к «токенизации»: другими словами, тексты сегментируются на «токены» (от англ. *tokens*), минимальные лингвистические единицы. Эта предварительная обработка данных необходима для разметки собранных текстов, такой как лемматизация, частеречная разметка (от англ. *POS-tagging*), синтаксическая разметка и т. д.

Большинство этих операций выполняются компьютерами, посредством специального программного обеспечения.

Одной из основных проблем, связанных с созданием веб-корпусов, является авторское право, поскольку даже оцифрованные тексты в Интернете защищены авторским правом, как и тексты в бумажном формате, поэтому их нельзя использовать без разрешения.

Важно отметить, что авторское право в отношении корпусной лингвистики касается не только отдельных текстовых материалов, содержащихся в корпусе, но и авторского права, защищающего весь лингвистический корпус.

Другой проблемой является нетспик, то есть языковая разновидность, используемая в Интернете, имеющая гибридные характеристики разговорного и письменного языка. Например, часто встречаются аббревиатуры, акронимы, дублирование букв и цитаты из комментариев или сообщений других пользователей Интернета.

У русского нетспика есть интересные особенности, такие как тенденция к транслитерации английских заимствований в зависимости от их произношения, а

также к образованию глаголов и прилагательных, происходящих от иностранных заимствований.

В четвёртой и последней главе данного исследования детально анализируются характеристики русского нетспика. Во-первых, исследование было проведено с использованием традиционного корпуса — Национального корпуса русского языка — и двух веб-корпусов — *ruTenTen11* и *ruTenTen17*, доступных на веб-платформе *Sketch Engine*. В тех случаях, когда результаты анализа оказывались сомнительными и противоречивыми, использовался третий веб-корпус – *Timestamped JSI web corpus 2014-2021 Russian*, также доступный в *Sketch Engine*. Двумя другими исследовательскими инструментами являлись бумажный словарь издательства «Дзаникелли» (Kovalev, 2014) и два Интернет-ресурса — Викисловарь и Академик.ру.

Для проведенного исследования были необходимы выделение и выбор лексем, пригодных к анализу, посредством некоторых сайтов, блогов и веб-словарей. Полный список источников можно посмотреть в разделе «Список словарей и сайтов, используемых для идентификации анализируемых лексем», в приложении к данной работе.

Из этих источников были отобраны более 80 лексем. Затем была создана таблица для того, чтобы в виде схемы детально записать данные, полученные в результате анализа лексем. В таблице анализа указывается род и склонение существительных, спряжение глаголов, этимология, тип лингвистической интерференции (калькирование и заимствование), возможное присутствие русских синонимов, производных и составных слов.

Данные анализа демонстрируют различные тенденции: во-первых, в главе были указаны различные транскрипции иностранных заимствований, здесь называемые «фонетическими транслитерациями». Затем были проиллюстрированы наблюдения, возникшие в отношении морфологических категорий существительных, прилагательных и глаголов. Наконец, были исследованы тенденции, касающиеся русских или русифицированных аббревиатур, акронимов, междометий и устойчивых выражений.

Очень частым явлением оказывается существование различных фонетических транслитераций одного и того же слова, например, написанных как

с гласной Э, так и с Е, а также с мягким знаком, твердым знаком или отсутствием знака.

Иногда составные слова пишутся слитно или разделяются дефисом в соответствии с оригинальным английским написанием или, напротив, в отличие от него. Некоторые слова пишутся с удвоением согласного, как в английском языке, но в других случаях существительные с удвоением согласного в английском языке транслитерируются на русский язык с одним согласным.

Некоторые английские заимствования с согласным «*s*» в русском языке пишутся как с глухим сибилянтом «с», так и со звонким сибилянтом «з». Другие заимствования, оканчивающиеся на «*d*» в английском языке, транслитерируются на русский языке как буквой «д», так и согласным «т», в соответствии с русскими правилами произношения, согласно которым все звонкие шумные согласные оглушаются в конце слова.

Что касается существительных, то самым интересным наблюдением является использование английских отглагольных имён существительных, обозначающих деятеля или действия и оканчивающихся на «-*er*», транслитерируемых и склоняемых как русские существительные мужского рода, оканчивающиеся на «-ер». Другими словами, для образования этих существительных не используются русские словообразовательные морфемы.

Другой интересной особенностью является использование транслитерированных английских существительных и прилагательных, к которым добавляется словообразовательный суффикс -ств(о), для создания абстрактных нейтральных существительных, обозначающих «союз или группу людей».

Наконец, интересный аспект касается транслитерированных заимствований или частичных калькирований одного и того же составного слова, имеющих одинаковое значение, но разные формы. Среди них цитируются существительные «селфи-стик» и «селфи-палка» и составные слова «фейк-ньюс», «фейкньюс» и «фейк-новости».

Что касается морфологической категории прилагательного, то часто встречаются прилагательные, происходящие от транслитерированных английских существительных. Кроме того, часто встречаются русские прилагательные, образованные от одного и того же транслитерированного английского

заимствования, имеющие одинаковое значение, но созданные с использованием разных словообразовательных суффиксов, в том числе -ов-/-ев-, -н-, -ск-. Среди них цитируются прилагательных «апгрейдный» и «апгрейдовый», «блогерный» и «блогерский», «лузерный» и «лузерский», «мейнстримный», «мейнстримовский» и «мейнстримовый», происходящие от английских существительных *upgrade*, *blogger*, *loser* и *mainstream*.

Это также происходит с глаголами, но, в отличие от прилагательных, различные словообразовательные суффиксы придают глаголу разнообразные оттенки значения. Например, от английского заимствования *upgrade* произошли пять различных русских глаголов: это глаголы «апгрейдировать», возвратный глагол «апгрейдиться», «апгрейднуть», «апгрейтить» и самый частотный из всех, т. е. «апгрейдить». Другой пример представлен глаголами «свайпать», «свайпить» и «свайпнуть», происходящими от английского слова *swipe*, обозначающего движение пальца по экрану. Это значение однократного и мгновенного действия объясняет, почему «свайпнуть» является самым частотным вариантом из трёх.

Затем были проанализированы некоторые глагольные формы, происходящие от английских транслитерированных заимствований, превращённых в возвратные глаголы с помощью постфикса -ся. Среди них самым экзотическим примером является глагол «сорриться», происходящий от английского слова *sorry*. В английском языке нет глагола *to sorry*, но используются такие формулы, как *to be sorry*, *to feel sorry* или *to say sorry*. Таким образом, этот глагол является доказательством глубокого процесса русификации, поскольку он адаптировался к модели русского возвратного глагола «извиняться – извиниться».

Проанализированные русские акронимы представляют собой транслитерации английских акронимов, таких как ИМХО, транслитерированный акроним английской фразы *In My Humble Opinion*. Напротив, сокращения часто касаются русских слов и среди них цитируются «днюха», от слова «день рождения», «жи́за» или «жиза́», от слова «жизнь».

В заключении этого исследования приводятся основные выводы в отношении примененных инструментов анализа.

Анализ показал, что традиционный бумажный словарь не является адекватным инструментом для поиска неологизмов и иностранных

заимствований; на самом деле, ни одна из проанализированных лексем не была найдена в словаре Дзаникелли (Kovalev, 2014).

Напротив, онлайн-словари Викисловарь и Академик.ру являются подходящими инструментами анализа; только в 5 случаях словари не показали ни одного из результатов поиска. Кроме того, было проведено статистическое сравнение онлайн-словарей Викисловарь с Академик.ру: 55 раз из 75 оба показывали результат поиска; в 16 случаях из 75 только Викисловарь содержал искомое слово, а в 4 случаях из 75 – только Академик.ру. Следовательно, можно утверждать, что из них более современным и адекватным инструментом для исследования русского нетспика является Викисловарь.

Что касается двух подкорпусов НКРЯ, то Основной корпус является более подходящим, чем Устный: в 56 случаях из 80 Основной корпус показывал вхождения искомого слова, в отличие от Устного корпуса, который только в 26 случаях из 80 показал вхождения искомого слова.

Несмотря на их несомненную полезность, эти два подкорпуса НКРЯ нельзя сравнивать с веб-корпусами, используемыми для исследований; *RuTenTen11* и *RuTenTen17*, несомненно, являются самыми эффективными инструментами для исследования неологизмов русского языка и английских заимствований: в 96,25% случаев результаты поиска показали вхождения искомого слова и различных производных и составных лексем.

Что касается лексического анализа, то наблюдаются разные заимствования английского происхождения, показывающие высокую степень интеграции в русском языке, поскольку они порождают новые производные и составные русифицированные слова.

Существительные являются самым частым заимствованием: в этом исследовании было проанализировано 41 существительное. Второй наиболее часто заимствуемой морфологической категорией являются глаголы: в ходе анализа было выявлено 40 различных глагольных форм, происходящих от английских существительных, и только в одном случае – от английского акронима.

За ними по частоте использования следуют прилагательные, образованные от транслитерированных заимствований и созданные с помощью различных

словообразовательных суффиксов; в этом исследовании было проанализировано 29 прилагательных.

Основные проблемы, возникающие в ходе исследования, касаются самих инструментов исследования: в частности, веб-корпуса иногда дают недостоверные и неточные результаты.

В заключение следует добавить, что, по моему мнению, данная работа выполнила поставленные цели, а в некоторых случаях результаты превзошли ожидания.

Я надеюсь, что это исследование может стать основой для более углублённого анализа, такого как, например, изучение управления глаголов, происходящих от английских заимствований, с целью проверить, адаптированы ли они к русскому языку посредством лингвистического калькирования, или же они сохраняют управление по модели английского языка.

Другое будущее исследование может касаться конкордансов и контекста использования проанализированных заимствований, чтобы понять, управляются ли эти существительные одними и теми же глаголами английского языка, или же они адаптированы к русскому языку.

Это лишь некоторые из примеров, которые показывают, насколько обширна и разнообразна эта область исследования: представленный в настоящей работе анализ является не конечной целью, а отправной точкой для будущих исследований.

Introduzione

La rete Internet è attualmente uno strumento essenziale ed indispensabile per rimanere connessi con il resto del mondo. Con il Covid-19 le nostre abitudini sono profondamente mutate e molti di questi cambiamenti sono stati possibili solo grazie ad Internet: la DAD 'Didattica A Distanza' e lo *smart working* hanno permesso a miliardi di persone di studiare o di adempiere alle proprie mansioni lavorative da remoto; durante il periodo di *lockdown* la rete Internet ha permesso di rimanere connessi con il resto del mondo, di fare attività fisica, gli acquisti e la spesa online. Solo mediante Internet è possibile scaricare il Green Pass e i menu di molti bar e ristoranti possono essere consultati esclusivamente con Codice QR ed una connessione Internet.

Questi sono solo alcuni degli esempi più evidenti che mostrano come Internet influenzi quotidianamente le nostre vite, ma esistono molti altri aspetti meno palesi ed ovvi, come i mutamenti linguistici. Questi mutamenti non sono determinati solo dal contatto tra i vari paesi, ma è il mezzo stesso di Internet che induce gli utenti a comunicare in tempi e spazi ridotti. L'uso del Web ha determinato la nascita di una vera e propria lingua a sé, il *netspeak*. Si aprono nuove prospettive di studio, che richiedono strumenti d'indagine all'avanguardia per analizzare quantità di dati enormi e potenzialmente infinite; la linguistica dei corpora si configura come disciplina ideale per realizzare questo scopo.

Il presente studio intende analizzare le caratteristiche del *netspeak* contemporaneo russo, utilizzando i corpora linguistici come principali strumenti di ricerca. I primi tre capitoli forniscono un quadro teorico di questa branca della linguistica: in particolare, il primo capitolo indaga la linguistica dei corpora, mentre il secondo si occupa delle caratteristiche e delle tipologie di corpora linguistici; infine, il terzo capitolo si occupa degli approcci più all'avanguardia della linguistica dei corpora, il *Web as Corpus* e il *Web for Corpus*.

L'ultimo capitolo è il frutto di un anno di ricerca e presenta i risultati delle analisi condotte su oltre cento lessemi. L'indagine si è focalizzata sui prestiti inglesi, sullo slang giovanile e sul *netspeak* russo contemporaneo, la lingua utilizzata nel Web, una vera e propria varietà linguistica, che presenta caratteristiche tipiche sia della lingua parlata, si di quella scritta. Tra le peculiarità rilevate, le più frequenti sono le

abbreviazioni, gli acronimi, la reduplicazione delle lettere, i messaggi ripetuti, le citazioni di altri commenti o di altri utenti. Inoltre, sono state osservate caratteristiche esclusive del solo *netspeak* russo, come la tendenza a traslitterare i prestiti inglesi così come vengono pronunciati, oppure l'esistenza di più trascrizioni per la medesima parola.

Tutte queste tendenze sono state analizzate e dettagliatamente riportate nel quarto ed ultimo capitolo, che ha portato alla luce fenomeni molto interessanti.

Nelle conclusioni sono riportate le principali considerazioni a cui questo studio è giunto, i limiti della ricerca e le prospettive d'indagine future.

1. La linguistica dei corpora

1.1 La linguistica dei corpora: una definizione preliminare

L'informatizzazione e in particolare la *rivoluzione digitale* hanno permesso importanti sviluppi in ogni ambito della vita umana.

Con l'avvento dei computer e l'invenzione di specifici software è stato possibile realizzare la cosiddetta *dematerializzazione*, ossia "il recupero su supporto informatico dei documenti e degli atti cartacei dei quali sia obbligatoria o opportuna la conservazione"[1]. Questo processo ha reso disponibili grandi quantità di testi e banche dati in formato elettronico.

Nell'ambito degli studi linguistici si è assistito ad una vera e propria fioritura della *corpus linguistics* o, secondo la formulazione nostrana, della *linguistica dei corpora*, tanto che alcuni autori parlano di un vero e proprio "renaissance of work in Corpus Linguistics" (Volk, 2002: 255).

In realtà questo approccio alla linguistica, inteso come studio della lingua mediante banche dati testuali rappresentative del linguaggio naturale, esisteva ben prima dell'introduzione dei computer, ma con l'informatizzazione quest'ambito di ricerca è mutato completamente, sia in termini di quantità di dati elaborabili, sia in termini di tempi – ormai brevissimi – di interrogazione di questi dati, tant'è che oggigiorno la definizione formale di corpus linguistico non può prescindere dal formato elettronico e computerizzato dei dati.

Prima di addentrarci nel cuore della questione, però, è necessario un inquadramento teorico; la letteratura a tal proposito offre differenti definizioni di linguistica dei corpora, a seguito ne sono citate tre:

> Corpus linguistics is the investigation of linguistic research questions that have been framed in terms of the conditional distribution of linguistic phenomena in a linguistic corpus. (Stefanowitsch, 2020: 56)

Il manuale *A Glossary of Corpus Linguistics* riporta:

> According to McEnery and Wilson (1996: 1) it is the study of language based on examples of "real life" language use and a methodology rather than an aspect of language requiring explanation or description. (Baker, Hardie, McEnery, 2006: 50)

[1] Sito del Governo Italiano "Comunicazione Pubblica in Rete": http://qualitapa.gov.it/sitoarcheologico/www.urp.it/sito-storico/www.urp.it/Sezione.jsp-idSezione=1874.html.

Infine, concludiamo questa breve rassegna con la definizione di V. P. Zacharov:

> Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использо-вания лингвистических корпусов с использованием компьютерных технологий. (Zacharov, 2013: 139)[2]

In altre parole, la linguistica dei corpora è una metodologia della linguistica che indaga i fenomeni della lingua attraverso analisi empiriche di tipo quantitativo e/o qualitativo, utilizzando come principale strumento i cosiddetti *corpora linguistici*, grandi database di testo in formato elettronico rappresentativi del linguaggio naturale, sia parlato che scritto.

I corpora sono dunque il cuore di questo metodo della ricerca linguistica e saranno i protagonisti del prossimo capitolo, ma prima di indagarne le caratteristiche, le principali tipologie e applicazioni pratiche, occorre illustrare i principali approcci della linguistica dei corpora, gli sviluppi diacronici di questa disciplina e l'attuale stato dell'arte.

1.2 Gli approcci *corpus-based* e *corpus-driven*

Esistono due principali approcci alla linguistica dei corpora, ancora oggi oggetto di una controversa *querelle* tra gli studiosi, le cui differenze sono state chiarite in modo puntuale nel lavoro di E. Tognini-Bonelli (Tognini-Bonelli, 2001): il *corpus-based* e il *corpus-driven*.

L'approccio *corpus-based* si avvale del corpus per testare, convalidare o confutare le intuizioni o aspettative del linguista, quantificare i fenomeni linguistici e ratificare teorie già esistenti, formulate prima dell'introduzione dei corpora come strumento di studio. In questo senso, il corpus si presenta come materiale di supporto, un database di dati qualitativi e quantitativi, mentre la teoria linguistica costituisce il fondamento della disciplina.

Un esempio di approccio corpus-based è il manuale per l'apprendimento della lingua inglese *Student Grammar of Spoken and Written English* (*SGSWE*), frutto del lavoro congiunto di D. Biber, S. Conrad e G. Leech e basato sul *Longman Student Grammar of Spoken and Written English* (*LGSWE*).

[2] "La linguistica dei corpora è una branca della linguistica computazionale che si occupa della elaborazione di principi generali per la creazione e l'utilizzo dei corpora linguistici, mediante l'utilizzo della tecnologia informatica". La traduzione è mia.

Il *LGSWE* fu pubblicato per la prima volta nel 1999, come risultato di sette anni di ricerca; secondo quanto vi è riportato, "LGSWE made important innovations in the method of grammatical study. It was based on a large, balanced corpus of spoken and written texts. These texts were electronically stored and analyzed with the aid of computers" (Biber, Conrad, Leech, 2002: 2). In particolare, quello a cui si fa riferimento è il *Longman Spoken and Written English Corpus* (*LSWE Corpus*), un corpus di circa 40 milioni di parole di esempi d'uso reale, rappresentativo dei quattro principali registri linguistici (conversazione, narrativa, articoli di giornale e prosa accademica).

Ecco un esempio di analisi condotta mediante l'utilizzo del *LSWE Corpus*:

The most common verbs controlling a complement *wh*-clause by semantic domain and register; occurrences per million words

over 1000 over 500 over 200 over 100
over 40 over 20

semantic domain	CONV	FICT	NEWS	ACAD
cognition				
know	■■■	■■■	■■	■
wonder	■	■		
think (about)	■	■		
remember	■	▮		
understand	▮	■		▮
guess	▮			
realize				▮
find (out)		▮		
speech act				
tell (NP)	■	■	▮	
ask (NP)	■	■	▮	
say	▮	▮		
explain			▮	▮
other communication				
show (NP)	▮	▮		▮

Figura 1. Esempio di analisi corpus-based presente in *SGSWE* (Biber, Conrad, Leech 2002: 325).

Nel secondo approccio, quello *corpus-driven*, il corpus linguistico funge da fondamento empirico dal quale è possibile estrarre dati e rilevare fenomeni linguistici, senza preconcetti o aspettative. Ciò si traduce nell'utilizzo di corpora grezzi, privi di qualsivoglia annotazione, affinché possano essere analizzati testi completamente incontaminati.

L'approccio corpus-driven, pertanto, tende a limitare le ipotesi aprioristiche circa un determinato fenomeno linguistico e eventuali teorie linguistiche o conclusioni sono formulate esclusivamente sulla base dei dati ricavati dal corpus indagato.

Un esempio d'indagine corpus-driven è lo studio *Un approccio corpus-driven al linguaggio dell'immigrazione,* condotto dalla linguista italiana E. Manca. In questa ricerca sono stati creati due corpora comparabili, basati su testi provenienti dalla pagina web *Portale dell'Immigrazione*, dedicato alle procedure di rilascio e rinnovo dei permessi e delle carte di soggiorno, dalla sezione *Immigrazione e Asilo* del sito del Ministero dell'Interno italiano e dal sito ufficiale del governo britannico *UK Visas and Immigration*.

Citando le parole della stessa autrice:

> L'intento finale è quello di applicare la metodologia della Corpus Linguistics ai testi giuridico-giudiziari scritti relativi all'immigrazione e utilizzati per facilitare la mediazione linguistico-culturale, in modo tale da individuare la corretta fraseologia relativa a tale linguaggio, sia in inglese che in italiano, e di elaborare una serie di equivalenti traduttivi tra le due lingue. (Manca, 2015: 287)

Lo studio è interessante perché permette di osservare come l'approccio corpus-driven – e più in generale l'impiego dei corpora – sia stato utilizzato per una ricerca riguardante la traduzione e mediazione linguistica nell'ambito di una tematica quanto mai attuale.

Nella prossima sezione vedremo lo sviluppo diacronico della linguistica dei corpora, con particolare riferimento al mondo anglofono.

1.3 Sviluppo diacronico della linguistica dei corpora nel mondo anglofono

La linguistica dei corpora è spesso considerata una metodologia recente della ricerca linguistica, sviluppatasi e divenuta popolare soprattutto negli anni '80 del secolo scorso. Tuttavia, le sue radici sono ben più profonde e si possono individuare in studi condotti prima dell'introduzione di corpora propriamente detti, in formato elettronico, e più precisamente risalenti ai primi tentativi di creazione di indici analitici, i cosiddetti *concordances*[3]. A tal proposito occorre menzionare il lavoro di Thomas Gybson, che nel 1535 pubblicò *The Concordance of the New Testament*, il primo indice analitico del

[3] *A Glossary of Corpus Linguistics* fornisce la seguente definizione di *concordance*: "Also referred to as key word in context (KWIC), a concordance is a list of all of the occurrences of a particular search term in a corpus, presented within the context in which they occur" (Baker, Hardie, McEnery, 2006: 42-44).

Nuovo Testamento; in seguito, nel 1550, John Merbecke (noto anche come Marbeck o Merbeck) pubblicò il primo indice dell'intera Bibbia in inglese, ma si trattava di un lavoro poco sistematico e preciso. Occorrerà attendere quasi duecento anni per la pubblicazione di un indice testuale che rimarrà incontrastato per circa un secolo, il *Curden's Concordance*[4]. Il titolo originale era *A Concordance to the Holy Scriptures* e fu realizzato da Alexander Cruden (1699-1770); fu pubblicato per la prima volta nel 1737, mentre le successive edizioni risalgono al 1761 e 1769.

Ulteriori indici analitici della Bibbia furono quello di Robert Young, pubblicato nel 1879 con il titolo *Young's Analytical Concordance to the Bible*[5], che include più di 300 mila riferimenti biblici in greco ed in ebraico, un indice del lessico del Nuovo Testamento ed anche una lista di nomi propri presenti nelle Sacre Scritture, e infine l'indice analitico della Bibbia di Re Giacomo, redatto da James Strong e pubblicato nel 1890 con il titolo *The Exhaustive Concordance of the Bible*.

Dal 1876 al 1926 circa furono condotti degli studi sull'acquisizione del linguaggio da parte dei bambini attraverso database testuali, basati su diari redatti dai genitori che riportavano accuratamente le parole e frasi del proprio bambino.

Nel 1897 Käding utilizzò un corpus di tedesco di 11 milioni di parole per analizzare la distribuzione di frequenza delle lettere o di sequenze di lettere in tedesco.

Sempre in ambito inglese, si ricordano gli studi pionieristici di Charles C. Fries (1887-1967), che nel 1922 raccolse ed analizzò 50 commedie per ogni decennio di letteratura britannica del periodo compreso tra il 1560 e il 1915 e 18 commedie di letteratura americana e britannica dal 1902 al 1918, raccogliendo in tutto 20.000 esempi d'uso di *shall* e *will*. Il fine era analizzare l'utilizzo effettivo di questi verbi e valutare l'accuratezza delle regole presenti nei manuali scolastici utilizzati per l'insegnamento della lingua (Fries, 2010:114); i risultati di questa ricerca, che potremmo definire "corpus-based", furono pubblicati nel 1925 in un articolo intitolato *The periphrastic future with shall and will in Modern English*[6].

[4] Per leggere una ristampa originale del 1858: https://www.unz.com/print/CrudenAlexander-1858.
[5] Disponibile al link: https://archive.org/details/analyticalconcor00younuoft/page/n5/mode/2up?view=theater.
[6] Disponibile al link: https://www.jstor.org/stable/457534?seq=43#metadata_info_tab_contents.

Questo studio è sorprendente se si considera il periodo in cui è stato condotto e la totale assenza di mezzi informatici e, sebbene non si possa certo parlare di corpus propriamente detto, ne è indubbiamente un precursore naturale.

Sempre a C. C. Fries si devono la raccolta e l'analisi sistematica di circa 50 ore di conversazioni telefoniche di 300 interlocutori, per un totale di oltre 250.000 parole, al fine di scoprire le caratteristiche dell'inglese parlato (Fries, 2010: 113). I risultati della ricerca furono pubblicati nel volume *The structure of English* (1952) e malgrado non si trattasse di un campione rappresentativo della popolazione anglofona, fu certo un tentativo di condurre un'analisi empirica mediante un "corpus" di lingua inglese parlata.

Generalmente gli anni '50 sono descritti come un momento di battuta d'arresto per la linguistica dei corpora a causa dei duri attacchi rivolti all'uso dei corpora nella ricerca linguistica dal linguista Noam Chomsky[7], già all'epoca molto influente.

In questa sede mi limiterò ad un breve elenco delle obiezioni mosse dallo studioso statunitense. Per un quadro approfondito della questione si rimanda al volume di McEnery e Wilson (McEnery, Wilson, 2001).

Secondo Chomsky:

- La linguistica non deve limitarsi a quantificare e descrivere le performance linguistiche di un parlante, ma indagare le sue competenze[8].
- Non è possibile basare l'analisi linguistica su dati finiti e quantificabili, come nel caso dei corpora linguistici, in quanto le lingue naturali hanno un carattere non-finito, sulla base della loro natura ricorsiva.
- Il linguista non deve rinunciare totalmente all'introspezione, vale a dire raccogliere dati sulla base delle proprie intuizioni linguistiche (solitamente relative alla propria lingua madre), altrimenti diventa difficile rilevare strutture agrammaticali o ambigue. In altre parole, secondo Chomsky, se un corpus non contiene una frase *x*, l'unico modo per stabilire se tale frase *x* sia o meno grammaticale è ricorrere al giudizio introspettivo del linguista stesso.

[7] Il sito web di Noam Chomsky: https://chomsky.info/.

[8] Chesi C. (2012: 241) dà la seguente definizione: "Da un lato, per competenza si intende «la conoscenza astratta del proprio linguaggio», ovvero quell'insieme di regole o principi che permettono a qualsiasi parlante nativo di determinare se un'espressione è ben-formata oppure no rispetto alla lingua di riferimento […]. Dall'altro lato, per performance linguistica si intende invece l'effettivo uso di questa conoscenza".

In seguito a queste critiche l'atteggiamento dei linguisti nei confronti dei corpora mutò, ma sarebbe sbagliato parlare di un rifiuto: se è vero che da una lato ci fu un allontanamento rispetto a questo metodo di indagine, è vero anche che nel periodo che va dagli anni '50 sino agli anni '80 continuarono ad essere condotti degli studi corpus-based, come nel campo dell'acquisizione del linguaggio e della fonetica, ambiti in cui è difficile condurre analisi introspettive e che necessitano di corpora quali fonti di dati oggettivi e attendibili. In generale, furono creati e pubblicati corpora che sono divenuti modelli e pietre miliari di questo approccio alla ricerca.

Un vero e proprio anno di svolta nella storia della linguistica dei corpora fu il 1964, quando venne pubblicata la prima versione del *Brown Corpus of American Written English*[9], un lavoro rivoluzionario di W. Nelson Francis ed H. Kučera presso la Brown University di Providence nel Rhode Island. Come osserva M. Barbera, il *Brown* "è certo il primo corpus a soddisfare in tutto e per tutto la moderna definizione formale" (Barbera, 2013: 11). Esso consisteva in 500 campioni di testo di circa 2000 parole ciascuno, appartenenti a 15 generi differenti – dalla divulgazione scientifica alla narrativa – per un totale di 1 milione di parole.

Sulla scia del *Brown Corpus* negli anni '70 fu creato il *Lancaster-Oslo-Bergen Corpus* (spesso abbreviato *LOB Corpus*)[10], frutto della collaborazione tra l'Università di Lancaster, l'Università di Oslo e il Norwegian Computing Center for the Humanities di Bergen. Essendo stato pensato come controparte britannica del *Brown Corpus*, la sua composizione era simile a quest'ultimo, sia a livello di dimensioni, sia in termini di generi testuali in esso contenuti; entrambi i corpora si basavano su 500 campioni testuali di circa 2000 parole, ma nel caso del *LOB Corpus* gli autori dei testi erano britannici e non statunitensi.

Nel 1987 fu pubblicato per la prima volta il *COBUILD English Language Dictionary*, acronimo di *Collins Birmingham University International Language Database.* Si tratta del primo dizionario avanzato pensato per gli apprendenti di lingua inglese e basato su reali esempi d'uso, ovvero sul *Collins COBUILD corpus*[11]. Il

[9] Il manuale del *Brown Corpus*: http://icame.uib.no/brown/bcm.html#tc.
[10] Il manuale del *LOB Corpus*: http://korpus.uib.no/icame/manuals/LOB/INDEX.HTM.
[11] Sito del *Collins COBUILD Corpus*: https://collins.co.uk/pages/elt-cobuild-reference-the-history-of-cobuild.

progetto per la creazione del corpus fu diretto dal noto linguista J. Sinclair presso l'Università di Birmingham e finanziato dagli editori del dizionario Collins.

Dagli anni ’90 il *COBUILD Corpus* è stato utilizzato come base per la creazione di libri di grammatica, come la *Cobuild English Grammar* e le *Cobuild English Guides*.

Per concludere, i corpora citati sono solo alcuni dei progetti realizzati negli anni ’60, ’70 ed ’80 del secolo scorso in ambiente anglofono nonostante la dura critica chomskyana e, a mio parere, queste evidenze confutano l’erronea ma diffusa convinzione secondo cui “negli anni Sessanta e Settanta si è avuta una battuta di arresto nello sviluppo della disciplina” (Chiari, 2007: 40).

Dopo questo breve excursus storiografico a proposito del mondo anglofono, nel prossimo paragrafo verrà illustrato il percorso intrapreso dalla controparte italiana.

1.4 Sviluppo diacronico della linguistica dei corpora nel panorama italiano

Se descrivere lo sviluppo diacronico della linguistica dei corpora anglofona risulta agevole, in quanto l’argomento viene trattato non solo dagli stessi autori anglofoni, ma anche da quelli italofoni, non si può dire altrettanto per la linguistica dei corpora italiana, poiché manca uno studio diacronico che si occupi in maniera puntuale del contesto italiano, dalle origini ad oggi.

Nel ricostruire lo sviluppo storico della disciplina appare evidente che ci sono dei vuoti storici che necessitano di essere colmati, con particolare riferimento al periodo che va dagli anni ’50 sino agli anni ’80 del secolo scorso, periodo di cui si hanno poche informazioni esaustive.

Sorgono spontanee alcune domande circa il rapporto in quegli anni tra la linguistica dei corpora anglofona e italiana: come venivano recepite in Italia le innovazioni d’oltreoceano? In quale misura la critica chomskyana influenzò lo scenario italiano? Qual era l’atteggiamento dei linguisti italiani nei confronti della linguistica dei corpora in quegli anni? Quali erano le risorse, sia materiali che economiche, di cui si disponeva per la ricerca? A quali ambiti di interesse venivano applicati gli studi della linguistica dei corpora in Italia?

Queste sono solo alcune delle domande sorte durante lo spoglio bibliografico. La lettura e la ricostruzione storica del panorama italiano potrebbero essere un punto di

partenza per una ricerca dedicata proprio a ricostruire questo periodo della storia italiana.

F. Sabatini, Presidente onorario dell'Accademia della Crusca, osserva come la tradizione italiana sia caratterizzata dalla creazione e l'utilizzo di corpora (intesi come ampie raccolte testuali) sin dai tempi delle dispute linguistiche tra volgare e latino come lingua letteraria, all'epoca di Dante. Nel suo saggio *La storia dell'italiano nella prospettiva della corpus linguistics*, Sabatini evidenzia come la lingua italiana contemporanea sia diretta discendente della lingua del XIV secolo – difatti il 61% delle parole dell'italiano contemporaneo è proprio di origine Trecentesca – affermatasi come standard linguistico sulla base di grandi raccolte di opere di autori dell'epoca, dei "corpora tardo-medioevali", per così dire (Sabatini, 2006: 32).

In realtà, questa tradizione italiana affonda le sue radici in un'epoca precedente a quella di Dante, in particolare nel periodo in cui visse Sant'Antonio da Padova (1195-1231): è a quest'ultimo che viene attribuito il *Concordantiae Morales*, il primo indice analitico della *Vulgata*, la traduzione latina della Bibbia di San Girolamo, risalente al V secolo.

Ritengo piuttosto significativo che Sant'Antonio da Padova non venga menzionato nella letteratura italiana, mentre invece è presente in diversi autori anglofoni – sia degli anni '70, che a noi contemporanei[12] – e russi (Solnyškina, Gatijatullina, 2020: 133). Inoltre, trovo altrettanto significativo che a fronte della ricerca in lingua italiana su Google "Sant'Antonio da Padova Concordantiae Morales" non emerga alcun risultato relativo all'opera, mentre se la ricerca viene effettuata in lingua inglese "Saint Anthony of Padua Concordantiae Morales" tutti i risultati riguardano l'opera, ma in siti stranieri. Nondimeno, esiste il libro *The Moral Concordances Of Saint Anthony Of Padua* che viene venduto anche in Italia, ma esclusivamente in lingua inglese.

Ritornando al nostro percorso storico della disciplina, occorre fare menzione di un personaggio illustre e carismatico nato alla fine del XIII secolo: Niccolò de' Rossi, ricordato per la sua incessante attività di autore e scrittore. Un illustre esempio della sua instancabile solerzia è rappresentato dal corpus manoscritto *Barberiniano latino 3953*[13],

[12] Per maggiori approfondimenti si veda: Pearsall, 1971: 88. Si veda, inoltre, un contributo più recente: O'Keeffe, McCarthy, 2010: 3.

[13] Per ammirare le immagini del manoscritto originale e i disegni miniati in esso contenuti: https://spotlight.vatlib.it/dante/feature/il-canzoniere-di-nicolo-de-rossi.

conservato nella Biblioteca apostolica Vaticana e risalente alla fine degli anni ’20 del XIV secolo, che contiene, oltre a 75 componimenti dello stesso de’ Rossi, rime del Duecento e inizio Trecento di autori come Dante, Guido Cavalcanti, Guido Novello da Polenta, Cino da Pistoia, Cecco Angiolieri, Musa da Siena e tanti altri.

Un altro corpus raccolto e trascritto da de’ Rossi fu il *Colombino 7.1.32* della Biblioteca capitolare di Siviglia, che contiene 4 canzoni e 434 sonetti composti da de’ Rossi in persona.

Dal de’ Rossi occorre fare un salto temporale di oltre due secoli, per approdare al 1612, anno di pubblicazione della prima edizione del *Vocabolario degli Accademici della Crusca*, un dizionario che raccoglieva 25.056 lemmi, 52.862 citazioni di 208 autori per un totale di 1.152.999 parole e la cui intenzione dichiarata, riporta Sabatini, era “pervenire alla formazione di un vero *corpus* universale, capace di rappresentare tutta la lingua” (Sabatini, 2006: 34).

Ma se quelli visti sino ad ora sono esempi di corpora *ante litteram* o, per utilizzare un’espressione russa, *doelektronnye korpusy* (‘corpora pre-elettronici’), bisognerà aspettare l’avvento del Padre gesuita Roberto Busa SJ (1913–2011), e più precisamente l’anno 1949, per una radicale svolta della disciplina: di fatto egli fu il primo ad addentrarsi nel sentiero dell’informatica linguistica.

Padre Busa si laureò in Filosofia nel 1946, presso la Pontificia Università Gregoriana, con una tesi di laurea su San Tommaso d’Aquino intitolata *La terminologia tomistica dell’interiorità*; ben presto comprese che per analizzare il concetto di “interiorità” in San Tommaso avrebbe dovuto studiare l’utilizzo e il significato di locuzioni come «in» e «intra», e fu così che iniziò ad annotare manualmente le loro occorrenze, giungendo ben presto a 10 mila schede. Questo approccio iniziò a mostrare i propri limiti, in primis di spazio: divenne evidente che fosse necessario implementarlo mediante computerizzazione.

Nel 1949 Padre Busa si recò in America e qui ebbe un incontro che cambiò non solo il corso della sua vita, ma anche, come si è detto, quello della disciplina: l’incontro con Thomas J. Watson Senior, il fondatore dell’IBM (*International Business Machines Corporation*)[14].

[14] Nel sito dell’IMB è possibile approfondire la storia dell’azienda dal 1880: https://www.ibm.com/ibm/history/history/history_intro.html.

L'idea del gesuita era di utilizzare le macchine IMB per analizzare automaticamente i testi latini di San Tommaso d'Aquino. Nonostante un iniziale rifiuto, Padre Busa ottenne un primo finanziamento per indicizzare l'opera omnia dell'aquinate, dapprima attraverso l'utilizzo di schede perforate, successivamente tramite l'elaborazione elettronica, dando vita all'*Index Thomisticus*, un proto-indice computerizzato in 56 volumi, frutto di trent'anni di lavoro, per un totale di quasi 11 milioni di parole, che fu "la prima applicazione del computer a studi linguistici (secondo quella disciplina che Busa chiamava «informatica linguistica» e che ora si chiama comunemente «linguistica computazionale», all'interno delle applicazioni del *Computer in Humanities*)" (Piccolo, Di Maio, 2014: 68).

Nel 1990 l'intero lavoro fu presentato in un CD-ROM che constava di tutte le opere di San Tommaso e quelle di 61 autori a lui coevi; oltre al compact disk era presente il software necessario per consultare i testi e fare operazioni di ricerca, nonché l'ipertesto contenente informazioni di tipo morfologico e lessicale di ogni parola dell'*Index*. Infine, nel 2005 l'*Index Thomisticus* fece il suo debutto nel web e ancora oggi è oggetto di perfezionamento ed implementazione[15].

Come accennato all'inizio di questo paragrafo, si hanno poche informazioni riguardanti la linguistica dei corpora a partire dalla seconda metà del XX secolo; in questa sede si è tentato di ricostruire le principali tappe della disciplina nel suddetto arco temporale.

Il primo corpus di italiano parlato risale al 1965: si tratta del *Corpus Stammerjohann*, contenente registrazioni di parlato spontaneo italiano raccolte a Firenze tra il 29 gennaio e il 24 febbraio 1965 da Harro Stammerjohann, per un totale di circa 47 ore. Successivamente H. Stammerjohann donò il corpus a E. Cresti, affinché venisse archiviato all'interno del *Corpus LABLITA;* infine, nel 2001 le registrazioni originali del corpus, incise su due bobine 4 piste, furono restaurate, digitalizzate ed archiviate elettronicamente, mediante trascrizione in formato CHAT[16], con allineamento testo-suono-parametri acustici ed annotazione (Signorini, Tucci 2004: 119).

Negli anni '70 videro la luce due importanti contributi: il *Corpus LIF* (*Lessico di frequenza della lingua italiana contemporanea*) ed il già citato *Corpus LABLITA*.

[15] Il sito ufficiale del *Corpus Thomisticus*: http://www.corpusthomisticum.org/.
[16] Il CHAT Transcription Format è lo stesso utilizzato dal progetto CHILDES. Il manuale aggiornato ad Agosto 2021 del CHAT è disponibile al link https://talkbank.org/manuals/CHAT.pdf.

Il *LIF*, frutto del lavoro del Centro Nazionale Universitario di Calcolo elettronico di Pisa, fu pubblicato per la prima volta nel 1971; si tratta di un corpus di testi provenienti dal mondo teatrale e cinematografico, oltre che da romanzi, periodici e sussidiari, scritti tra il 1945 e il 1968 per un totale di mezzo milione di parole, da cui sono stati individuati i 5000 lemmi più frequenti.

Nel 1973 fu pubblicato il *LABLITA* (*LABoratorio Linguistico del dipartimento di ITAlianistica* dell'Università di Firenze), un corpus bilanciato di parlato spontaneo, la cui creazione si deve in gran parte al contributo dei linguisti E. Cresti e M. Moneglia; esso presenta un allineamento testo-suono e una trascrizione in formato CHAT, ed è annotato sia per parti del discorso (*POS-tagging*), sia prosodicamente. Il *LABLITA* ha realizzato numerosi progetti, tra questi occorre citare *C-ORAL-ROM* (*Integrated Reference Corpora for Spoken Romance Languages*), un insieme di corpora di lingua parlata delle principali lingue romanze, come francese, italiano, portoghese e spagnolo[17]. La sezione italiana del *C-ORAL-ROM* è formata da quattro sotto-corpora: un corpus costituito da 80 ore di parlato spontaneo, per un totale di 600.000 parole trascritte; un *learner corpus* della lingua italiana di 650.000 parole circa, corrispondente a 95 ore di parlato; un corpus della lingua cinematografica, contenente 21 ore di trascrizione di film italiani degli anni 1948-1994, per un totale di 115.000 parole, ed infine un corpus di linguaggio dei media (radio e TV) di 92.000 parole.

Un altro progetto del *LABLITA* è il *CorDIC* (*Corpora Didattici Italiani di Confronto*) che si propone come strumento didattico per il confronto tra le varietà scritta e orale dell'italiano[18]; il *CorDIC* è articolato nel *Corpus CorDIC-scritto* e nel *Corpus CorDIC-parlato*, strettamente comparabili e costituiti da circa 200 testi per un totale di circa 500.000 parole ciascuno.

Infine, un ultimo progetto del *LABLITA* è *RIDIRE* (*RIsorsa Dinamica Italiana di REte*), un corpus lessicale dinamico "di 1.514.631.794 token, costruito a partire da 2.010 siti con 3.767.668 pagine web" (Barbera, 2015: 125), per un totale di oltre 1,3 miliardi di parole. Il corpus è suddiviso nei cosiddetti *domini semantici*, come lo sport, il cinema, la religione, la letteratura, il design, il cibo, la moda e in *domini funzionali*, che comprendono l'informazione, l'economia e gli affari, l'amministrazione e la

[17] Per maggiori informazioni sul progetto C-ORAL-ROM: http://www.elda.org/en/proj/coralrom.html.
[18] Per maggiori informazioni sul progetto CorDIC LABLITA: http://corporadidattici.lablita.it/.

legislazione. L'aspetto più interessante è che l'apprendente può consultare il corpus nel suo insieme, oppure analizzare il linguaggio specifico di un singolo dominio.

Proseguendo nel nostro excursus storico della disciplina, gli anni '80 del secolo scorso hanno visto la nascita di ulteriori progetti: il cosiddetto *Progetto di Pavia*, nato nel 1985, un corpus di italiano L2 che raccoglie 120 ore di parlato di 20 apprendenti di italiano (madrelingue albanese, arabo, chichewa, cinese, inglese, morè, tedesco, tigrino) (Chini, 2016: 4), e il *Corpus La Repubblica*, nato anch'esso nel 1985, un corpus di testi di italiano giornalistico tratti dal quotidiano "La Repubblica" degli anni 1985-2000, per un totale di oltre 3,5 milioni di *token*[19].

Nel 1988 il lavoro congiunto di A. Ciliberti, D. Zorzi e G. Aston ha dato vita al *PIXI Corpus* (*Pragmatics of Italian/English Cross-Cultural Interaction*), che consta della trascrizione dettagliata di 450 dialoghi spontanei tra venditore ed acquirente, avvenuti in alcune librerie del sud-est dell'Inghilterra e del nord Italia. Il progetto mirava ad analizzare gli aspetti pragmatici e la struttura del discorso nell'ambito dell'interazione interculturale inglese-italiano e a identificarne le somiglianze e differenze.

Infine, nel 1989 fu pubblicato il *VELI* (*Vocabolario Elettronico della Lingua Italiana*) di T. De Mauro, un vocabolario elettronico di frequenza di circa 10.000 parole basato sull'omonimo *Corpus VELI*, una raccolta di testi tratti dalle testate giornalistiche *Europeo*, *Domenica del Corriere*, *Il Mondo* e dall'agenzia ANSA.

Negli anni '90 assistiamo ad una piena maturazione della disciplina. I progetti ed i corpora pubblicati in questo decennio sono numerosi e di vario genere, come il *LIP* (*Lessico di frequenza dell'italiano parlato*) del 1993, il *CoLFIS* (*Corpus e Lessico di Frequenza dell'Italiano Scritto*)[20] e il *LIR* (*Lessico di frequenza dell'italiano radiofonico*), entrambi del 1995, e infine il *CORIS/CODIS* (*Corpus di Italiano Scritto*)[21] e il *CiT* (*Corpus di Italiano Televisivo*)[22] del 1998, solo per citarne alcuni[23].

Dagli anni 2000 ad oggi la disciplina ha compiuto passi da gigante: grazie al rapido sviluppo tecnologico ed alla nascita di specifici software attualmente è possibile

[19] Per maggiori informazioni sul *Corpus La Repubblica*: https://corpora.dipintra.it/.
[20] Per maggiori informazioni sul *CoLFIS*: http://www.ge.ilc.cnr.it/strumenti.php.
[21] Per maggiori informazioni sul *CORIS/CODIS*: http://corpora.dslo.unibo.it/coris_ita.html.
[22] Per maggiori informazioni sul *CiT*: http://www.culturitalia.info/ARCHIVIO/s_spina/cit/demo.htm.
[23] Per una rassegna dettagliata di banche dati, corpora e archivi testuali italiani si veda: http://old.accademiadellacrusca.org/it/link-utili/banche-dati-dellitaliano-scritto-parlato.html, oltre all'indice (aggiornato al 25 Marzo 2004) di M. Barbera: http://www.bmanuel.org/clr/clr3_fi.html#Italian.

raccogliere, elaborare, organizzare ed analizzare enormi quantità di dati, in tempi brevissimi. La linguistica dei corpora è sempre più presente non solo a livello nazionale, grazie a workshop, convegni, pubblicazioni e studi riguardanti questo metodo d'indagine, ma anche all'estero, grazie ai contributi degli studiosi italiani pubblicati in prestigiose riviste specializzate, come *The International Journal of Corpus Linguistics* (*IJCL*)[24], e alla partecipazione a importanti conferenze, come *EURALEX* (*European Association for Lexicography*)[25] e *International Corpus Linguistics Conference*.

Probabilmente saranno molti i futuri contributi e gli sviluppi tecnologici di questo approccio alla ricerca; è auspicabile che in futuro possa nascere un'Associazione Italiana di Linguistica dei Corpora, al pari dell'*American Association for Corpus Linguistics* (*AACL*), dell'*Asia Pacific Corpus Linguistics Association* (*APCLA*)[26], o ancora della *Japan Association for English Corpus Studies* (*JAECS*)[27], della *Korean Association for Corpus Linguistics* (*KACL*)[28] e infine dell'*Asociación Española de Lingüística de Corpus* (*AELINCO*)[29], poiché ad oggi in Italia esiste la *Società di Linguistica Italiana* (*SLI*)[30] ed anche la *Associazione Italiana di Linguistica Computazionale* (*AILC*)[31], ma non esiste una associazione specificatamente dedicata alla *corpus linguistics*, sintomo del fatto che in Italia questo approccio alla disciplina non ha ancora acquisito una completa autonomia ed affermazione.

Nel prossimo paragrafo volgeremo lo sguardo ad Est, in particolare alla linguistica dei corpora in Russia, per avere una panoramica della disciplina nel paese della lingua oggetto di studio in questa ricerca.

1.5 La linguistica dei corpora in Russia

La *korpusnaja lingvistika* ('linguistica dei corpora') russa è una metodologia d'indagine piuttosto recente, diffusasi soprattutto negli anni 2000; infatti, come afferma la studiosa russa K. Irgizova, "[…] la scienza dei corpora in Occidente ha iniziato a svilupparsi 40

[24] Per maggiori informazioni sull'IJCL: https://benjamins.com/catalog/ijcl.
[25] Per maggiori informazioni sull'EURALEX: https://euralex.org/.
[26] Il sito ufficiale dell'*APCLA*: https://apcla.net/.
[27] Il sito ufficiale della JAECS: https://www.jaecs2020.org/.
[28] Il sito ufficiale della KACL: http://kacl.or.kr/.
[29] Il sito ufficiale dell'*AELINCO*: http://www.aelinco.es/es.
[30] Il sito ufficiale della SLI: https://www.societadilinguisticaitaliana.net/.
[31] Il sito ufficiale dell'AILC: https://www.ai-lc.it/.

anni prima rispetto che in Russia"[32] e oggi in Russia "[…] non ci sono molti studi autonomi dedicati allo studio dell'applicazione delle ultime tecnologie alla linguistica"[33] (Irgizova, 2019: 2).

Uno dei primi lavori basati sull'utilizzo di un corpus di testi fu il *Častotnyj slovar' russkogo jazyka*[34] (Dizionario di frequenza della lingua russa) ad opera della linguista L. N. Zasorina (1929–2016), elaborato tra gli anni '60 e '70 del secolo scorso e pubblicato nel 1977. Il dizionario si basava su un corpus di testi appartenenti a quattro differenti generi (testi su temi socio-politici, di narrativa, scientifici, divulgativi e teatrali), per un totale di 1 milione di parole.

Successivamente, nel 1985, nell'allora Unione Sovietica nacque il *Mašinnyj Fond Russkogo Jazyka*[35] sotto la direzione dell'accademico A. P. Eršov e sostenuto dal Presidium dell'Accademia delle Scienze e dal Comitato statale del Consiglio dei Ministri dell'URSS per la scienza e la tecnologia (GKNT SSSR).

Il fondo conteneva testi teatrali, di prosa e poesia russa dei secoli XIX-XX, un corpus di giornali russi degli anni '90 del secolo scorso, alcuni dizionari russi ed infine testi di storia e folklore russo.

La ricerca fu molto attiva e produttiva tra il 1985 e il 1992, tuttavia i finanziamenti diminuirono progressivamente nel corso degli anni '90, periodo in cui l'interesse virò verso altri tipi di ricerca, come quella riguardante la realizzazione del Corpus Nazionale della lingua russa, che tratteremo a breve.

Il primo vero e proprio corpus di lingua russa fu l'*Uppsal'skij Korpus Russkogo Jazyka* (Corpus di Uppsala della lingua russa), creato in Svezia tra gli anni '80 e '90 presso il Dipartimento di studi di slavistica dell'Università di Uppsala su iniziativa del Professor L. Lönngren; esso apparteneva ai cosiddetti corpora russi di Tubinga, del progetto "Linguistische Datenstrukturen. Theoretische und empirische Grundlagen der Grammatikforschung" dell'Università di Tubinga (Zacharov, 2013: 2).

[32] "[…] наука о корпусах на Западе начала развиваться на 40 лет раньше, чем в России". La traduzione in italiano è mia.
[33] "[…] существуют не так много отдельных исследований, посвященных изучению применения новейших технологий в языкознании". La traduzione in italiano è mia.
[34] Per maggiori informazioni sul Dizionario di frequenza della lingua russa (1977) consultare il link: http://project.phil.spbu.ru/lib/data/slovari/zasorina/zasorina.html.
[35] Sito ufficiale del *Mašinnyj Fond Russkogo Jazyka*: http://cfrl.ruslang.ru/.

Il *Corpus di Uppsala*[36] è composto da circa 600 testi russi specialistici e letterari, per un totale di 1 milione di parole. I testi specialistici appartengono al periodo compreso tra il 1985 e il 1989, mentre i testi letterari sono degli anni 1960-1988.

Al fine di rendere il corpus più rappresentativo e vario sono stati scelti dei testi riguardanti differenti aree tematiche, tra le quali ricordiamo: economia, società, educazione, storia, cultura, diritto, ingegneria, informatica, medicina, ambiente, agricoltura, ricerca spaziale, biologia, sport e molte altre, per un totale di 25 aree tematiche.

È bene sottolineare che nonostante sia un corpus di lingua russa, non si tratta di un progetto russo, bensì svedese, dell'Università di Uppsala; per il primo vero e proprio corpus creato in territorio nazionale bisognerà attendere il periodo tra il 2000 e il 2002, quando il laboratorio di lessicologia e lessicografia generale ed informatica[37] della Facoltà di filologia dell'Università statale di Mosca (MGU) creò il primo *Komp'juternyj korpus tekstov russkich gazet konca XX veka* (Corpus informatico di testi della stampa russa della fine del XX secolo)[38].

Secondo quanto scritto nel sito ufficiale, questo corpus, che comprende oltre 11 milioni di parole, contiene articoli di giornale pubblicati tra il 1994 e il 1997 e provenienti da 13 "giornali e riviste (come *MN* e *Novaja Gazeta*), di "sinistra" (come *Zavtra, Pravda* e *Pravda-5*) e di "destra", giornali centrali e locali, generici e specializzati (come *Literaturnaja gazeta*)"[39]. Questa varietà permette di offrire un corpus relativamente rappresentativo, bilanciato e soprattutto oggettivo dei testi giornalistici dell'epoca.

Un altro corpus di lingua russa, nato al di fuori dei confini russi e, più precisamente, a Helsinki in Finlandia, è il *Chel'sinskij annotirovannyj korpus* (Corpus Annotato di Helsinki)[40]. Il progetto, durato circa 10 anni dal 2001 al 2012, è nato grazie all'iniziativa del professor A. Mustajoki e del coordinatore del progetto M. V. Kopotev,

[36] Per maggiori informazioni si veda il sito *Swedish National Service Data*: https://snd.gu.se/en/catalogue/study/ext0071.
[37] Лаборатория общей и компьютерной лексикологии и лексикографии.
[38] Per consultare il *Corpus Informatico dei testi della stampa russa della fine del XX sec*. si veda: http://www.philol.msu.ru/~lex/corpus/.
[39] "[…] ежедневных и неежедневных (МН, Новая газета), "левых" (Завтра, Правда, Правда-5) и "правых", центральных и местных, общих и профессионально ориентированных (Литературная газета) газет" dal sito di MGU: http://www.philol.msu.ru/~lex/corpus/corp_descr.html. La traduzione in italiano è mia.
[40] Per consultare il sito ufficiale del CHANKO: http://h248.it.helsinki.fi/hanco/index.html.

che all'epoca operavano presso il Dipartimento di lingue e letterature slave e baltiche dell'Università di Helsinki.

L'*Helsinki Corpus* consta di circa 100 mila parole provenienti da testi tratti dal settimanale socio-politico *Itogi*, pubblicato in Russia negli anni 1996-2014, e presenta annotazioni su morfologia e sintassi, oltre a informazioni metalinguistiche sui testi.

Tutti i corpora sino ad ora menzionati hanno costituito delle tappe importanti per lo sviluppo della linguistica dei corpora russa, ma nessuno di questi si proponeva come strumento rappresentativo della lingua russa, in grado di competere con i grandi corpora americani e britannici del XXI secolo.

Come ha osservato S. Šaroff nel suo discorso pronunciato in occasione della conferenza di linguistica dei corpora, tenutasi all'Università di Lancaster nell'aprile del 2003, e successivamente pubblicato nel 2006,

> From the viewpoint of corpus linguistics, Russian is one of few major world languages that lack a comprehensive corpus of modern language use. However, the need for constructing such a corpus is growing in the corpus linguistics community both in Russia and in the rest of the world. (Šaroff, 2006: 169)

Il linguista, in queste poche righe, aveva fotografato con acuta lucidità lo stato dell'arte della linguistica dei corpora russa all'inizio del secolo, che muoveva i primi passi verso una vera e propria affermazione. Lo stesso Šaroff tentò di creare il primo equivalente russo del *British National Corpus*[41], il *Bol'šoj korpus russkogo jazyka* (Il grande corpus della lingua russa)[42].

Secondo il progetto presentato nel 2003 da Šaroff, il corpus avrebbe dovuto contenere testi rappresentativi delle principali varietà del russo moderno, per un totale di 100 milioni di parole, con annotazione per parti del discorso (*POS-Tagging*) e lemmatizzazione.

Sarebbe stato presente anche un sotto-corpus, il *Russian Standard*, costituito da 10 milioni di parole provenienti da testi di narrativa moderna, al fine di rappresentare la lingua letteraria standard; infatti, secondo quanto affermava il linguista, in Russia la lingua letteraria influenza quella dei parlanti nativi a tal punto da essere considerata una fonte autorevole e questo spiegherebbe le alte percentuali di testi di narrativa contenuti

[41] Sito ufficiale del BNC: https://www.english-corpora.org/bnc/.
[42] Nel suo articolo S. Šaroff afferma: "The objective of the project presented in the paper is to develop the Russian equivalent of the BNC, namely the *Russian Reference Corpus* (BOKR, BOl'šoj Korpus Russkogo jazyka) (Šaroff, 2006: 169).

nel *Corpus Uppsala* e in quello utilizzato da Zasorina per la creazione del *Dizionario di frequenza della lingua russa*. Inoltre, i testi si sarebbero distinti per stile del linguaggio (neutrale, ufficiale, popolare, regionale, ecc.) e per i cosiddetti domini di conoscenza (scienze naturali, scienze applicate, scienze sociali, politica, commercio, arte, ecc.).

Secondo il progetto, il *BOKR* sarebbe stato completato entro la fine del 2004, anno in cui vide ufficialmente la luce, ma con il nome di *Nacional'nyj Korpus Russkogo Jazyka* (Corpus Nazionale della Lingua Russa)[43], frutto del lavoro di un team di linguisti provenienti da San Pietroburgo, Mosca, Voronež, Kazan', Saratov e da altre sedi di istituti scientifici e accademici della Russia.

Secondo i dati presenti nella seguente tabella (*Figura 3*), tratta dal sito ufficiale di *NKRJa* ed aggiornata al 23 agosto 2021, attualmente il corpus contiene più di 2,6 milioni di testi, per un totale di oltre 1 miliardo di parole (tali cifre sono cerchiate in blu in *Figura 2*).

I. Распределение текстов по подкорпусам

Подкорпус	Число текстов	Число предложений	Число словоупотреблений	% словоупотреблений
Основной корпус	126 852	28 079 366	337 025 184	33.3%
- в том числе со снятой омонимией	2 170	519 726	6 003 393	0.6%
Газетный корпус СМИ 2000-х гг.	986 924	23 571 224	332 645 828	32.8%
Газетный региональный корпус	52 845	1 804 876	22 934 473	2.3%
Синтаксический корпус	734	88 512	1 250 923	0.1%
Диалектный корпус	1 080	67 932	395 440	0.0%
Обучающий корпус	229	65 666	664 747	0.1%
Параллельный корпус	4 121	10 539 891	140 253 613	13.8%
Поэтический корпус	93 406	1 237 505	12 820 698	1.3%
Устный корпус	4 210	1 877 238	13 399 937	1.3%
Акцентологический корпус	1 332 281	13 274 096	133 123 658	13.1%
Мультимедийный корпус	1 098	954 317	5 114 547	0.5%
Русский мультипарк	18	32 101	201 555	0.0%
Англо-русский мультипарк	30	21 421	229 338	0.0%
Древнерусский	27	117 449	573 250	0.1%
Берестяные грамоты	885	3 717	19 427	0.0%
Старорусский	6 214	341 202	8 136 258	0.8%
Церковнославянский	1 160	381 827	4 476 006	0.4%
Всего:	**2 612 114**	**82 458 340**	**1 013 264 882**	**100%**

Figura 2. Distribuzione dei testi per sotto-corpora.

Nel sito non esiste la versione in lingua inglese di questa tabella ma, come si può osservare dalla prima colonna, *NKRJa* si è progressivamente arricchito di numerosi sotto-corpora; oltre ad *Osnovnoj korpus*, troviamo *Gazetnyj korpus SMI 2000-ch godov*,

[43] Sito ufficiale di *NKRJa*: https://ruscorpora.ru/new/index.html.

Gazetnyj regional'nyj korpus, Sintaksičeskij korpus, Dialektnyj korpus, Obučajuščij korpus, Parallel'nyj korpus, Ustnyj korpus, Akcentologičeskij korpus, Mul'timedijnyj korpus, Russkij mul'tipark, Anglo-Russkij mul'tipark, Drevnerusskij, Berestjanye gramoty, Starorusskij ed infine *Cerkovnoslavjanskij*. Per un'analisi più dettagliata di questo corpus si rimanda al quarto capitolo, dove si parlerà di *NKRJa* come strumento utilizzato per questa ricerca.

Il corpus *NKRJa* ha segnato un vero e proprio punto di svolta per la linguistica dei corpora russa, superando i propositi del progetto originario che fissavano le dimensioni a 100 milioni di parole, similarmente al *British National Corpus*; esso si è affermato come strumento di ricerca costantemente ampliato ed aggiornato, al pari di modelli d'oltreoceano, come il *Corpus of Contemporary American English*[44] (*Figura 3*) e, confrontando i generi linguistici dei due corpora, si potrebbe affermare che *NKRJa* ha persino superato il database americano.

Genre	# texts	# words	Explanation
Spoken	44,803	127,396,932	Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Oprah)
Fiction	25,992	119,505,305	Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and fan fiction.
Magazines	86,292	127,352,030	Nearly 100 different magazines, with a good mix between specific domains like news, health, home and gardening, women, financial, religion, sports, etc.
Newspapers	90,243	122,958,016	Newspapers from across the US, including: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle, etc. Good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
Academic	26,137	120,988,361	More than 200 different peer-reviewed journals. These cover the full range of academic disciplines, with a good balance among education, social sciences, history, humanities, law, medicine, philosophy/religion, science/technology, and business
Web (Genl)	88,989	129,899,427	Classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages (by Serge Sharoff). Taken from the US portion of the GloWbE corpus.
Web (Blog)	98,748	125,496,216	Texts that were classified by Google as being blogs. Further classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages. Taken from the US portion of the GloWbE corpus.
TV/Movies	23,975	129,293,467	Subtitles from OpenSubtitles.org, and later the TV and Movies corpora. Studies have shown that the language from these shows and movies is even more colloquial / core than the data in actual "spoken corpora".
	485,179	1,002,889,754	

Figura 3. La composizione del Corpus of Contemporary American English.

Attualmente esistono numerosi corpora linguistici russi di vario genere; un forte impulso è stato dato dalla creazione dei cosiddetti *web* corpora, ovvero quei corpora creati attraverso dati testuali provenienti da siti internet e successivamente elaborati, come, per esempio, *ruTenTen*[45], che è stato utilizzato per questa ricerca e di cui si

[44] Sito ufficiale del *COCA*: https://www.english-corpora.org/coca/.
[45] Consultabile nel sito ufficiale di *Sketch Engine*: https://www.sketchengine.eu/#blue.

tratterà nel terzo capitolo. Sebbene in Russia la linguistica dei corpora abbia attecchito tardivamente, si è affermata in vent'anni con una forza ed una velocità sorprendenti, non solo mediante la creazione dei corpora, ma anche a livello teorico con numerosi contributi da parte dei linguisti russi e con la loro partecipazione a congressi nazionali ed internazionali.

Secondo una mia personale valutazione, se i contributi teorici fossero pubblicati anche in lingua inglese, gli autori russi riuscirebbero ad affermarsi maggiormente nel contesto internazionale, in quanto permetterebbero ad un pubblico molto più vasto, che non si limiti ai soli russofoni e studiosi di slavistica, di conoscerli; per ora non ci rimane che attendere gli sviluppi futuri della disciplina.

1.6 Conclusioni del capitolo

In questo capitolo sono state fornite diverse definizioni di linguistica dei corpora come metodologia di indagine e sono stati individuati i due principali approcci alla disciplina, quello *corpus-based* e quello *corpus-driven*.

Successivamente si è passati alla rassegna del percorso storico compiuto dalla linguistica dei corpora nel mondo anglofono, la capostipite che ha dato origine all'intera disciplina; si è tentato di ricostruire la linguistica dei corpora in Italia e nel far questo è emerso che la tradizione italiana tende a far risalire le origini della disciplina all'epoca di Dante, come indicato da F. Sabatini, mentre nei contributi di studiosi anglofoni o russofoni si indica come capostipite dell'intera disciplina il *Concordantiae Morales* di Sant'Antonio da Padova. Infine, la linguistica dei corpora russa è quella che tra le tre si è sviluppata più tardivamente, ma in breve tempo ha colmato il gap che la separava dal "mondo occidentale", sebbene l'ostacolo linguistico sia ancora presente.

Nel prossimo capitolo ci addentreremo nel cuore pulsante della disciplina: verranno esaminati il concetto di *corpus linguistico*, le caratteristiche dei corpora e le varie tipologie esistenti.

2. I corpora linguistici

2.1 Il corpus linguistico: una definizione preliminare

Il sostantivo *corpus*, al plurale *corpora*, ha origini molto antiche: deriva dal sostantivo latino neutro di III declinazione *corpŭs, corporis*. La quarta edizione aggiornata de *IL vocabolario della lingua latina* di L. Castiglioni e S. Mariotti riporta diverse definizioni della parola, tra cui: individuo, essere vivente, sostanza, materia, organismo, complesso unitario, raccolta di scritti. Proprio quest'ultimo significato si è conservato sino ai giorni nostri, tuttavia "raccolta di testi" ha un'accezione tanto generica quanto incompleta: potrebbe indicare un corpus testuale, un'antologia, un archivio di testi o addirittura una libreria. Per essere definito *linguistico*, un corpus deve soddisfare determinate caratteristiche che vedremo dettagliatamente nel prossimo paragrafo, ma prima di tutto occorre fornire un inquadramento teorico della questione, analizzando alcune definizioni che ci offre la letteratura a riguardo.

> Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi. (Barbera, 2013: 18)

> Nonostante il termine corpus si applichi a qualunque raccolta di testi parlati o scritti, nella linguistica contemporanea si considera corpus solamente una raccolta in formato digitalizzato, annotato, interrogabile, distribuito in qualche forma, con precisi requisiti di estensione, eventuale bilanciamento, rappresentatività, e di riferimento delle varietà che intende rappresentare. (Chiari, 2012: 91)

> В. П. Захаров под лингвистическим или языковым корпусом текстов понимает большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач[46]. (Saženin, 2013: 121)

> In the first instance, a "corpus" is simply any collection of written or spoken texts. However, when the term is employed with reference to modern linguistics, it tends to bear a number of connotations, among them machine-readable form, sampling and representativeness, finite size, and the idea that a corpus constitutes a standard reference for the language variety it represents. (Lüdeling, Kytö, 2008: v)

[46] "V. P. Zacharov per corpus di testi linguistico o della lingua intende un complesso di dati linguistici grande, in formato elettronico, standardizzato, strutturato, annotato e filologicamente attendibile, destinato alla risoluzione di problemi linguistici concreti". La traduzione è mia.

In altre parole, un *corpus linguistico* è una raccolta di dati testuali autentici in formato elettronico, provenienti da fonti scritte, orali o multimediali, rappresentativi di una lingua o una varietà di essa, al fine di condurre indagini linguistiche empiriche ripetibili, sia qualitative che quantitative.

Nelle quattro definizioni di corpus linguistico, sono ricorrenti alcune parole chiave, quali "formato elettronico/digitalizzato", "dati autentici" e "rappresentatività". Queste sono le prime tre caratteristiche indispensabili, la *conditio sine qua non* affinché un corpus si possa definire *linguistico*. In realtà, in base all'autore, in letteratura sono indicate anche altre caratteristiche, che illustreremo nel prossimo paragrafo.

2.2 Le caratteristiche del corpus linguistico

Per dare un riferimento teorico il più completo e accurato possibile, in questo paragrafo verranno analizzate le principali caratteristiche dei corpora, come il formato elettronico, l'autenticità dei dati, la rappresentatività, il bilanciamento, la finitezza e le dimensioni.

2.2.1 Il formato elettronico del dato linguistico

> Corpus Linguistics, in the sense of using natural language samples for linguistics, is much older than computer science. […] But the advent of computers has changed the field completely. (Volk, 2002: 355)

Questa citazione di M. Volk è appropriata per iniziare una disamina riguardante il dato linguistico informatizzato, giacché in poche righe coglie perfettamente il cuore della questione. Infatti, come abbiamo visto nel capitolo precedente, i corpora linguistici esistevano ben prima dell'invenzione dei computer, ma l'informatizzazione ha cambiato radicalmente la concezione di corpus linguistico e ha tracciato una linea netta tra i corpora pre-elettronici (secondo l'espressione russa, *doelektronnye korpusy*) e quelli informatizzati.

In primo luogo, la digitalizzazione ha permesso di immagazzinare crescenti quantità di dati linguistici e questo ha favorito la nascita di corpora sempre più grandi; dai cosiddetti corpora di prima generazione si è giunti agli attuali corpora di terza generazione. Secondo quanto riportato in *A Glossary of Corpus Linguistics* i corpora di prima generazione sono:

> a series of relatively small corpora that were created using a similar model. These include the Brown Corpus of American English (1961), the Lancaster–Oslo/Bergen (LOB) corpus

> of British English (1961), the Kolhapur Corpus of Indian English[47] (1978), the Wellington Corpus of Written New Zealand English[48] (1986) and the Australian Corpus of English[49] (1986). (Baker, Hardie, McEnery, 2006: 72)

In altre parole, i corpora di prima generazione risalgono agli anni ’60 e ’70 e solitamente contengono 1 milione di parole circa, come nel caso del *LOB Corpus* o del *Kolhapur Corpus*.

I corpora di seconda generazione degli anni ’80 e ’90 sono talvolta definiti *mega-corpora* “because of their large size (for example 100 million words or more). Examples of second generation corpora include the British National Corpus (BNC), the Bank of English (BoE) and the Longman Corpus Network” (Baker, Hardie, McEnery, 2006: 142).

Infine, i corpora di terza generazione sono quelli creati a partire dagli anni 2000 e contengono miliardi di parole. Esempi di questo tipo sono i web corpora, come il *Russian Web 2017* (*ruTenTen17*) di Sketch Engine che contiene oltre 9 miliardi di parole, oppure la versione *ruTenTen11* che ne contiene quasi il doppio (14,5 miliardi) e, infine, *English Web 2020* (*enTenTen20*), che raccoglie ben 38,1 miliardi di parole.

È bene ricordare che le dimensioni di un corpus dipendono da una serie di fattori: se il corpus viene creato per condurre un’analisi linguistica circoscritta ad un preciso arco temporale, una determinata varietà linguistica o uno specifico argomento, le dimensioni ridotte risultano adeguate e non inficiano le caratteristiche di rappresentatività e bilanciamento. Un esempio a tal proposito è il *Brexit Corpus*[50], creato mediante articoli Web dei giornali *BBC*, *Telegraph*, *Daily Mail* e il *Guardian*, e commenti in Internet, pubblicati su Twitter, nei blog o nei forum online tra il 19 e il 21 giugno 2016. Questo corpus, proprio in virtù della specificità dell’argomento e del breve arco temporale a cui si riferisce, non necessita di contenere miliardi di parole, tuttavia raggiunge la ragguardevole cifra di quasi 108,5 milioni di parole; inoltre, esiste anche la versione *Brexit corpus without retweets*, di circa 4,8 milioni di *token*[51].

Ritornando alla rivoluzione della disciplina, il formato elettronico ha reso possibile interrogare in maniera più veloce ed avanzata il contenuto del corpus. Se una

[47] Il manuale del *Kolhapur Corpus*: http://korpus.uib.no/icame/manuals/KOLHAPUR/INDEX.HTM.
[48] Il manuale del *Wellington Corpus*: http://korpus.uib.no/icame/wsc/INDEX.HTM.
[49] Il manuale dell’*Australian Corpus*: http://korpus.uib.no/icame/manuals/ACE/INDEX.HTM. Per consultare l’*Australian Corpus*: https://www.ausnc.org.au/corpora/ace.
[50] Per maggiori informazioni sul *Brexit Corpus*: https://www.sketchengine.eu/brexit-corpus/#toggle-id-1.
[51] Per *token* si intende una singola unità linguistica, molto spesso corrisponde ad una parola.

volta le concordanze erano compilate manualmente, come nel caso di *The Concordance of the New Testament* di T. Gybson (1535), oppure di *The Exhaustive Concordance of the Bible* di J. Strong (1890) viste nel precedente capitolo (§ 1.3), oggi è possibile crearle attraverso software specifici e si possono consultare in meno di un minuto con un solo click.

Oltre alle concordanze, è possibile ricercare la frequenza d'uso di una parola, oppure la frequenza d'uso di tipo grammaticale (ossia la frequenza dei tempi verbali, delle preposizioni o dei complementi a seguito di un verbo, ecc.), nonché stilare liste di frequenza, condurre indagini sulla collocazione[52] delle parole, sulla loro collisione o colligazione[53], analizzare la distribuzione di una parola nel corpus e la sua dispersione, che misura l'occorrenza di una parola o di una frase in un determinato corpus o sotto-corpus.

Queste sono solo alcune delle possibili applicazioni dei corpora permesse dall'informatizzazione del dato testuale. Tuttavia, questa sola caratteristica non è sufficiente per rendere un corpus fruibile; nel prossimo sotto-paragrafo si prenderà in esame un altro concetto chiave, quello di "autenticità del dato linguistico".

2.2.2 L'autenticità del dato linguistico

Il concetto di autenticità è stato "il principio guida dell'intera disciplina fin dai suoi albori" (Barbera, 2013: 43), eppure oggi gli autori tendono a non menzionarlo.

È bene spiegare di cosa si tratti sia per la sua importanza, sia in virtù dell'attuale dibattito tra gli studiosi circa l'autenticità o meno dei dati internet. Per citare le parole di Baroni e Bernardini:

> One of the tenets of corpus linguistics is the requirement to observe language as it is produced in authentic settings, for authentic purposes, by speakers and writers whose aim is not to display their language competence, but rather to achieve some objective through language. […] corpus linguists require collections of authentic texts (spoken and/or written) (Baroni, Bernardini, 2006: 9).

[52] *A Glossary of Corpus Linguistics* riporta la seguente definizione di c*ollocation*: "[…] certain words are more likely to occur in combination with other words in certain contexts. A collocate is therefore a word which occurs within the neighbourhood of another word" (Baker, Hardie, McEnery, 2006: 36).

[53] La *colligazione* o *collisione* è un tipo di collocazione che riguarda le relazioni sintattico-grammaticali. *A Glossary of Corpus Linguistics* offre la seguente definizione di *colligation*: "A form of collocation which involves relationships at the grammatical rather than the lexical level. For example, nouns tend to colligate with adjectives while verbs tend to colligate with adverbs. We can also apply colligation to phrases or words" (Baker, Hardie, McEnery, 2006: 36).

In altre parole, si definiscono autentici quei dati linguistici genuini, frutto dell'interazione umana scritta, orale o multimediale; sono autentici i dati linguistici prodotti a scopo comunicativo e che non sono in nessun caso manipolati dal linguista, o elicitati al fine di condurre un'analisi linguistica e per dimostrare la competenza linguistica dei parlanti.

Il criterio di autenticità è più facilmente rispettato nel caso delle fonti scritte, mentre per quanto riguarda i dati orali occorre tenere in considerazione una serie di fattori: in primo luogo i parlanti possono sentirsi in soggezione sapendo che la loro conversazione è registrata per condurre degli studi linguistici; il loro parlato non risulterà più totalmente genuino ed autentico, al tempo stesso è illegale e poco etico registrare qualcuno senza il suo permesso. Solitamente è necessario un consenso scritto e durante tutta la conversazione il dispositivo di registrazione è ben visibile, per ricordare ai parlanti che è in corso una registrazione; questo inevitabilmente determina un certo grado d'inautenticità dei dati[54].

In secondo luogo, i dati orali registrati vengono trascritti e ciò determina l'omissione o la semplice descrizione di importanti aspetti, tipici della comunicazione orale, quali il tono di voce, l'intonazione, le espressioni facciali o la gestualità. Anche in questo caso possiamo parlare di una "manipolazione" dei dati orali, che inficia l'autenticità degli stessi.

Il terzo ed ultimo fattore riguarda tanto le fonti scritte quanto quelle orali ed ha a che fare con l'errore: è il linguista a stabilire se un testo contiene o meno degli errori e dovrà prendere delle decisioni in tal senso, che necessariamente intaccheranno l'autenticità del testo.

Come accennato all'inizio di questo paragrafo, è in corso un dibattito tra gli studiosi circa l'autenticità o meno dei dati internet: secondo W. H. Fletcher, "Web pages are typically anonymous and Web server location is no certain guide to origin, so it is difficult to establish authorship and provenance and to assess the reliability, representativeness and authoritativeness of texts, both for their linguistic form and their content" (Fletcher, 2004: 2). Oltre a questo, nel Web esistono innumerevoli profili fittizi

[54] Al riguardo occorre menzionare le parole di A. Stefanowitsch: "The ethical and legal problems in recording unobserved spoken language cannot be circumvented, but their impact on the authenticity of the recorded language can be lessened in various ways – for example, by getting general consent from speakers, but not telling them when precisely they will be recorded" (Stefanowitsch, 2020: 27).

o cosiddetti “fake”, dei quali non si può effettivamente stabilire l’attendibilità ed autenticità dei commenti. Non si può stabilire con certezza se un utente abbia realmente l’età, il genere, la nazionalità, l’istruzione o la professione dichiarate al momento della creazione del profilo, la “persona” del profilo può addirittura non esistere.

Infine, il problema dell’autenticità del Web è spesso legato a quella che M. Gatto definisce *authorativeness*, ossia “autorevolezza” (Gatto, 2008: 13): le pagine web spesso contengono significative quantità di errori grammaticali, parole scritte in modo errato, errori di punteggiatura, linguaggio ripetitivo o frammentario, assenza di coesione del testo e via dicendo. Questo materiale non è affidabile dal punto di vista linguistico e non è rappresentativo di una determinata lingua, ma piuttosto della lingua utilizzata nel mondo del Web. Nel prossimo paragrafo analizzeremo il concetto di rappresentatività, quale terza caratteristica indispensabile del corpus linguistico.

2.2.3 La rappresentatività del corpus linguistico

La rappresentatività di un corpus linguistico è strettamente legata ad altre caratteristiche indispensabili, quali il bilanciamento, le dimensioni e la finitezza.

Barbera definisce la rappresentatività come “un campione, un *sample*, della lingua analizzata che ne riproduca idealmente, seppur ‘in miniatura’, tutte le caratteristiche, pur nell’impossibilità di avere, in ultima analisi, le stesse uguali ed identiche caratteristiche della lingua oggetto di analisi” (Barbera, 2013: 44).

In altre parole, un corpus linguistico è rappresentativo quando riflette, sia qualitativamente che quantitativamente, una determinata varietà linguistica e quando la distribuzione dei fenomeni linguistici al suo interno è identica a quella del linguaggio nel suo insieme, al fine di poter fare delle generalizzazioni riguardanti la lingua oggetto d’analisi.

Per molti linguisti, padri della disciplina come Biber (1993) e Atkins, Clear, Ostler (1992), sarebbe proprio la rappresentatività a differenziare un corpus linguistico da una semplice raccolta di testi disponibili in formato elettronico o tratti da un archivio digitale.

Naturalmente è impossibile includere in un corpus tutte le occorrenze linguistiche, sia scritte che parlate, pertanto, al fine di rappresentare al meglio una lingua nella sua totalità, in fase di progettazione del corpus, i linguisti dovranno stabilire chiaramente la

popolazione target che intendono rappresentare, l'utilizzo a cui il corpus è destinato, i metodi di campionamento, i testi da includere nel corpus, il numero di parole per campione testuale, le tipologie di testo e così via. Queste scelte "strutturali" influiscono sul livello di rappresentatività, al tempo stesso la progettazione del corpus dipende strettamente da ciò che esso vuole rappresentare.

Non tutti gli autori ritengono sia davvero possibile raggiungere un certo grado di rappresentatività[55]; ad esempio, secondo A. Stefanowitsch:

> For a corpus to be representative (or "balanced"[56]), its text categories should accurately reflect both quantitatively and qualitatively the language varieties found in the speech community whose language is represented in the corpus. However, it is clear that this is an ideal that is impossible to achieve in reality for at least four reasons. (Stefanowitsch, 2020: 29)

Il primo motivo è che si possono conoscere i parametri di distribuzione di genere, età o istruzione di un campione di popolazione, ma non il rapporto di distribuzione della lingua scritta rispetto a quella orale, della lingua letteraria rispetto a quella giornalistica. In secondo luogo, non si può essere certi che tutte le manifestazioni di una lingua rappresentino il sistema linguistico allo stesso modo. In terzo luogo, le comunità linguistiche non sono omogenee, pertanto basare il criterio di rappresentatività su una proporzione matematica potrebbe non fornire una rappresentazione realistica della lingua. Infine, ci sono delle varietà linguistiche impossibili da campionare, come le conversazioni riguardanti le confessioni religiose o quelle tra avvocato e cliente, psicologo e paziente ecc. Stefanowitsch conclude asserendo: "Given the problems discussed above, it seems impossible to create a linguistic corpus meeting the criterion of representativeness" (Stefanowitsch, 2020: 30).

Dopo aver analizzato il concetto di rappresentatività, si approfondirà il concetto di bilanciamento e si tenterà di spiegare la relazione tra queste due caratteristiche.

2.2.4 Il bilanciamento del corpus linguistico

[55] Questa obiezione fu mossa per la prima volta da N. Chomsky, il quale riteneva che per quanto grande potesse essere un corpus linguistico, non sarebbe mai stato rappresentativo in quanto "Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list" (Chomsky, 1962: 159).

[56] L'autore utilizza i termini rappresentatività e bilanciamento come sinonimi, così come Barbera (Barbera, 2007: 50). Altri autori, invece, utilizzano i due termini distintamente (Lew, 2009: 7) e (Wynne, 2005: 14). Infine, c'è chi ritiene che "the word representative has tended to fall out of discussions, to be replaced by the meeker balanced" (Kilgarriff, Grefenstette, 2003: 342).

A Glossary of Corpus Linguistics fornisce una definizione piuttosto generica di corpus bilanciato: "A corpus that contains texts from a wide range of different language genres and text domains, so that, for example, it may include both spoken and written, and public and private texts" (Baker, Hardie, McEnery, 2006: 18). In questa sede si cercherà di fornire una definizione più specifica di bilanciamento e di proporre un esempio pratico a riguardo.

Un corpus linguistico si definisce bilanciato quando include al suo interno un ampio spettro di categorie testuali differenti, rappresentative della lingua o della varietà linguistica oggetto d'indagine. In altre parole, è proprio il bilanciamento di un corpus a garantire la rappresentatività dello stesso ed è per questo motivo che, a nostro avviso, occorre trattare i due termini distintamente e non usarli come sinonimi intercambiabili (cfr. Nota 56).

Un corpus ben bilanciato può includere varie tipologie testuali scritte (come articoli di giornale, libri, poesie, diari, ecc.) e trattare argomenti differenti (come l'arte, il commercio e la finanza, la cultura, le scienze sociali, la religione, la cucina, la moda, ecc.), può ricomprendere al suo interno testi prodotti in archi temporali differenti, di autori di età, genere, nazionalità differenti. E la medesima cosa avviene per i dati orali.

Naturalmente, il bilanciamento dipende in primis dalla tipologia di corpus linguistico: un corpus generico (§ 2.3.2) solitamente contiene sia fonti scritte che orali, mentre un corpus specialistico non necessariamente comprende entrambe.

Un corpus bilanciato è il *British National Corpus*[57], in cui la parte costituita da fonti scritte corrisponde a circa il 90% del totale, mentre le fonti orali corrispondono al 10%.

Nelle *Tabelle 1, 2* e *3*[58] sono riportati alcuni dei criteri di progettazione utilizzati nella creazione del *BNC*, al fine di fornire un quadro indicativo di un corpus bilanciato.

[57] Secondo quanto riportato dal sito ufficiale: "The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins".
Consultabile al link: http://www.natcorp.ox.ac.uk/corpus/index.xml.

[58] Le tabelle sono tratte dal seguente sito: http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html.

Written Medium					
	texts	**w-units**	**%**	**s-units**	**%**
Book	1411	50293803	57.18	2887523	57.88
Periodical	1208	28609494	32.52	1487644	29.82
Miscellaneous published	238	4233135	4.81	287700	5.76
Miscellaneous unpublished	249	3538882	4.02	220672	4.42
To-be-spoken	35	1278618	1.45	104665	2.09

Tabella 1. Tipologie testuali del *BNC.*

Publication date					
	texts	**w-units**	**%**	**s-units**	**%**
Unknown	162	1831585	1.86	126416	2.09
1960-1974	46	1718449	1.74	119510	1.98
1975-1984	169	4730889	4.80	257962	4.28
1985-1993	3672	90082860	91.58	5522396	91.63

Tabella 2. Variazione diacronica dei testi del *BNC.*

Written Domain					
	texts	**w-units**	**%**	**s-units**	**%**
Imaginative	476	16496420	18.75	1352150	27.10
Informative: natural & pure science	146	3821902	4.34	183384	3.67
Informative: applied science	370	7174152	8.15	356662	7.15
Informative: social science	526	14025537	15.94	698218	13.99
Informative: world affairs	483	17244534	19.60	798503	16.00
Informative: commerce & finance	295	7341163	8.34	382374	7.66
Informative: arts	261	6574857	7.47	321140	6.43
Informative: belief & thought	146	3037533	3.45	151283	3.03
Informative: leisure	438	12237834	13.91	744490	14.92

Tabella 3. Le tipologie di argomenti contenuti nel *BNC.*

Come si può osservare, il *British National Corpus* è piuttosto ricco dal punto di vista della variazione linguistica diacronica, diafasica, diamesica e diatopica[59] ed in tal senso può definirsi ben bilanciato e rappresentativo dell'inglese britannico della fine del XX secolo, sia parlato che scritto[60]. Nel prossimo paragrafo si prenderanno in considerazione altre caratteristiche dei corpora, citate più di rado dagli autori.

2.2.5 La finitezza e le dimensioni del corpus linguistico

Altre due caratteristiche importanti dei corpora linguistici sono la finitezza e le dimensioni; in questo paragrafo verranno illustrate insieme, poiché come è stato lucidamente puntualizzato "da un lato la 'rappresentatività' implica selezione, e quindi implicitamente finitezza; dall'altro le dimensioni 'più ampie possibili' di un corpus, talvolta citate tra i criteri definitori di un corpus, portano a vagheggiare una dimensione idealmente infinita" (Barbera, Corino, Onesti, 2007: 51). In questa sede cercheremo di chiarire se queste caratteristiche sono tra loro complementari o in contrapposizione.

La finitezza, secondo alcuni autori, fa parte delle caratteristiche indispensabili dei corpora, al pari di autenticità, rappresentatività e formato elettronico del dato linguistico; ad esempio, McEnery e Wilson ritengono che un corpus linguistico sia per definizione "a body of text of a finite size" (McEnery, Wilson, 2006: 30).

Un corpus, per quanto grande, viene creato mediante precisi criteri di selezione che implicano un numero finito di testi; infatti, poiché mediante i corpora vengono eseguite delle analisi di tipo statistico, è necessario che l'insieme di dati linguistici su cui si opera, detto campione[61], sia finito. Per di più "la finitezza di un corpus ne garantisce la possibilità di operare entro confini scientificamente ed univocamente

[59] Per variazione diacronica si intende "la variazione lungo l'asse del tempo e nella storia, che dà luogo a diversi stati di lingua successivi, e che, come si è detto, viene più precisamente designata con il termine di mutamento (o cambiamento) linguistico". La variazione diafasica "si manifesta attraverso le diverse situazioni comunicative e consiste nei differenti modi in cui vengono realizzati i messaggi linguistici in relazione ai caratteri dello specifico contesto presente nella situazione; viene quindi anche detta variazione situazionale". Per variazione diamesica "si intende la capacità di una lingua di variare a seconda del mezzo o canale adottato, sia esso scritto (grafico-visivo) o parlato (fonico-acustico)". Infine, per variazione diatopica "(dal gr. diá «attraverso» e tópos «luogo») si intende la variazione linguistica su base geografica". Fonte: https://www.treccani.it/enciclopedia/variazione-linguistica_(Enciclopedia-dell'Italiano)/.

[60] Questa è la *mission* dichiarata nel sito ufficiale del *BNC*, nella sezione intitolata "About the BNC" al link: http://www.natcorp.ox.ac.uk/.

[61] Il campione statistico viene definito come un "Gruppo di unità elementari che formano un sottoinsieme della popolazione". Per maggiori informazioni a riguardo consultare il seguente sito: https://www.treccani.it/enciclopedia/campione-statistico_%28Dizionario-di-Economia-e-Finanza%29/.

stabiliti dal linguista, non solo a livello di bilanciamento del materiale in esso contenuto (che non potrebbe essere tenuto sotto controllo in un corpus "aperto"), ma anche a livello di completa ripetibilità, *ceteris paribus*, degli esperimenti" (Barbera, Corino, Onesti, 2007: 51).

La questione della finitezza del corpus è una delle principali obiezioni mosse nei confronti di quegli studiosi che considerano il World Wide Web un corpus linguistico: il Web, infatti, è una fonte potenzialmente infinita di dati linguistici, ma a causa delle sue dimensioni colossali è impossibile compiere delle semplici indagini statistiche riguardanti la quantità di dati linguistici presenti o la percentuale di pagine in una determinata lingua[62]. Secondo molti linguisti, tra cui Sinclair (2005), Barbera (2013) e molti altri, il Web non può essere considerato un corpus linguistico, proprio in virtù delle sue dimensioni non finite, ma di questo ci occuperemo specificatamente nel Capitolo III, dedicato al *Web as Corpus*.

Per quanto riguarda la questione delle dimensioni, questo requisito è mutato nel corso del tempo grazie all'evoluzione tecnologica: se negli anni '60 e '70 del secolo scorso venivano considerati di dimensione standard i corpora contenti 1 milione di parole, come il *LOB Corpus*, negli anni '80 e '90 si parlava già di mega-corpora di oltre 100 milioni parole ed oggi con i web corpora si superano le decine di miliardi di parole (§ 2.2.1).

Attualmente molti autori sostengono che "there is no maximum size" (Wynne, 2005: 15), o ancora che "Assuming that one of the main features of a representative corpus is its size, then a 100-million *token* corpus, considered a standard at the beginning of this century, now appears in many cases to be insufficient to receive relevant statistical data. […] we call a corpus 'very large' if its size exceeds 10 billion tokens" (Benko, Zacharov, 2016: 80) e infine "In the age of the World Wide Web, corpus size is practically limited only by technical considerations" (Stefanowitsch, 2020: 37).

[62] Il sito https://www.worldwidewebsize.com/ riporta una stima delle dimensioni del World Wide Web in termini di pagine Web. Secondo i dati aggiornati al 22 ottobre 2021 (data di consultazione) il cosiddetto "Indexed Web" (quello che non include il cosiddetto *Deep Web*) raggiungerebbe i 4,74 miliardi di web-page, ma non viene fornita una stima del numero di *type* o *token* contenuti nel Web. Tuttavia questi dati sembrano esigui se confrontati con quelli riportati da Gatto, secondo cui alla fine di Gennaio 2005 il Web contava almeno 11,5 miliardi di pagine, per oltre 8 trilioni di parole (Gatto, 2009: 20). Per altre possibili stime sulle dimensioni del web si veda (Kilgarriff, Grefenstette, 2003: 337).

Questo non deve portare il lettore all'errata conclusione che *bol'še - lučše,* che in russo significa "di più è meglio". Sicuramente la dimensione è una caratteristica fondamentale che garantisce la rappresentatività di un corpus; tuttavia le dimensioni dipendono da vari fattori, come la tipologia di corpus che si vuole creare e il tipo d'indagine che si vuole condurre. Come viene puntualizzato in *A Glossary of Corpus Linguistics* alla voce *size* (dimensioni), per uno studio sulla prosodia sono sufficienti 100.000 parole di parlato spontaneo; per un'analisi morfologica delle forme verbali è necessario un corpus di almeno mezzo milione di parole, che però non sarebbe sufficiente per condurre un'analisi lessicografica (Baker, Hardie, McEnery, 2006: 146).

A fronte delle cifre riportate in *A Glossary of Corpus Linguistics*, sorge spontaneo chiedersi quali siano le dimensioni minime e massime di un corpus linguistico. Per quanto riguarda la prima domanda, pochi autori forniscono una risposta: Barbera asserisce che 250 mila *token* si potrebbero considerare le dimensioni minime, ma corpora come il *SUSANNE*[63], di meno di 130 mila parole, oppure come il *Corpus Taurianense*[64], di circa 260 mila *token*, sono la dimostrazione che "un corpus piccolo, ma ben controllato ed accuratamente annotato, può giocare un ruolo assai importante nello sviluppo della linguistica" (Barbera, 2013: 45). Chiari, parlando di estensione dei corpora per le analisi lessicali, definisce corpora di piccole dimensioni quelli contenenti dalle 15 alle 100 mila parole (Chiari, 2007: 41).

Per quanto riguarda le dimensioni massime, diversi autori trattano la questione: secondo O'Keeffe e McCarthy le dimensioni dipendono dalla rappresentatività, ossia se è stato raccolto materiale sufficiente per rappresentare la lingua oggetto di indagine, e da questioni pratiche, come i limiti di tempo (O'Keeffe, McCarthy, 2010: 32). Anche secondo M. Gatto, "A corpus aimed at being representative of general language certainly needs to be quite large if it is to provide evidence for a sufficiently large number of language items" (Gatto, 2014: 14); secondo A. Stefanowitsch, "it must be large enough to contain a sample of instances of the phenomenon under investigation

[63] Il *SUSANNE Corpus* comprende un sottoinsieme di circa 130.000 parole del Brown Corpus dell'inglese americano ed è accuratamente annotato. Per maggiori informazioni sul corpus consultare il sito: https://www.grsampson.net/SueDoc.html. *SUSANNE* è attualmente interrogabile online tramite Sketch Engine.

[64] Il *Corpus Taurinense* è un corpus di italiano antico, costituito da 22 testi fiorentini della seconda metà del XIII secolo, completamente annotati e disambiguati e di piccole-medie dimensioni, per un totale di 257.185 *token*. Per consultarlo online si veda http://www.bmanuel.org/projects/ct-HOME.html. Per il manuale *Schema e storia del Corpus Taurinense. Linguistica dei corpora dell'italiano antico* si veda Barbera, 2009.

that is large enough for analysis", e aggiunge "it must be large enough to contain sufficiently large samples of every grammatical structure, vocabulary item, etc." (Stefanowitsch, 2020: 38). Per concludere, le dimensioni sia minime che massime dipendono dalla rappresentatività del corpus.

A mio avviso, *in medio stat virtus*, "la virtù sta nel mezzo": sulla scia di molti autori italiani, ritengo che per condurre un'analisi linguistica meticolosa sia più importante la qualità dei dati, piuttosto che la loro quantità; infatti, spesso risultano più adeguati corpora di dimensioni contenute, ma accuratamente annotati (§ 2.3.1), piuttosto che corpora contenenti miliardi di parole ma poco precisi dal punto di vista del controllo e della annotazione dei testi, come nel caso di alcuni web corpora.

2.2.6 Altre caratteristiche del corpus linguistico

Altre importanti caratteristiche dei corpora sono la ripetibilità, la replicabilità o riproducibilità[65] e la omogeneità dei dati linguistici.

I risultati di un'analisi condotta su un corpus linguistico devono essere necessariamente ripetibili, nel senso che le ricerche condotte dallo stesso linguista, con lo stesso metodo di misurazione, strumento di misura e nelle medesime condizioni di utilizzo dello strumento, devono dare sempre lo stesso identico risultato.

Inoltre, questi risultati devono essere replicabili o riproducibili: in altre parole, un qualsiasi ricercatore esterno, utilizzando gli stessi strumenti e metodi di indagine, dovrà ottenere misurazioni identiche allo studio originale.

Il concetto di omogeneità dei dati linguistici, invece, viene menzionato raramente[66]; secondo quanto riportato in *A Glossary of Corpus Linguistics*, "When discussing corpus design, a corpus is said to be homogenous if the text it contains has been drawn from one source or a narrow range of sources" (Baker, Hardie, McEnery, 2006: 85). Ne consegue che i corpora generali avranno un basso grado di omogeneità, poiché contengono un'ampia varietà di fonti testuali, al contrario i corpora specialistici risulteranno più omogenei nei contenuti. A differenza delle caratteristiche fondamentali come "formato elettronico", "autenticità" e "riproducibilità" del dato linguistico,

[65] Per uno studio approfondito dei concetti di *replicability*, *repeatability*, e *reproducibility* si veda K. Bretonnel Cohen *et alii*, 2018. Inoltre, si veda Barbera, Corino, Onesti (2007: 51) per il concetto di ripetibilità; McEnery, Hardy (2012: 16) per il concetto di replicabilità ed infine Hundt, Nesselhauf, Biewer (2007: 11) per quello di riproducibilità.

[66] L'omogeneità viene menzionata da Baker, Hardie, McEnery (2006: 85) e Wynne (2005: 20).

l'omogeneità non è una peculiarità distintiva dei corpora linguistici e le scelte riguardanti l'omogeneità o eterogeneità dei dati devono essere fatte coerentemente, in relazione al corpus che si va a creare.

Dopo aver analizzato le caratteristiche del corpus linguistico, nel prossimo paragrafo illustreremo le tipologie di corpora esistenti.

2.3 Le tipologie di corpora linguistici

Esistono numerose tipologie di corpora, sulla base della variazione linguistica diacronica, diafasica, diamesica e diatopica, del tipo di annotazione linguistica, della tipologia di dati linguistici contenuti e così via.

Nei prossimi sotto-paragrafi analizzeremo le differenti tipologie, ma per uno sguardo d'insieme è possibile consultare la tabella riassuntiva *Le tipologie di corpora linguistici* in Appendice a pagina 141.

2.3.1 Corpora annotati e grezzi

Lo sviluppo informatico ha consentito la creazione di software specifici in grado di arricchire i dati testuali con differenti tipi di annotazione, al fine di rendere i corpora dei raffinati strumenti di ricerca, idonei a realizzare indagini linguistiche complesse.

In primis occorre citare la "tokenizzazione" e il markup, perché è proprio grazie a queste due operazioni preliminari che una raccolta di testi in formato elettronico può dirsi propriamente "corpus linguistico". Il *token* è la singola unità linguistica o l'elemento linguistico minimo; pertanto, per tokenizzazione s'intende un processo automatico mediante il quale i testi di un corpus vengono "segmentati" in singoli *token* e separati da uno spazio, chiamato *blank*[67].

Il markup, invece, fornisce informazioni extra-testuali o, per utilizzare le parole di Barbera, "soprasegmentali", in contrapposizione a quelle "segmentali" dei *token* (Barbera, 2013: 28); in altre parole, è il più alto livello di annotazione. Esistono due differenti tipi di markup, il metadato e il markup testuale. I metadati, chiamati anche informazioni metalinguistiche o markup esterno[68], aggiungono informazioni extra-

[67] Si noti che "possono esservi varie gradazioni e sfumature di (più o meno) non tokenizzato e di (più o meno) tokenizzato, il discrimine tra le due categorie è non di meno netto e sempre tracciabile" (Barbera, Corino, Onesti, 2007: 28).
[68] Barbera suggerisce di distinguere il markup in *markup* esterno e *markup* interno e filologico (Barbera, 2013: 28).

testuali, come i dati anagrafici di un autore, quando, da chi e dove è stato pubblicato un testo, in che lingua è scritto, il genere testuale e così via.

Il markup testuale, chiamato anche markup interno e filologico, fornisce informazioni interne al testo, come: l'utilizzo del corsivo, del grassetto o di caratteri speciali; l'utilizzo di discorso diretto o indiretto, di abbreviazioni o acronimi; le informazioni circa i capitoli, i paragrafi e i numeri di pagina. Questo tipo di markup è molto importante per i corpora di lingua parlata, poiché fornisce informazioni sugli accenti, sull'utilizzo di un linguaggio dialettale o di una varietà non-standard, indica i momenti di sovrapposizione tra due o più parlanti, le pause, le esitazioni, le autocorrezioni e le false partenze, tutte caratteristiche tipiche del parlato.

L'annotazione o etichettatura linguistica (dall'inglese *tagging*[69]) è un particolare tipo di *markup* che aggiunge al testo informazioni linguistiche di varia natura. Le tipologie più frequenti di annotazione sono le seguenti:

- Lemmatizzazione: è un processo automatico di riduzione delle forme flesse al loro rispettivo lemma, ossia la forma base di una parola.
- *POS-tagging* (dall'inglese *Part Of Speech-tagging*): chiamato anche "annotazione morfosintattica"; è il processo di annotazione del corpus per parti del discorso, come nomi, verbi, aggettivi, avverbi, pronomi, ecc.
- *Parsing*: chiamato anche *syntactic parsing*, *treebanking* o *treebank annotation*[70]. È un tipo di annotazione che aggiunge informazioni di tipo sintattico a livello del sintagma e del periodo; ad esempio, individua i sintagmi nominali, i sintagmi verbali ecc., oppure indica la struttura della frase e le dipendenze sintattiche dei vari periodi.
- Annotazione semantica: chiamata anche *semantic annotation* o *semantic tagging*[71]; aggiunge informazioni di tipo semantico e consente di condurre

[69] Secondo quanto riporta *A Glossary of Corpus Linguistics* il termine *tagging* è definito come "A more informal term for the act of applying additional levels of annotation to corpus data" (Baker, Hardie, McEnery, 2006: 154).

[70] *Treebanking* è utilizzato da Lüdeling, Kytö, 2008: 421; *Treebank* viene utilizzato da Stefanowitsch, 2020: 43; infine, *Treebank* e *Treebanking annotation* viene utilizzato da McEnery, Hardie, 2012: 42 e *ivi* p. 252.

[71] Per osservare nel concreto come l'annotazione semantica viene elaborata ed applicata ad un corpus di testi, si veda lo studio di V. Starko riguardante il Corpus Regionale Generale Annotato della Lingua Ucraina (Starko, 2020). Un altro studio molto interessante, che rivela le potenzialità dell'annotazione semantica, riguarda delle analisi semantiche effettuate per conto di un'azienda che sviluppa dispositivi per il controllo e la gestione dello stress. Questa ricerca ha analizzato un campione di commenti postati su

operazioni di disambiguazione nei casi di polisemia, di omografi e omofoni, e di indagini circa l'uso metaforico delle parole, come la metonimia e la sineddoche, l'utilizzo dell'ironia, ecc.

- *Error-tagging*: utilizzato soprattutto nei cosiddetti *learner corpora*, corpora costituiti da testi (o trascrizioni di parlato) prodotti da apprendenti di una lingua L2. L'annotazione dell'errore indica dove lo studente ha commesso l'errore, il tipo di errore, da cosa è determinato, ecc.

Esistono inoltre annotazioni specifiche per i corpora di parlato, come l'annotazione fonetica e fonologica, in cui vengono inserite informazioni circa i fonemi contenuti all'interno del testo o l'annotazione ortoepica, riguardante la corretta pronuncia delle parole; l'annotazione prosodica aggiunge informazioni circa l'accento, l'intonazione, la lunghezza delle pause; l'annotazione pragmatica riguarda invece il significato, il contesto e l'interpretazione dell'atto comunicativo.

Un'interessante tipologia di annotazione è quella gestuale e del linguaggio dei segni: si tratta di un'annotazione del linguaggio non verbale riguardante lo sguardo, le espressioni facciali, come quelle della bocca o delle sopracciglia, il movimento degli occhi, delle mani e della testa.

L'annotazione può essere di tre differenti tipi: il primo tipo è l'annotazione automatica, svolta mediante specifici software il cui livello di errore è solitamente basso, ma può variare in base al tipo di annotazione. Il secondo tipo di annotazione è quella eseguita manualmente da un team di linguisti e ciò implica un certo rischio di errore umano e tempi di realizzazione più lunghi. Infine, il terzo tipo è l'annotazione semi-automatica, svolta in parte da software e programmi appositi, in parte da linguisti; questa tipologia è sicuramente più accurata, poiché permette un doppio controllo sulla qualità dell'annotazione e implica un impegno accettabile in termini di tempo e lavoro.

Dopo questo excursus sull'annotazione del testo è possibile distinguere i corpora annotati e grezzi: i corpora annotati sono arricchiti da un qualche livello di annotazione, al fine di condurre analisi complesse sui dati testuali. Questi corpora sono innumerevoli, tra quelli già citati troviamo il *Brown Corpus*, il *BNC* ed anche il *Nacional'nyj Korpus*

Twitter per comprendere, mediante annotazione ed analisi semantica, le pratiche di gestione dello stress da parte degli utenti (Jussila J., Alkhammash E., Saleh Alghamdi N., Madhala P., Ayoub Khan M., 2022).

Russkogo Jazyka (da ora abbreviato *NKRJa*); proprio quest'ultimo detiene una singolare varietà di annotazioni: ad esempio, il *Korpus poetičeskich tekstov* di *NKRJa* ('Corpus dei testi poetici'), inaugurato nel 2006 e contenente testi poetici scritti dal XVIII secolo ai giorni nostri, presenta, oltre all'annotazione metatestuale, morfologica e semantica, una particolare annotazione poetica che indica il tipo di rima, di strofa, di ritmo, di metro e così via.

Un corpus che include l'*error-tagging* è il *Russian Error-Annotated Learner English Corpus* (*REALEC*)[72], con oltre 3,3 milioni di parole provenienti da testi in inglese scritti da studenti madrelingua russi della *Vyssšaja Škola Ekonomiki*[73]. Questo corpus presenta quattro differenti livelli di annotazione: l'errore a livello di regola grammaticale, la possibile causa dell'errore, il grado di "danno" grammaticale o linguistico causato dall'errore e il grado di danno pragmatico causato dall'errore (se e in che misura l'errore inficia la comprensione del testo) (Kuzmenko, Kutuzov, 2014: 90).

I corpora non annotati, chiamati anche "grezzi", dall'aggettivo inglese *raw*, non presentano alcun tipo di annotazione e sono il principale strumento utilizzato dall'approccio corpus-driven della linguistica dei corpora (§ 1.2), secondo cui l'annotazione inficia la purezza e l'unità del dato testuale.

2.3.2 Corpora generici e specialistici

I corpora generici vengono chiamati anche corpora di riferimento, dall'inglese *reference corpora*, o *multi-purpose corpora*. Nel sito di *NKRJa* è presente questa definizione relativa al Corpus stesso, che ben si presta a descrivere i corpora di riferimento nel loro insieme:

> Он характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. […] Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию)[74].

[72] Il sito ufficiale del *Russian Error-Annotated Learner English Corpus* è consultabile al link: https://realec.org/.

[73] Chiamata anche, secondo la denominazione inglese, *Higher School of Economics* (*HSE*). Il sito ufficiale è disponibile al link: https://www.hse.ru/en/.

[74] "È caratterizzato da rappresentatività, o bilanciamento dei testi. Ciò significa che il corpus contiene, se possibile, tutti i tipi di testi scritti e orali presenti nella lingua attuale (narrativa di vario genere, pubblicistica, testi didattici, scientifici, commerciali, colloquiali, dialettali, ecc.), e che tutti questi testi

In altre parole, un corpus di riferimento deve essere sufficientemente ampio da ricomprendere tutte le varietà significative di una lingua, il lessico e le strutture morfosintattiche tipiche della lingua parlata e scritta; inoltre, deve includere varie tipologie testuali e di parlato (narrativa, divulgazione scientifica, testi accademici, poetici, conversazioni, parlate dialettali, ecc.), riguardanti svariati argomenti (dalla finanza alla moda, dalla letteratura alla cucina, ecc.). I corpora generici solitamente hanno grandi dimensioni e sono progettati in modo da includere al loro interno un'ampia varietà di sotto-corpora e in virtù di questo divengono uno strumento indispensabile per la creazione di dizionari, tesauri o grammatiche.

Un esempio è il già citato *NKRJa*, che contiene numerosi sotto-corpora, come il corpus del russo antico, il corpus accentato, quello dialettale, multimediale, poetico, ecc.[75]. Un altro esempio è il *General'nyj Internet-Korpus Russkogo Jazyka* (abbreviato *GIKRJa*)[76], un web-corpus generale della lingua russa di Internet, contenente 20 miliardi di parole. Questo corpus, creato nel 2015 mediante specifici software di download e markup testuale automatici, contiene testi tratti da *Runet* (dal russo *Russkojazyčnyj Internet* o *Russkij Internet*, letteralmente 'Internet di lingua russa' e 'Internet russo'), in particolare, da *VKontakte*[77], *LiveJournal*[78], *Blogs@Mail.ru*[79] e dalla rivista *Žurnalnyj Zal*[80].

Come evidenziato da A. Č. Piperski, è possibile utilizzare *GIKRJa* per analizzare segmenti della lingua, come le varietà regionali della lingua russa presenti nei già citati social media. Ad esempio, gli utenti di *LiveJournal*, al momento dell'iscrizione,

---

sono possibilmente inclusi nel corpus in proporzione alla loro percentuale nella lingua del rispettivo periodo. […] In secondo luogo, il corpus contiene speciali informazioni aggiuntive riguardanti le caratteristiche dei testi in esso contenuti (il cosiddetto markup, o annotazione)". Citazione tratta dal sito di *NKRJa*: https://ruscorpora.ru/new/corpora-intro.html. La traduzione è mia.

[75] Per consultare tutte le tipologie di sotto-corpora contenuti in *NKRJa* si veda il sito: https://ruscorpora.ru/new/corpora-intro.html.

[76] Per tutte le informazioni riguardanti il *GIKRJa* si veda: http://www.webcorpora.ru/.

[77] *VKontakte* (anche abbreviato *VK*) è il principale social network in Russia, nato nel 2006. Secondo i dati riportati nel sito ufficiale, nel primo trimestre del 2021 *VK* ha raggiunto i 73 milioni di utenti. Per maggiori informazioni si veda il sito al link: https://vk.com/press/q1-2021-results.

[78] Dal russo *Živoj Žurnal*. È un servizio di social network, nato nel 1999, che unisce le caratteristiche dei social network a quelle di una piattaforma per blogger, in cui gli utenti possono creare contenuti, come un diario, un blog o una rivista. Consultabile al link: https://www.livejournal.com/.

[79] Blogs@Mail.ru è una piattaforma che consente a qualsiasi utente con una casella di posta Mail.ru di aprire un blog personale. La piattaforma è nata nel 2005, ma nel 2014 è stata definitivamente chiusa.

[80] È un sito che al suo interno contiene i numeri di diverse riviste di varia natura, tra le quali *Novyj Mir*, *Družba Narodov*, *Znamja*, *Inostrannaja literatura* e molte altre ancora. Il sito ufficiale è consultabile al link: https://magazines.gorky.media/.

indicano la propria regione di residenza e d'istruzione. Sulla base di questi dati sono state individuate 16 varietà regionali (dei territori della Federazione Russa e dell'Ucraina), corrispondenti a 16 sotto-corpora, tra cui citiamo le varietà di San Pietroburgo (24,15%), Mosca (9,58%), Kiev (9,21%), Repubblica di Tatarstan (2,78%), e molte altre (Piperski, 2013).

Per quanto riguarda i corpora specialistici, *A Glossary of Corpus Linguistics* fornisce la seguente definizione: "A corpus which has been designed for a particular research project, for example, lexicography for dictionary compilation, or to study particular specialist genres of language: child language, English for Academic Purposes etc." (Baker, Hardie, McEnery, 2006: 147). Nello specifico, si definiscono specialistici o *special-purpose corpora* quei corpora contenenti dati testuali riguardanti uno specifico ambito di competenza o un determinato gruppo sociolinguistico, un periodo storico, una precisa tipologia testuale. In virtù del loro carattere altamente specifico e dell'attenta selezione testuale, questi corpora sono generalmente di dimensioni contenute rispetto ai corpora di riferimento e non sono adatti per condurre indagini linguistiche di carattere generale. Al contrario, sono lo strumento ideale per le analisi riguardanti uno specifico argomento, linguaggio, periodo storico o gruppo sociolinguistico.

Un esempio è il *Korpus Russkich Učebnych Tekstov* (*KRUT*)[81], una raccolta di testi russi per un totale di 3,1 milioni di parole scritti da studenti frequentanti una università o un master. Gli studenti provengono da diverse facoltà e i testi sono di vario genere, come tesine, tesi di laurea, saggi, abstract e compiti per casa. Il corpus presenta informazioni metalinguistiche (come l'età, il sesso dell'autore, se l'autore è bilingue, l'anno di studio, ecc.), annotazioni morfosintattiche e, soprattutto, l'*error-tagging*, che indica il tipo di errore linguistico e il motivo dell'errore (lessicale, stilistico, grammaticale o sintattico). Si tratta di un corpus specialistico in quanto raccoglie dati linguistici tratti da specifici generi testuali, di una specifica fascia di età e periodo storico, di un determinato gruppo socioculturale e sociolinguistico.

2.3.3 Corpora diacronici e sincronici

[81] È noto anche come Corpus of Russian Student Texts (abbreviato CoRST). Per maggiori informazioni si veda il sito: http://web-corpora.net/learner_corpus.

Esistono dei corpora appositamente progettati per condurre ricerche sulla variazione diacronica, intesa come "la variazione lungo l'asse del tempo e nella storia, che dà luogo a diversi stati di lingua successivi" (§ 2.2.4), ossia i corpora diacronici, sincronici ed una particolare categoria, i corpora storici.

La parola "diacronia" è di origine greca, composta dal prefisso διά- (diá-) e dal sostantivo χρόνος (chrónos), e significa letteralmente "attraverso il tempo". I corpora diacronici vengono progettati per indagare il mutamento linguistico di una lingua, di una particolare varietà linguistica, del lessico o di un fenomeno linguistico in un ampio arco temporale.

La parola "sincronia", dal greco σύνχρονος, composta dal prefisso σύν- (syn-) e dal sostantivo χρόνος (chrónos), significa "insieme nel tempo" o "contemporaneamente". I corpora sincronici contengono dati testuali appartenenti a una determinata, e solitamente ristretta, finestra temporale, al fine di condurre indagini linguistiche riguardanti una specifica fase della lingua.

Una particolare categoria è costituita dai corpora storici. Lüdeling e Kytö puntualizzano al riguardo:

> With the passage of time, every corpus will eventually turn into one that can be used for historical study, but strictly speaking a 'historical corpus' is one which is intentionally created to represent and investigate past stages of a language and/or to study language change. (Lüdeling, Kytö, 2008: 242)

Un corpus sincronico, quindi, può essere considerato un corpus storico se viene appositamente creato per condurre indagini linguistiche di tipo storico, focalizzate su un determinato periodo nel passato; al tempo stesso, un corpus sincronico può riguardare la contemporaneità e in questo caso non può essere definito "storico". La medesima considerazione può applicarsi ai corpora diacronici[82].

Esempi di corpora sincronici sono *NKRJa*, così come il già citato *GIKRJa* (§ 2.3.2). Un corpus diacronico, invece, è il *Regensburgskij diachroničeskij korpus russkogo jazyka* o, secondo la denominazione ufficiale, il *Regensburg Russian*

[82] Al riguardo anche *A Glossary of Corpus Linguistics* puntualizza: "While many historical corpora only include texts from a single time period (for example, the Newdigate Letters), one type of historical corpus (called a diachronic corpus) would include texts sampled from different times over a longer period of history. The Helsinki Corpus is an example of this type." (Baker, Hardie, McEnery, 2006: 85). Altri autori ritengono invece che "Synchronic data is contemporary or recent material, used for general, contemporary dictionaries. Diachronic data covers historical data, and may include words and usages that are now considered archaic or have fallen out of use altogether" (O'Keeffe, McCarthy, 2010: 430). In altre parole, i corpora sincronici riguarderebbero solo la lingua contemporanea.

Diachronic Corpus (*RRuDi*)[83], nato da un progetto della Facoltà di Lingue Slave dell'Università di Ratisbona. La nuova versione include oltre 100 mila parole e contiene le opere di Kirill Turovskij[84] e testi come il *Domostroj*[85], *Choždenie Bogorodicy po mukam*[86], *Povest' vremennych let*[87] e *Novgorodskaja pervaja letopis'*[88].

Un altro esempio di corpus diacronico è il già menzionato *NKRJa*, che contiene testi scritti a partire dal XVIII. Inoltre, il sotto-corpus *Istoričeskij korpus* ("Corpus storico") contiene testi scritti tra il XI e il XVIII secolo. Questo sotto-corpus è suddiviso in quattro distinte sezioni: *Drevnerusskij korpus*[89], *Berestjanye gramoty*, *Starorusskij korpus*, *Cerkovslavjanskij korpus*, oppure, secondo la suddivisione fornita da Mitrenina, *Church-Slavonic* (4,7 milioni di parole), *Middle Russian* (3,1 milioni di parole), *Old Russian* (443 mila parole), e *Birchbark corpora* (19 mila parole) (Mitrenina, 2014: 456). Infine, nello stesso sito di *NKRJa* sono indicate numerose pagine web di corpora diacronici e storici russi[90].

2.3.4 Corpora dinamici e statici

I corpora dinamici sono chiamati anche corpora di monitoraggio, dall'inglese *dynamic corpora* e *monitor corpora*[91] (quest'ultima definizione fu coniata da Sinclair nel 1982). La letteratura offre molteplici definizioni al riguardo:

---

[83] Per maggiori informazioni consultare il sito ufficiale dell'Università di Ratisbona: https://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/netzwerke/regensburger-korpora/index.html. Si veda, inoltre, l'articolo di O. Mitrenina relativo al *RRuDi Corpus* (Mitrenina, 2014: 457).

[84] Kirill Turovskij o Cirillo da Turov (1130-1182) fu un vescovo e teologo della Rus' di Kiev e Santo della Chiesa ortodossa russa. La sua festa, secondo il calendario ortodosso, si celebra il 28 Aprile. Per maggiori informazioni si consulti il sito *Akademic*: https://dic.academic.ru/dic.nsf/ruwiki/107271.

[85] Il *Domostroj* è un fondamento della letteratura russa del XVI secolo, comparso alla fine del XV secolo a Novgorod e redatto dall'arciprete Sil'vestr. Contiene un insieme di regole e istruzioni sulla vita mondana, familiare e spirituale dell'epoca. Per consultare l'intera opera si veda: https://azbyka.ru/otechnik/Silvestr/domostroj/.

[86] Letteralmente "Il pellegrinaggio della Vergine Maria attraverso i tormenti". Si tratta di un testo apocrifo, citato da Ivan Karamazov nel romanzo di Dostoevskij *I fratelli Karamazov*. Per consultare il testo integrale si veda: http://drevne-rus-lit.niv.ru/drevne-rus-lit/text/hozhdenie-bogorodicy-po-mukam/hozhdenie-bogorodicy-po-mukam-original.htm.

[87] Nota anche come "Cronaca degli anni passati", attribuita al monaco Nestor di Pečerska. Per consultare l'opera integrale si veda: https://azbyka.ru/otechnik/Nestor_Letopisets/povest-vremennyh-let/.

[88] Nota anche come "Cronaca di Novgorod" o "Prima cronaca di Novgorod".

[89] Per consultare tutti i testi contenuti nella sezione *Drevnerusskij korpus* si veda il link: https://ruscorpora.ru/new/search-old_rus.html.

[90] Per maggiori informazioni consultare il link: https://ruscorpora.ru/new/corpora-other.html alla sezione *Diachroničeskie korpusa russkogo jazyka* e *Staroslavjanskie korpusa*.

[91] Anche in russo si utilizzano le medesime definizioni: *otkrytyj* ('aperto'), *dinamičeskij* ('dinamico'), *monitornyj* o *monitoringovyj* ('di monitoraggio') *korpus*, in contrapposizione a *statičeskij korpus* ('corpus statico').

> A corpus that grows continually, with new texts being added over time so that the dataset continues to represent the most recent state of the language as well as earlier periods. (McEnery, Hardie, 2012: 246)

> A dynamic corpus is one which is continually growing over time, as opposed to a static corpus, which does not change in size once it has been built. Dynamic corpora are useful in that they provide the means to monitor language change over time – for this reason they are sometimes referred to as monitor corpora. (Baker, Hardie, McEnery, 2006: 64)

> Корпус мониторный – постоянно пополняемый и обновляемый корпус текстов, создаваемый в целях мониторинга представляемого корпусом подъязыка или языка в целом[92]. (Zacharov, Bogdanova, 2013: 139)

In altre parole, i corpora di monitoraggio sono costantemente aggiornati mediante l'aggiunta, a intervalli regolari, di dati testuali sempre nuovi ed in tal senso, per citare le parole di Rossini Favretti, "il corpus assume una configurazione dinamica" (Rossini Favretti 2001: 371). Solitamente si tratta di corpora di grandi dimensioni, adatti per condurre analisi diacroniche sul mutamento linguistico. L'aggiunta di nuovi dati implica un lavoro accorto da parte del linguista, al fine di mantenere il corpus bilanciato, rappresentativo e omogeneo, sulla base della tipologia di monitor corpora; inoltre, non bisogna giungere all'errata conclusione che la costante integrazione testuale sia in contrasto con il principio di finitezza (§ 2.2.5), poiché i nuovi dati testuali inseriti sono sempre in un numero finito, in conformità al suddetto principio.

I corpora statici, chiamati anche corpora a campione chiuso o, dall'inglese, *sample*, *snapshot* e *static corpora*[93], sono corpora di grandezza finita non soggetti ad aggiornamento; in virtù della loro natura offrono uno *snapshot*, ossia una "istantanea", di una particolare varietà linguistica, periodo storico, gruppo socio-linguistico o tipologia testuale.

NKRJa è un corpus dinamico, poiché è soggetto ad un periodico aggiornamento dei dati: consultando la sezione *Glavnaja*[94] ('Principale') sarà possibile trovare innumerevoli annunci riguardanti l'ampliamento del corpus principale e dei vari sotto-corpora. Per citare i tre più recenti:

[92] "Monitor corpus: un corpus di testi costantemente ampliato ed aggiornato, creato al fine di monitorare la sotto-lingua o la lingua nel suo insieme rappresentata dal corpus". La traduzione è mia.
[93] McEnery e Hardie utilizzano i termini *sample corpus* e *snapshot corpus* (McEnery, Hardie, 2012: 250). *A Glossary of Corpus Linguistics* parla di *sample corpus* (Baker, Hardie, McEnery, 2006: 141) e *static corpus* (Baker, Hardie, McEnery, 2006: 152). O'Keeffe e McCarthy utilizzano la definizione *sample corpus*, mentre il sostantivo *snapshot*, con il significato di "istantanea", viene utilizzato in riferimento a diverse tipologie di corpora (O'Keeffe, McCarthy, 2010). In Lüdeling e Kytö si trovano tutte e tre le definizioni ed anche il ridondante *static sample corpus* (Lüdeling, Kytö, 2008: 394).
[94] Per consultare tutte le notizie si veda il sito al link: https://ruscorpora.ru/new/.

2 ноября 2021 года: Общий объем корпуса превысил 1 млрд словоформ! […][95].

2 октября 2021 года: Мультимедийный корпус пополнен до 5,4 млн словоупотреблений. […][96].

31 августа 2021 года: Корпус региональной и зарубежной прессы пополнен до 23 млн словоупотреблений. […][97].

Un altro corpus dinamico è il *Dinamičeskij korpus tekstov po sovremennoj publicistike (90-e gody)*, creato presso il Dipartimento di Lessicografia Sperimentale dell'Istituto di Lingua Russa dell'Accademia Russa delle Scienze[98], al fine di analizzare i cambiamenti nel linguaggio dei media e nel discorso politico durante il periodo della *perestrojka* e quello successivo alla caduta dell'Unione Sovietica. Sfortunatamente ad oggi le uniche informazioni su questo corpus sono reperibili nel manuale scolastico di A. N. Baranov, uno degli autori del corpus (Baranov, 2001: 131).

2.3.5 Corpora con fonti scritte, orali e multimediali

Tradizionalmente la variazione diamesica riguarda il canale di trasmissione di un messaggio, che può essere scritto o orale. Recentemente sono stati creati anche corpora multimediali, che hanno reso il panorama della linguistica dei corpora ancora più variegato.

I corpora della lingua scritta sono realizzati mediante fonti esclusivamente scritte (diari, articoli di giornale, lettere, manuali, narrativa, riviste, ecc.), al contrario, i corpora della lingua parlata sono composti da fonti orali trascritte (programmi televisivi e radiofonici, conversazioni, monologhi, conferenze, interviste, lezioni accademiche, discorsi pubblici, ecc.).

In realtà, la distinzione tra queste due tipologie non è netta: molti corpora contengono sia fonti scritte che orali e in virtù di questa caratteristica vengono chiamati corpora misti. Esempi di corpora misti sono i *reference corpora* (§ 2.3.2), che mirano a rappresentare tutte le varietà significative di una lingua, il lessico e le strutture

[95] "2 Novembre 2021: L'estensione totale del corpus ha superato il miliardo di *token*! […]".
[96] "2 Ottobre 2021: il *Corpus multimediale* ha raggiunto i 5,4 milioni di *token*. […]".
[97] "31 Agosto 2021: il *Corpus della stampa regionale ed estera* ha raggiunto i 23 milioni di *token*. […]".
[98] Il progetto è stato finanziato tra il 1996 e il 1997 dal *Rossijskij fond fundamental'nych issledovanij* ("Fondo russo per le ricerche di base"). Per consultare il sito ufficiale e approfondire i vari progetti svolti dal team di ricerca si veda il sito al link: https://www.ruslang.ru/node/251.

morfosintattiche tipiche della lingua, sebbene nella maggior parte dei casi le fonti scritte superino di gran lunga quelle orali.

In generale, i corpora della lingua scritta sono molto più frequenti rispetto a quelli della lingua parlata. Ciò dipende da problemi relativi all'acquisizione e alla trascrizione delle fonti orali: in primo luogo, acquisire dati orali richiede tempi più lunghi e ciò si traduce in costi maggiori. In secondo luogo, come già anticipato (§ 2.2.2), i linguisti devono ottenere il consenso scritto dei parlanti, ma al tempo stesso la consapevolezza di essere registrati può intaccare l'autenticità del dato linguistico. La trascrizione, poi, non solo richiede tempo, risorse tecnologiche ed economiche, ma aumenta anche il potenziale rischio di errore; inoltre, la trascrizione potrebbe intaccare l'autenticità dei dati, in quanto vengono a mancare determinate caratteristiche tipiche del parlato, come il linguaggio non verbale, la gestualità, le espressioni facciali, ecc. Infine, non esiste un sistema di trascrizione unico, sebbene esistano diversi standard di trascrizione.

Questi sono solo alcuni dei motivi che rendono preferibile la creazione dei cosiddetti *written corpora*; tuttavia, anche per quest'ultima tipologia esistono delle oggettive difficoltà di compilazione: basti pensare all'attuale dibattito riguardante il linguaggio del Web, che presenta caratteristiche tipiche sia della lingua scritta, sia del parlato, tanto da valergli le denominazioni di *netspeak, netlish, netlingo, weblish* e *globespeak.* Quella del Web sembra quindi essere una vera e propria lingua a sé. Le caratteristiche peculiari di questa lingua rendono difficile stabilire se si possa parlare propriamente di corpora di lungua scritta piuttosto che di lingua parlata. Di questo tema ci occuperemo nel prossimo capitolo dedicato al *Web as Corpus*.

Una particolare tipologia di corpora di parlato è rappresentata dai cosiddetti corpora audio, contenenti campioni di linguaggio parlato in forma di segnale acustico, talvolta arricchiti dalla trascrizione ortografica.

Infine, i corpora multimediali o multimodali costituiscono una nuova frontiera della linguistica dei corpora: si tratta di corpora contenenti registrazioni audio e video di atti comunicativi, che permettono di condurre studi avanguardistici riguardanti il linguaggio dei segni, la comunicazione verbale e non verbale e lo studio delle emozioni.

Un esempio a riguardo è il *Russkojazyčnyj emocional'nyj korpus* (*REC*), "annotato tenendo conto delle informazioni riguardanti la mimica, i movimenti delle mani, delle sopracciglia, ecc. Esso permette di studiare le strategie dell'interazione e del

conflitto emotivo, il comportamento comunicativo ininterrotto, le esitazioni e le interruzioni del parlato, ecc."[99] (Zacharov, Bogdanova, 2013: 78). Il corpus è suddiviso in due differenti sezioni: la prima include 295 videoregistrazioni di esami universitari orali della durata di 26-60 minuti ciascuno, per un totale di quasi 30 ore; la seconda sezione comprende videoregistrazioni di conversazioni riguardanti il pagamento delle utenze a uno sportello pubblico. Ha differenti livelli di annotazione: della mimica degli occhi e delle labbra, dei movimenti delle mani o della parte del corpo che esegue il movimento, della traiettoria compiuta dal movimento, ecc. (Kotov, Gopkalo, 2011: 212).

Altri corpora multimediali russi sono il *Mul'timedijnyj korpus dialektnych tekstov «Žiznennyj krug»*[100] ('Corpus multimediale di testi dialettali «Cerchio della vita»'), il *Mul'timedijnyj Saratovskij dialektologičeskij tekstovyj korpus* (abbraviato *SDK*)[101] ('Corpus testuale multimediale dialettologico di Saratov'), il *Korpus russkogo žestovogo jazyka*[102] ('Corpus della lingua dei segni russa') ed infine il *Mul'timedijnyj russkij korpus* (*MURKO*)[103] ('Corpus multimediale russo'), un sotto-corpus di *NKRJa* costituito da frammenti di film dagli anni '30 agli anni 2000, con relative trascrizioni del testo. Oltre alla tradizionale annotazione morfologica, semantica e metalinguistica, esso presenta anche un'annotazione gestuale, ortoepica[104] e quella riguardante l'accento e la tipologia di atto linguistico.

Tra i corpora orali della lingua russa occorre menzionare il *Corpus of Russian Spontaneous Speech* (*CoRuSS*), che contiene 30 ore di registrazione di russo spontaneo, parlato da 60 madrelingua tra i 16 e i 77 anni. Il corpus è arricchito da annotazione

[99] "Русскоязычный эмоциональный корпус (*REC*), размеченный с учетом данных о мимике, движениях рук, бровей и т. д., позволяет изучить стратегии эмоционального взаимодействия и конфликта, непрерывное коммуникативное поведение, хезитации и речевые сбои и др.". Per maggiori informazioni si veda anche il contributo di A. A. Kotov e O. C. Gopkalo riguardante il *REC* (Kotov, Gopkalo, 2011).

[100] Chiamato anche *Mul'timedijnyj korpus regional'nich tekstov «Žiznennyj krug»*. Per maggiori informazioni si veda (Zacharov, 2014: 7), (Dračeva, Zadumina, 2014) e (Dračeva, Zubova, 2015).

[101] Citato anche in (Dračeva, Zadumina, 2014: 45), per maggiori informazioni si veda il sito ufficiale dell'Università di Saratov: https://www.sgu.ru/structure/philological/narrech.

[102] Il progetto è stato realizzato tra il 2012 e il 2014. Comprende più di 230 video-testi, tradotti ed annotati, prodotti da 43 persone che parlano la lingua dei segni russa. Per maggiori informazioni consultare il sito al link: http://rsl.nstu.ru/site/project.

[103] Citato in Zacharov, Bogdanova (2013: 77) e Zacharov (2013: 12). Si vedano anche i contributi di E. A. Grišina (2011, 2015), che con il suo lavoro ha permesso la nascita e lo sviluppo del *MURKO*.

[104] Per annotazione ortoèpica s'intende quel tipo di annotazione che riguarda la corretta pronuncia delle parole. Per maggiori informazioni circa il significato di questo aggettivo si veda al link: https://www.treccani.it/vocabolario/ortoepico/.

ortografica, prosodica e fonetica, nonché dalla segnalazione delle esitazioni e degli eventi non-linguistici (Kačkovskaja, Kočarov, Skrelin, Vol'skaja, 2016: 1951). Un altro corpus orale è il *Korpus ustnoj reči* ('Corpus del linguaggio orale'), un sotto-corpus di *NKRJa* che contiene 13,4 milioni di parole[105], per un arco temporale che va dagli anni '30 del secolo scorso ad oggi. Infine, per questioni di spazio si citano soltanto, senza il dovuto approfondimento, il *SibLing Corpus of Russian Dialogue Speech*[106] (Kačkovskaja *et alii*, 2020) e *Odin rečevoj den'*[107] ('Un giorno di lingua parlata').

2.3.6 Corpora monolingui e multilingui

Un corpus monolingue, come suggerisce l'aggettivo stesso, contiene dati testuali in una sola lingua, come il *Manuscript Corpus*[108], il *Sankt-Peterburgskij korpus agiografičeskich tekstov*[109] ('Corpus di testi agiografici di San Pietroburgo'), i già citati *Regensburgskij diachroničeskij korpus russkogo jazyka* (§ 2.3.3), il *Korpus russkogo žestovogo jazyka* e il *Corpus of Russian Spontaneous Speech* (§ 2.3.5).

Un corpus multilingue, invece, contiene testi in due o più lingue. Di questa tipologia si distinguono i corpora paralleli e comparabili.

Un corpus parallelo contiene dati testuali originali nella cosiddetta *Source Language* (SL) e la loro traduzione in una o più lingue, definite *Target Language* (TL) (Olohan 2004: 24-25, citato da Gandin, 2005: 134).

---

[105] Secondo i dati di *NKRJa* aggiornati al 23 Agosto 2021. Si veda al link: https://ruscorpora.ru/new/corpora-stat.html. Inoltre, questa notizia è riportata anche nella sezione *Glavnaja* ("Principale"), dell'1 giugno 2021: https://ruscorpora.ru/new/.

[106] Si tratta di un corpus contenente 90 dialoghi di 10 coppie di fratelli e sorelle dello stesso sesso e di età compresa tra i 23 e 40 anni, incluse 4 coppie di gemelli monozigoti. Questi parlanti sono stati registrati nel corso di alcune conversazioni con differenti tipi di interlocutori: il proprio fratello o sorella; un amico stretto; una persona sconosciuta, coetanea e dello stesso genere; una persona sconosciuta, coetanea, ma di genere differente; una persona sconosciuta, dello stesso genere, più grande d'età e con una più alta posizione lavorativa. Lo studio si focalizza sul cambiamento del tono di voce in base alle differenti situazioni sociali e al grado di familiarità con l'interlocutore (Kačkovskaja *et alii*, 2020).

[107] *ORD* è un progetto nato nel 2007 grazie all'Istituto di Ricerche Filologiche dell'Università statale di San Pietroburgo, con l'obiettivo di raccogliere registrazioni di parlato reale e quotidiano. Per maggiori informazioni si veda (Zacharov, 2013: 13), (Zacharov, 2014: 7) e (Kačkovskaja, Kočarov, Skrelin, Vol'skaja, 2016: 1949).

[108] Si tratta di un corpus realizzato dal Laboratorio di Ricerca Filologica Computerizzata presso l'Università Statale dell'Udmurtia (Federazione Russa), contenente manoscritti in slavo ecclesiastico (inclusi alcuni documenti scritti in glagolitico) e russo medioevale e testi di Lomonosov. Per consultare il sito ufficiale: http://mns.udsu.ru/.

[109] Il corpus, creato dal Dipartimento di Linguistica Matematica della Facoltà di Filologia dell'Università Statale di San Pietroburgo, raccoglie testi agiografici dei secoli XV-XVII. Secondo i dati riportati nel sito, attualmente comprende 50 manoscritti, per un totale di 500 mila *token*. Per consultare il sito ufficiale dello *SKAT*: http://project.phil.spbu.ru/scat/page.php?page=project.

S. Gandin distingue quattro differenti modelli di corpora paralleli, che dimostrano quale raffinato lavoro di progettazione si celi dietro questa particolare tipologia[110].

- Il modello uni-direzionale prevede dati testuali in una *SL* e la loro traduzione in una *TL*, in un rapporto, per così dire, *one-to-one*.
- Il modello bi-direzionale prevede dati testuali in due specifiche *SL* e la traduzione combinata nelle suddette due lingue.
- Il modello a stella è costituito da testi in una sola *SL*, ma traduzioni in due o più *TL*.
- Il modello a diamante contiene testi in tre o più *SL* e le rispettive traduzioni combinate.

Nel condurre indagini linguistiche mediante corpora paralleli, può risultare molto utile l'allineamento[111] del testo originale e della corrispettiva traduzione, al fine di avere una costante corrispondenza tra i due testi a differenti livelli della lingua (corrispondenza di parole, frasi, periodi, paragrafi, capitoli, ecc.).

Come nel caso dell'annotazione, anche per l'allineamento testuale esistono vari software e programmi, che suddividono e distribuiscono automaticamente il testo. Tuttavia, nonostante il lavoro venga svolto dal computer, esiste una percentuale di errore e per questo motivo l'allineamento automatico è spesso seguito dal controllo manuale da parte del linguista.

I corpora paralleli sono una risorsa indispensabile per la creazione di dizionari bilingui e software di traduzione automatica nell'ambito del *Natural Language Processing* (*NLP*). Sono anche un ottimo strumento per l'apprendimento di una lingua e per le traduzioni di linguaggi altamente specialistici, come quello del diritto internazionale e delle istituzioni e organizzazioni politiche.

Per quanto riguarda i corpora comparabili, si riportano differenti definizioni:

> In the prototypical parallel corpus […] the relationship lies in shared meaning. By contrast, what links the collections of texts in comparable corpora is that they have been put together according to the same type of criteria (texts of a certain size, on a set topic, from a given period, etc.). (O'Keeffe, McCarthy, 2010: 487)

> Another type of parallel corpus (sometimes called a 'comparable corpus') consists of different texts in each language: it is merely the sampling method that is the same. (Baker, Hardie, McEnery, 2006: 126)

[110] Le seguenti informazioni sono una rielaborazione di Gandin, 2005: 134 che a sua volta si rifà a Johansson 2003: 138-140.
[111] Per completezza, il termine tecnico utilizzato in russo per *allineamento* è *vyravnivanie*.

> Parallel corpus is a collection of text pairs consisting of source language text and translated target language text and there is a strict translation relationship between the two languages. Comparable corpus is a non-translated text pair collection with different language but similar content. (Zong, Hong, 2019: 1)

In altre parole, ciò che accomuna i corpora comparabili non è la traduzione da una lingua all'altra, ma le modalità di campionamento affini.

Esempi di corpora paralleli sono quelli contenuti in *NKRJa*, che attualmente comprendono le seguenti lingue: inglese, armeno, baschiro[112], bielorusso, bulgaro, buriato[113], ceco, cinese, italiano, estone, finlandese, francese, lettone, lituano, polacco, spagnolo, svedese, tedesco, ucraino ed un corpus multilingue, per un totale di 140 milioni di parole. Esiste anche il *Mul'timedijnyj parallel'nyj korpus*[114] nella versione inglese-russo. Infine, un web-corpus parallelo è l'*OPUS2*, consultabile tramite la piattaforma Sketch Engine, che contiene traduzioni allineate in circa 40 lingue, tra le quali anche il russo[115].

2.3.7 I learner corpora

Il learner corpus è una speciale tipologia di corpus linguistico, che presenta tutte le caratteristiche tradizionali dei corpora, ma a differenza di questi ultimi i dati testuali, orali e scritti, non sono prodotti da madrelingua, ma da apprendenti di una lingua. La maggior parte presenta vari livelli di annotazione e, soprattutto, l'annotazione dell'errore.

Questo approccio si è sviluppato tra la fine degli anni '80 e l'inizio degli anni '90 del secolo scorso, al fine di condurre degli studi riguardanti l'acquisizione di una seconda lingua e i fattori che influenzano il processo di acquisizione (Lüdeling, Kytö, 2008: 259).

Lüdeling e Kytö distinguono i learner corpora in *commercial learner corpora* ed *academic learner corpora*. I commercial learner corpora sono creati da grandi case

[112] La lingua baschira o baškira è parlata dalla popolazione che vive nella Repubblica di Baschiria.
[113] Si tratta di un dialetto mongolo parlato da una etnia mongola della Siberia, stanziata nella regione intorno al lago Bajkal.
[114] Abbreviato *Mul'tipark*, consultabile al link: https://ruscorpora.ru/new/search-multiparc_rus.html.
[115] Per maggiori informazioni a riguardo si veda al link: https://www.sketchengine.eu/opus-parallel-corpora/#toggle-id-1.

editrici e solitamente sono più ampi e più variegati rispetto agli academic learner corpora. Un esempio di commercial learner corpus è *Longman Learners' Corpus*[116].

Gli academic learner corpora, invece, sono più numerosi rispetto alla prima tipologia e vengono creati in ambito accademico; un esempio al riguardo è l'*International Corpus of Learner English* (*ICLE*)[117] (Lüdeling, Kytö, 2008: 261).

Le applicazioni d'uso sono molteplici: questi corpora possono essere utilizzati dagli studenti per comprendere l'errore linguistico ed evitarlo in futuro o per acquisire maggiore consapevolezza riguardo il proprio livello linguistico; sono utili per gli insegnati che vogliono creare lezioni mirate, basandosi sugli errori più frequenti da parte degli apprendenti; infine, sono utili per i ricercatori, per condurre indagini relative ai meccanismi di apprendimento di una lingua.

Per quanto riguarda la lingua russa ricordiamo il già citato *Korpus russkich učebnich tekstov* (§ 2.3.2) e l'*Učebnyj Mul'timodal'nyj Korpus* ('Learner corpus multimodale'), che contiene 28 brevi videoregistrazioni di dialoghi spontanei prodotti da russofoni e apprendenti di lingua russa di nazionalità cinese e tedesca[118].

Il *Saint Petersburg EFL Learner Corpus* (*SPbEFL LC*) è un learner corpus relativamente piccolo, contenente testi orali e scritti di varia natura (saggi, lettere personali, monologhi e dialoghi). Il campione comprendeva 90 studenti delle scuole superiori di San Pietroburgo (Federazione Russa) e 12 loro coetanei immigrati con un livello di russo intermedio o avanzato[119].

Infine, il *Russkij Učebnyj Korpus*[120], noto anche come *Russian Learner Corpus* (*RLC*), raccoglie dati orali e scritti di due particolari categorie di parlanti russi: coloro

[116] Per maggiori informazioni si veda il link: http://global.longmandictionaries.com/longman/corpus#aa.
[117] Nel 2009 è nata la seconda versione aggiornata del corpus, chiamata *ICLEv2*, contenente 3,7 milioni di parole scritte da studenti *EFL* (dall'inglese "English as a Foreign Language", da non confondere con "English as Lingua Franca", abbreviato *ELF*) madrelingua di 16 differenti nazionalità (Bulgaria, Cina, Repubblica Ceca, Finlandia, Francia, Germania, Italia, Giappone, Russia, ecc.). La terza versione, l'*ICLEv3*, presentata nel 2020, contiene 5,7 milioni di parole e presenta dati testuali prodotti da madrelingua di ben 25 differenti nazionalità (26 secondo il sito ufficiale). A differenza della seconda versione, distribuita commercialmente in formato CD-ROM (al sito: https://uclouvain.be/en/research-institutes/ilc/cecl/iclev2.html), è possibile consultare online l'*ICLEv3* come versione di prova al link: https://corpora.uclouvain.be/cecl/icle/trial/. Per maggiori informazioni si consulti il manuale dell'*ICLEv3* (Granger, Dupont, Meunier, Naets, Paquot, 2020: i).
[118] Si veda il link https://studbooks.net/2148397/literatura/uchebnyy_multimodalnyy_korpus per maggiori informazioni.
[119] Per maggiori informazioni si veda Kamšilova, 2017.
[120] Per maggiori informazioni si veda: http://web-corpora.net/RLC. Inoltre, per consultare i numerosi progetti della *Škola lingvistiki* ('Scuola di linguistica') della *Vysšaja Škola Ekonomiki* si veda: https://ling.hse.ru/resources.

che studiano il russo come lingua straniera e i cosiddetti *eritažnye govorjaščie*[121]. Questa importante risorsa è stata realizzata dalla *Škola lingvistiki* ('Scuola di linguistica') presso la *Vysšaja Škola Ekonomiki*, grazie all'operato della prof.ssa E. Rakhilina e ad oggi è un corpus fondamentale per gli studi di linguistica acquisizionale e di metodologia della didattica.

2.4 Conclusioni del capitolo

Nel corso di questo capitolo sono state fornite diverse definizioni di corpus linguistico. Successivamente sono state analizzate le caratteristiche dei corpora: il formato elettronico, l'autenticità e rappresentatività del dato linguistico, il bilanciamento, le dimensioni, la finitezza, l'omogeneità, la ripetibilità e riproducibilità. Infine, si è passati alla rassegna delle varie tipologie di corpora, riportando numerosi esempi di corpora russi. Per una lista esaustiva dei corpora russi o di lingua russa, si veda la sezione *I corpora della lingua russa* a pagina 143 dell'Appendice.

Il prossimo capitolo riguarderà la nuova frontiera della linguistica dei corpora, il *Web as Corpus*. Verranno illustrate le modalità di creazione dei web corpora, si affronterà il dibattito riguardante il *netspeak*, la lingua di Internet, e le questioni relative al copyright. Infine, si concluderà con un confronto tra i corpora tradizionali e i web corpora.

[121] Con l'espressione "parlante ereditario" (*heritage speaker*) si fa riferiamo a una persona che ha imparato una lingua in modo informale, mediante esposizione, non formalmente come in un ambiente scolastico.

3. Il *Web as Corpus* e i web corpora

3.1 Il *Web as Corpus*: la nascita di un nuovo approccio alla disciplina

Nei precedenti capitoli sono stati illustrati i progressi compiuti dalla linguistica dei corpora, grazie all'invenzione dei computer ed alla successiva e graduale informatizzazione dei dati linguistici.

Dai primi esperimenti compiuti da Padre Roberto Busa negli anni '50 si è passati, nel corso di un decennio, alla nascita del primo corpus in formato elettronico, il *Brown Corpus of American Written English* di W. N. Francis e H. Kučera, e dai cosiddetti corpora di prima generazione degli anni '60 e '70, contenenti 1 milione o più di parole, si è passati a quelli di seconda generazione, contenenti 100 milioni o più di parole.

Nel 1992 Michael Rundall pubblicò tre articoli nella rivista *English Today*, intitolati *The corpus revolution*[122], in cui sottolineava che l'avvento dei computer aveva portato alla prima *corpus revolution*, nonostante all'epoca la creazione di corpora mediante computer richiedesse lungo tempo e fosse molto costosa. Nel 2008 lo stesso Rundall pubblicò un quarto articolo a riguardo, intitolato *The corpus revolution revisited*, in cui confermava quanto detto un decennio prima e aggiungeva che "It is fair to say, then, that the arrival of the Web has sparked a second Corpus Revolution" (Rundall, 2008: 26). La diffusione di Internet ha reso accessibili informazioni digitali globali determinando, a mio parere, non solo una seconda *corpus revolution*, ma più in generale una *cultural revolution*.

Nel 2003 Adam Kilgarriff e Gregory Grefenstette pubblicarono un articolo dal titolo provocatorio *Introduction to the Special Issue on the Web as Corpus*. In questo articolo si intendeva dimostrare che il Web può essere considerato un corpus linguistico e in quanto tale può essere interrogato al fine di condurre indagini linguistiche; i due linguisti concludevano l'articolo con la famosa frase "Our take on the Web is that it is a

[122] Si fa riferimento a Rundell M. (1992), "The corpus revolution", *English Today*, Vol. 8: 2 (1992), pp. 9-14 al link DOI: https://doi.org/10.1017/S026607840000626X. Il secondo articolo è Rundell M. (1992), "The corpus revolution", *English Today*, Vol. 8: 3 (1992), pp. 21-32 al link DOI: https://doi.org/10.1017/S0266078400006520. Il terzo è Rundell M. (1992), "The corpus revolution", *English Today*, Vol. 8: 4 (1992), pp. 45-51 al link DOI: https://doi.org/10.1017/S0266078400006751. Sfortunatamente questi tre articoli non sono pubblicamente consultabili, pertanto le informazioni bibliografiche a riguardo sono inserite in questa nota e non in bibliografia. Esiste tuttavia un quarto articolo, intitolato *The corpus revolution revisited*, utilizzato come fonte per questo capitolo e indicato in bibliografia, nonché pubblicamente consultabile tramite internet: per maggiori informazioni si veda la bibliografia alla voce "Rundell M. (2008)".

fabulous linguists' playground. We hope the special issue will encourage you to come on out and play!" (Kilgarriff, Grefenstette, 2003: 345). Non solo il loro invito è stato ampiamente accolto, ma l'espressione volutamente provocatoria "Web as Corpus" è diventata una vera e propria formula della linguistica dei corpora, tanto da essere presente in numerosi autori come Volk (2002), Baroni e Bernardini (2006), Lüdeling e Baroni (2007), Ferraresi (2009), Lew (2009), Gatto (2009, 2014), tanto per citarne alcuni. Tutti questi contributi contengono nel loro titolo e trattano del "Web as Corpus".

L'intuizione di Kilgarriff e Grefenstette, inoltre, spinse alcuni linguisti a pensare ad un nuovo nome per indicare questa branca della linguistica dei corpora, che andava delineandosi all'inizio degli anni 2000: infatti, nel 2004 il linguista britannico David Crystal pubblicò il manuale *Language and the Internet* (Crystal, 2004), in cui introdusse per la prima volta la definizione di "Internet Linguistics"[123], formulazione ripresa l'anno successivo in *The scope of Internet linguistics*[124]. Nel 2005, il linguista svedese Gunnar Bergh propose la variante "Web Linguistics"[125], utilizzata in seguito da Lüdeling e Kytö (Lüdeling, Kytö, 2008: 309).

In altre parole, la *Web Linguistics* o *Internet Linguistics* è una branca della linguistica dei corpora, che si avvale del World Wide Web per condurre ricerche di tipo linguistico.

La Web Linguistics ha differenti approcci, indicati dagli autori in numero variabile in base al loro pensiero. Hundt, Nesselhauf e Biewer indicano due differenti approcci, il *Web for Corpus* e il *Web as Corpus*. Per citarli:

> (a) With the help of commercial crawlers or internet-based search engines such as WebCorp, the web can be used as a corpus itself ('Web as corpus') – as a heuristic tool but also in a more systematic way. The heuristic use could be referred to as 'data sniffing', the systematic application as 'data testing'.
> (b) The www can alternatively be used as a source for the compilation of large offline monitor corpora ('Web for corpus building'). (Hundt, Nesselhauf, Biewer, 2007: 2)

Anche Lüdeling e Kytö propongono la medesima suddivisione:

[123] La formula "Internet Linguistics" viene utilizzata 5 volte in tutto, mentre "Corpus Linguistics" solo 2 volte.
[124] Si tratta di un contributo esclusivamente online di David Crystal per l'American Association for the Advancement of Science meeting del febbraio 2005, consultabile nel sito ufficiale di D. Crystal al link https://www.davidcrystal.com/GBR/Books-and-Articles?itemId=807.
[125] Cito la sua metafora culinaria: "Adding a bit of culinary zest to the present paper, finally, we may sum up this outing into Web linguistics, with its particular focus on large-scale language collections and state-of-the-art search technology, by a short concluding statement which seems to capture the essence of the discussion: the Web is best enjoyed in carefully cut slices, preferably based on the raw capacity of Google and spiced according to taste with the fine-tuned linguistic facilities of Web-Corp". (Bergh, 2005: 45)

> In Web linguistics there are two main approaches to the exploitation of online data. One is referred to as Web for Corpus (WfC), or Corpora from the Web, and is concerned with corpus compilation of textual data from the Web; the other is known as Web as Corpus (WaC) and involves direct utilization of the Web as a corpus (e. g. de Schryver 2002). It is worth noting, however, that very often the term WaC is used to indicate all aspects of Web linguistics, even those that, according to the definition given above, should go under the label of WfC. (Lüdeling, Kytö, 2008: 315)

Barbera, Corino[126] e Onesti, invece, indicano tre differenti approcci:

> (1) il materiale del web reso corpus in un determinato taglio temporale, considerando le informazioni di un insieme molto ampio di testi ma comunque finito e stabile; (2) l'idea di elaborare le informazioni su materiale 'aperto', sulla rete in continuo movimento, non creando un vero e proprio corpus ma applicando ai dati tools di estrazione e crawling; (3) un ibrido delle due precedenti […] paragonabile ad una collezione di monitor corpora molto ravvicinati. (Barbera, Corino, Onesti, 2007: 44)

Infine, Baroni e Bernardini indicano quattro differenti approcci, citati successivamente anche da Gatto, che per questioni di spazio riassumeremo qui brevemente[127]: il primo è il *Web as a corpus surrogate*[128], in cui il Web viene utilizzato come fosse il surrogato di un corpus linguistico, interrogabile tramite un motore di ricerca commerciale; nel secondo approccio, il *Web as a corpus shop*, il Web è uno strumento mediante cui è possibile selezionare e scaricare i testi desiderati, al fine di creare un corpus linguistico tradizionale, in altre parole è l'approccio *Web for Corpus* di cui parleremo nel dettaglio in questo capitolo (§ 3.3). Il terzo approccio, chiamato *Web as corpus proper*, utilizza il Web come un vero e proprio corpus, al fine di indagare lo stato di una lingua o varietà linguistica nel Web, in uno specifico arco temporale; questo approccio altri non è che il già citato *Web as Corpus*. Infine, l'ultimo approccio, il cosiddetto M*ega-corpus/mini-Web*, è il più radicale, quello che Barbera, Corino e

[126] Nel suo articolo *Didattica delle lingue corpus-based*, Corino sottolinea, rifacendosi a Barbera (2013), l'esistenza di due principali approcci: "(1) il materiale del web reso corpus in un determinato taglio temporale, considerando le informazioni di un insieme molto ampio di testi ma comunque finito e stabile; (2) l'idea di elaborare le informazioni su materiale 'aperto', sulla rete in continuo movimento, non creando un vero e proprio corpus ma applicando ai dati tools di estrazione e crawling" (Corino, 2014: 234).

[127] Per una lettura più approfondita della questione si rimanda a Baroni, Bernardini (2006: 10) e Gatto (2009: 4; 2014: 37).

[128] Secondo quanto riporta A. Ferraresi, molti studenti e traduttori professionisti utilizzano Internet per compiere ricerche di tipo linguistico, in quanto "using search engines relieves translators from the task of learning to use new software programs, such as concordancers, and, depending on the task, may provide enough information and thus make the download of reference materials superfluous" (Ferraresi, 2009: 1). Infatti, secondo dei dati statistici riportati dallo stesso Ferraresi, nel 2007 il 95% dei traduttori professionisti e studenti intervistati affermavano di utilizzare Internet per i loro lavori di traduzione (ad esempio, per ricerche di tipo terminologico), mentre meno della metà ricorreva all'uso o alla compilazione di corpora linguistici.

Onesti indicano come terzo approccio, quello ibrido: si tratta di creare dei *mini-Web* o *mega-corpora* come strumenti che coniugano le caratteristiche tipiche del Web, ovvero le grandi dimensioni e il costante aggiornamento dei dati, e quelle tipiche dei corpora tradizionali, come l'annotazione o la possibilità di effettuare interrogazioni avanzate (Baroni, Bernardini, 2006).

Tra questi approcci il più controverso è il *Web as Corpus*: attualmente è in corso un acceso dibattito tra i sostenitori del *Web as Corpus* (in seguito abbreviato in *WaC*) e i sostenitori dell'approccio opposto, il *Web for Corpus* (d'ora in avanti *WfC*).

Nel prossimo paragrafo saranno illustrate nel dettaglio le principali critiche mosse nei confronti dell'approccio *WaC* e le posizioni dei vari autori a riguardo.

3.2 Il *Web as Corpus*?

Che cos'è un corpus linguistico? Quali sono le caratteristiche che rendono "linguistico" un corpus? E in virtù della sua natura, il Web può essere considerato un corpus linguistico?

Nel secondo capitolo è stata fornita una definizione precisa di corpus linguistico, inteso come "una raccolta di dati testuali autentici in formato elettronico, provenienti da fonti scritte, orali o multimediali, rappresentativi di una lingua o una varietà di essa, al fine di condurre indagini linguistiche empiriche ripetibili, sia qualitative che quantitative" (§ 2.1). È proprio a partire dalla definizione di ciò che è un corpus linguistico e delle caratteristiche che rendono "linguistico" un corpus, che gli studiosi hanno mosso le loro critiche nei confronti dell'approccio *WaC*, in relazione all'autenticità e l'autorevolezza dei dati, le dimensioni, la finitezza, la rappresentatività e il bilanciamento del Web, nonché la ripetibilità e la riproducibilità dei risultati della ricerca.

3.2.1 L'autenticità e l'autorevolezza del Web

Come è stato già osservato nel precedente capitolo (§ 2.2.2), si definiscono autentici i dati linguistici genuini, frutto dell'interazione umana scritta, orale o multimediale, prodotti a scopo comunicativo e non manipolati dal linguista, elicitati al fine di condurre un'analisi linguistica e per dimostrare la competenza linguistica dei parlanti.

In questo senso i dati linguistici presenti nel Web possono essere considerati generalmente autentici; per citare le parole di M. Gatto, sostenitrice dell'approccio *WaC*, "authenticity is the most obvious strength in the similarity between a corpus and the web" (Gatto, 2014: 43). Anche in Hundt, Nesselhauf e Biewer, nel paragrafo intitolato "Why use the web as corpus?", al primo posto è indicato "Freshness and spontaneity: the content of compiled corpora ages quickly, but texts on contemporary issues and authentic examples of current, nonstandard, or emerging language usage thrive online" (Hundt, Nesselhauf, Biewer, 2007: 27).

Tuttavia, molti linguisti obiettano al riguardo, chiamando in causa l'inaffidabilità dei dati linguistici presenti nel Web, legata all'impossibilità di stabilire con certezza la reale identità degli utenti. Infatti, riportando quanto sostenuto da Fletcher, "Web pages are typically anonymous and Web server location is no certain guide to origin, so it is difficult to establish authorship and provenance and to assess the reliability, representativeness and authorativeness of texts" (Fletcher, 2004: 2).

In altre parole, la critica nei confronti dell'approccio *WaC* riguarda l'esistenza di numerosi profili fittizi o *fake*[129], nei quali vengono falsificate le informazioni personali relative all'identità, professione, età, genere, istruzione, nazionalità, ecc.; in generale, la non-autenticità del campione linguistico oggetto d'esame inficia inevitabilmente l'attendibilità dei risultati.

Secondo Gatto, invece, il problema dell'autenticità dei dati linguistici nel Web andrebbe individuato nella mancanza di *authoritativeness*[130], o autorevolezza, degli utenti, legata alla presenza di numerosi errori ortografici e grammaticali, dell'uso improprio di alcune parole o espressioni, del linguaggio frammentario, ripetitivo, ovvero di tutto ciò che nella Web Linguistics prende il nome di "*noise*". Kilgarriff e Grefenstette difendono l'autenticità ed autorevolezza del Web sostenendo che "the Web is a dirty corpus, but expected usage is much more frequent than what might be considered noise" (Kilgarriff, Grefenstette, 2003: 342).

[129] Nel sito online del Commissariato di Pubblica Sicurezza è indicata l'attuale normativa riguardante la falsificazione dei dati personali, la creazione dei profili *fake* e molto altro: https://www.commissariatodips.it/approfondimenti/social-network/approfondimenti-normativi/index.html.

[130] Gatto parla della questione in questi termini: "authenticity in the web is in fact often related to problems of 'authoritativeness'. Everyday experience suggests that 'authentic' in the web often means inaccurate (misspelt words, grammar mistakes, improper usage by non-native speakers), i.e. texts are almost certainly authentic in terms of their communicative intent but mat be unreliable and lacking authority at the level of code" (Gatto, 2014: 43).

Chiari invece obietta che "la presenza massiccia di pagine amatoriali, spesso scritte da utenti che non governano bene la lingua […], finisce per costituire una rappresentazione molto sbilanciata delle caratteristiche linguistiche dei testi" (Chiari, 2007: 55).

Quindi, da una parte la mancanza di autorevolezza inficia, seppur in minima parte, la qualità e l'autenticità dei dati linguistici, dall'altra non si può ignorare che anche le pagine amatoriali, così come quelle scritte dagli utenti che non governano bene una lingua straniera, costituiscono una realtà linguistica, presente nel Web così come nella realtà. L'unico modo per far fronte a queste problematiche è selezionare accuratamente i testi ed effettuare su di essi specifiche operazioni di *cleaning* e *filtering*, mediante dei software appositamente creati, dei quali si parlerà a breve (§ 3.3). Solo così il linguista potrà operare su un campione autentico e al tempo stesso autorevole.

3.2.2 Le dimensioni e la finitezza del Web

La dimensione e la finitezza sono due caratteristiche fondamentali, che influenzano non solo il bilanciamento del corpus linguistico, ma anche la ripetibilità e riproducibilità dei risultati della ricerca. Infatti, una delle motivazioni più solide su cui si basa la critica del *WaC* riguarda proprio le dimensioni del Web ed il suo costante mutamento.

Quanto è grande il World Wide Web? Molti autori negli anni, sin dalla nascita di questo approccio alla linguistica, hanno cercato di rispondere a questa domanda; primi fra tutti Kilgarriff e Grefenstette, che nel loro articolo riportavano: "Lawrence and Giles […] estimated that, in 1999, there were 800 million indexable Web pages available" (Kilgarriff, Grefenstette, 2003: 337).

In seguito, R. Lew tentò di stimare approssimativamente le dimensioni di Google, utilizzando i dati forniti dal motore di ricerca, relativi all'agosto 2008 (circa 8 miliardi di pagine, stimate per difetto), e utilizzando l'algoritmo proposto 10 anni prima dagli stessi Lawrence e Giles. Sulla base di questi calcoli concludeva: "puts a rough estimate of the total (indexed and unindexed) textual resources at five trillion (5,000,000,000,000) word tokens" (Lew, 2009: 4).

Bisogna ricordare che la stima di Lew si riferiva al solo motore di ricerca Google e non al World Wide Web in generale. Alcuni autori, come Fletcher, sostenevano che il

World Wide Web contenesse "over ten billion publicly-accessible online documents provide comprehensive coverage of the major languages and language varieties, and span virtually all content domains and written text types" (Fletcher, 2004: 1). In quello stesso anno il linguista svedese G. Bergh scrisse un articolo, che sarebbe stato pubblicato solo l'anno successivo, in cui proponeva cifre differenti: "As regards the size of the material on the Web, a rough estimate indicates that there are currently (December 2004) about eight billion Web pages available" (Bergh, 2005: 25).

Infine, Gatto nel suo manuale ricorda la video-lezione *Theorizing from data*, tenuta nel 2008 dal direttore di ricerca di Google, Peter Norvig, che affermava "the size of the internet amounts to 100 trillion words" (Gatto, 2014: 46).

È indubbio che il Web sia caratterizzato da dinamicità ed anarchia incontrollate: ogni giorno vengono create nuove pagine web e al tempo stesso altre vengono eliminate. Inoltre, accanto alle cosiddette *indexed pages*, ossia quelle pagine web mappate dai motori di ricerca, esiste il cosiddetto Web sommerso, costituito dal *deep Web* e dal *dark Web*, di cui non si hanno informazioni precise riguardo le dimensioni.

La dinamicità del Web implica inevitabilmente mancanza di finitezza e questo non è un dettaglio trascurabile: come osservato nel precedente capitolo, "poiché mediante i corpora vengono eseguite delle analisi di tipo statistico, è necessario che l'insieme di dati linguistici su cui si opera, detto campione, sia finito" (§ 2.2.5). La finitezza garantisce non solo di operare entro certi confini, ma anche un certo grado di bilanciamento e soprattutto assicura la replicabilità e riproducibilità dei risultati di ricerca, di cui parleremo a breve (§ 3.2.4).

Quali che siano le dimensioni del Web, le posizioni degli autori al riguardo sono piuttosto omogenee: Chiari afferma che "la dimensione del web non solamente è indeterminata, è anche in certo modo indeterminabile. Da questa prospettiva non si tratta né di un corpus statico, né di un corpus dinamico controllato" (Chiari, 2007: 56).

Anche Hund, Nesselhauf e Biewer sostengono che il principale problema dell'approccio *WaC* è che "we still know very little about the size of this 'corpus', the text types it contains, the quality of the material included or the amount of repetitive 'junk' that it 'samples'. Furthermore […] replicability of the results is impossible" (Hundt, Nesselhauf, Biewer, 2007: 3). Dello stesso avviso è Barbera, il quale afferma che l'approccio *WaC* "si scontra però con il problema della finitezza: il WWW è sempre

in movimento, non si può considerare né definito (almeno non nel senso di consentire la ripetibilità degli esperimenti) né finito" (Barbera, 2013: 21).

Gatto, pur essendo una sostenitrice del *WaC,* ammette che "the problem of the web's non-finiteness brings with uncertainties and doubts concerning its value in a research field, like corpus linguistics, which is based on the investigation of quantitative data (Gatto, 2014: 45).

Una voce fuori dal coro è R. Lew, secondo cui le grandi dimensioni del Web permettono di studiare a livello statistico non solo le parole poco frequenti, ma anche quelle collocazioni che difficilmente comparirebbero in un corpus tradizionale o in un dizionario, e a questo proposito conclude: "this quantitative aspect appears then to score a point in favour of the World Wide Web, when seen in opposition to traditional corpora" (Lew, 2009: 7).

Dopo questa rassegna è possibile affermare che la critica riguardante le dimensioni e la mancanza di finitezza del Web è una delle più solide rispetto all'approccio *WaC* e trova un generale favore unanime. Una soluzione al riguardo è rappresentata dall'approccio *WfC*, che prevede la selezione finita di campioni testuali tratti dal Web, garantendo non solo il principio di finitezza, ma anche quello di bilanciamento, rappresentatività, replicabilità e riproducibilità dei dati.

3.2.3 La rappresentatività e il bilanciamento del Web

Un corpus, per esser definito linguistico, deve essere rappresentativo di una lingua, ossia deve riprodurre idealmente, seppur in miniatura, tutte le peculiarità di una determinata varietà linguistica. Una caratteristica ancora più importante è il bilanciamento di un corpus: un corpus è bilanciato quando include al suo interno un ampio spettro di categorie testuali differenti, rappresentative della lingua. Ne consegue che è proprio il bilanciamento a garantire la rappresentatività di un corpus.

Alla luce di quanto detto sorge spontaneo chiedersi se e in che misura il Web sia rappresentativo. In virtù delle sue grandi dimensioni ed eterogeneità, il Web potrebbe essere potenzialmente rappresentativo, ma la mancanza di bilanciamento ne inficia la rappresentatività: è vero che nel Web trova spazio una gamma molto ampia di generi testuali, ma è altrettanto vero che alcuni di questi generi, come la scrittura accademica, la narrativa, i blog e le chat, sono sovra-rappresentati, mentre altri generi, come le

conversazioni telefoniche e i dialoghi quotidiani, sono sotto-rappresentati o non rappresentati affatto. Lo stesso vale per le lingue e varietà linguistiche presenti: la lingua più rappresentata è sicuramente l'inglese, mentre alcune varietà linguistiche, come quelle regionali, rimangono una categoria sotto-rappresentata.

Quando nel 2003 Kilgarriff e Grefenstette hanno proposto per la prima volta di considerare il Web un corpus linguistico, hanno minimizzato la questione relativa alla rappresentatività asserendo: "First, 'representativeness' begs the question 'representative of what?' […], we do not know with any precision what existing corpora might be representative of" (Kilgarriff, Grefenstette, 2003: 340), e avevano risolto la questione affermando: "The Web is not representative of anything else. But neither are other corpora, in any well-understood sense" (Kilgarriff, Grefenstette, 2003: 343).

Sulla scia di questi due autori, Gatto aveva affermato:

> Certainly the web cannot be considered a representative sample of language use in general, but its scope, variety, and above all its immense size seem to legitimize the opinion that these characteristics can counterbalance the limits of representativeness. […] the web's impossibility of being representative of nothing else but itself (Kilgarriff and Grefenstette 2003) does not altogether destroy its value as a source of linguistic information from a corpus linguistics perspective. (Gatto, 2014: 45)

In altre parole, secondo Gatto, sebbene il Web non sia un campione rappresentativo della lingua, le sue grandi dimensioni ed eterogeneità controbilancerebbero la mancanza di rappresentatività. Tuttavia, gli autori sembrano trascurare la mancanza di bilanciamento del Web, tanto che non la menzionano affatto, ma è bene sottolineare che in assenza di bilanciamento la rappresentatività da loro citata non è in alcun modo garantita.

La linguistica dei corpora è una disciplina scientifica che studia il linguaggio mediante indagini statistiche su dati certi, con risultati chiari e ripetibili. Proprio in virtù della sua scientificità ritengo che alla base delle argomentazioni vi debbano essere motivazioni di natura altrettanto rigorosa e non ci si possa nascondere dietro la logica dei sillogismi, riducendo la questione ad affermazioni di carattere filosofico, senza alcun fondamento scientifico.

3.2.4 La ripetibilità e la riproducibilità del Web

Come in ogni esperimento di carattere scientifico, anche nella linguistica dei corpora è necessario che le ricerche siano ripetibili e riproducibili, al fine di verificare i

risultati ottenuti. Come è già stato osservato, in questo tipo di indagini statistiche il linguista deve operare entro confini ben definiti, il che si traduce nell'adozione di un campione d'analisi necessariamente finito e definito; tuttavia, la mancanza di finitezza del Web e i suoi costanti mutamenti rendono impossibile ripetere o riprodurre le indagini linguistiche.

Con questa problematica si sono scontrati diversi autori, tra cui Kilgarriff, che in seguito ad un esperimento di ricerca, ripetuto a distanza di un giorno e che ha portato a risultati differenti, aveva osservato: "The engines will give you substantially different counts, even for repeats of the same query. […] The reasons are that queries are sent to different computers, at different points in the update cycle, and with different data in their caches" (Kilgarriff, 2007: 148). Quindi, i risultati di ricerca ottenuti sono differenti non solo perché quotidianamente vengono create, ed al tempo stesso eliminate, nuove pagine web, portando a dei cambiamenti nel campione di ricerca, ma variano anche in base al luogo, al *data center*[131] a cui i dati vengono inviati o al *device*[132] utilizzato.

Infatti, al fine di rendere sempre più performanti e personalizzati i risultati delle ricerche, i *browser*[133] geolocalizzano la posizione dell'utente, individuano se quello che sta usando è un *device* aziendale o personale, mediante i cookie[134] vengono memorizzate le informazioni sulle abitudini dell'utente riguardante lo storico dei siti web visitati.

L'opinione dei linguisti al riguardo è unanime e può essere riassunta con quanto viene riportato in Hundt, Nesselhauf e Biewer: "When using the web as a corpus – especially when it is accessed through a commercial search engine – it is virtually impossible to test for reproducibility (Hundt, Nesselhauf, Biewer, 2007: 11).

Una possibile soluzione per far fronte al problema è quella di operare su un numero finito di testi tratti dal Web, ossia utilizzando l'approccio *Web for Corpus*.

---

[131] Il *data center* o CED (Centro Elaborazione Dati) è una infrastruttura fisica usata dalle aziende per elaborare, organizzare, proteggere e conservare i dati dei computer, governare e gestire le apparecchiature aziendali. È il cuore pulsante delle aziende e del business, ciò che ne garantisce il funzionamento.

[132] Secondo quanto riportato dal sito ufficiale di Garzanti Linguistica, per *device* si intende il "dispositivo elettronico; si dice in particolare di dispositivi e apparecchi ad alta tecnologia e di piccole dimensioni (smartphone, e-book reader, tablet PC ecc.)". Link: https://www.garzantilinguistica.it/ricerca/?q=device.

[133] Il *browser* è un programma o applicazione che permette di inoltrare domande di ricerca, navigare in Internet, visualizzare le pagine di ricerca ecc. I browser più noti e utilizzati sono Google Chrome, Internet Explorer, Mozilla Firefox, Safari.

[134] I *cookie* in informatica sono dei file di informazioni che vengono salvati nel computer dell'utente durante la navigazione in Internet, al fine di memorizzare le preferenze dell'utente, i dati di registrazione, i siti visitati, ecc.

3.2.5 Conclusioni: il *Web as Corpus*?

In questo paragrafo chiariremo definitivamente il pensiero degli autori a sostegno o in opposizione all'idea del *Web as Corpus*.

Sostenitori di questo approccio sono i già citati Kilgarriff e Grefenstette, secondo i quali se per corpus linguistico si intende una qualsiasi raccolta di testi, creata per condurre studi linguistici o letterari, allora "The answer to the question 'Is the web a corpus?' is yes" (Kilgarriff, Grefenstette, 2003: 334), e successivamente aggiungono "The Web is clearly a multilingual corpus" (Kilgarriff, Grefenstette, 2003: 337).

Fletcher crede che il Web possa essere considerato un corpus linguistico, ma evidenzia i pro e i contro di questo approccio, sostenendo che:

> Both as a corpus and as a source of texts for corpora the Web offers significant benefits in its virtually comprehensive coverage of major languages, content domains and written text types, yet its usefulness is limited by the generally unknown origin and reliability of online texts and by the sheer amount of "noise" on the Web. (Fletcher, 2004: 1)

Successivamente però aggiunge che, secondo la sua generale impressione, il Web si presenta come una fonte di dati linguistici tanto utili, quanto affidabili e che, in base alla sua esperienza, il Web è una risorsa attendibile, se utilizzata per l'analisi di specifiche parole o frasi (Fletcher, 2004).

Tuttavia, nonostante le sue convinzioni in favore del *WaC*, bisogna sottolineare che Fletcher utilizza l'approccio *Web for Corpus*: infatti, l'obiettivo dichiarato del suo articolo è "to analyze language samples from the Web, not to investigate the language of the Web in general" (Fletcher, 2004: 2), e nelle conclusioni afferma che "this paper has surveyed a number of techniques and algorithms for downloading, preprocessing and evaluating texts from the Web for inclusion in a corpus" (Fletcher, 2004: 9).

Anche Lew sostiene l'approccio *WaC*, asserendo che i corpora, intesi come raccolta di testi autentici in formato elettronico, sono una preziosa risorsa di dati e in tal senso il World Wide Web può essere visto come un grande corpus dinamico (Lew, 2009). Infine, nelle conclusioni del suo articolo scrive:

> Based on the above comparison of traditional electronic text corpora and the textual resources of the World Wide Web, it can be concluded that the WWW, despite its noisiness and poor balancing, can be an attractive and useful tool for on-line language reference. Its main virtues lie in the impressive size of the resource, and the speed with which it can be trawled using a general-access search engine. (Lew, 2009: 15)

Secondo altri autori, invece, non è possibile considerare il Web un corpus linguistico. Un esempio al riguardo è Sinclair, che nel paragrafo intitolato "What is not a corpus?" risponde in modo piuttosto categorico:

> There are many collections of language text that are nothing like corpora. The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present it is quite mysterious […] and it is not at all clear what population is being sampled. (Sinclair, 2005: 21)

Altri autori mostrano un atteggiamento più moderato: tra questi occorre citare Barbera, Corino e Onesti, secondo i quali la risposta è perlopiù negativa, in quanto non si tratta di un corpus in senso stretto, ma può essere una fonte di dati testuali per creare i cosiddetti web corpora o utilizzare l'approccio *mini-Web/mega-corpus* (Barbera, Corino, Onesti, 2007). A riguardo si è espressa la stessa E. Corino, affermando che:

> L'esplorazione delle risorse web come 'mega corpus' risponde all'insufficienza quantitativa della base di dati per affrontare problematiche linguistiche specifiche sempre più complesse, e al sempre più rapido 'invecchiamento' dei materiali rispetto al continuo evolversi del linguaggio. (Corino, 2014: 234)

A mio parere, sarebbe necessario fare chiarezza in primis sulla definizione *Web as Corpus*, che viene spesso utilizzata in modo sregolato e in relazione non allo specifico approccio, ma come sinonimo di Web o Internet Linguistics, per andare ad indicare indistintamente tutti gli approcci visti in § 3.1.

In secondo luogo, ritengo che il Web sia una fonte ancora perlopiù sconosciuta, incontrollata, instabile e tutti questi limiti non possono essere ignorati, né minimizzati; pertanto, sulla base delle mie ricerche, non sono propensa a considerare il Web un corpus linguistico. Al contrario, utilizzare il Web come "portale per estrarre conoscenza", come nel caso dell'approccio *Web for Corpus*, è un approccio sensato e al tempo stesso altamente produttivo e fruttuoso. Infatti, selezionando i dati testuali, è possibile ottenere corpora di dimensioni finite, bilanciati, rappresentativi e, soprattutto, accuratamente annotati; in altre parole, è meglio puntare sulla qualità dei dati testuali, rispetto alla quantità.

Nel prossimo paragrafo verranno illustrate le procedure per la creazione di Web Corpora, sofisticati strumenti di ricerca di ultima generazione.

3.3 Il *Web for Corpus* e i Web corpora

Dietro i Web corpora si celano numerose procedure che trasformano progressivamente il materiale testuale "grezzo" del web in uno strumento che permette indagini linguistiche complesse. Questo paragrafo si pone l'obiettivo di offrire al lettore una breve panoramica di questo processo[135], per far non solo comprendere come vengono creati i Web corpora, ma anche per far riflettere sulla natura estremamente raffinata e complessa degli stessi.

In primo luogo occorre selezionare i dati testuali dal Web. Questa operazione viene effettuata principalmente attraverso due modalità: il *search engine*, ossia il motore di ricerca, oppure il *web crawling*.

La ricerca svolta attraverso il *search engine* può essere effettuata mediante interrogazione manuale o attraverso appositi programmi; esistono anche dei software che selezionano automaticamente un certo numero di URL di base da cui iniziare la ricerca, come la risorsa completamente gratuita *BootCaT*[136], oppure si può partire da una lista di parole con frequenza medio-alta o selezionate sulla base di *KWiCFinder* (dall'acronimo "Key Word in Context"[137]).

Il secondo metodo è quello del *web crawling*, espressione che deriva dal verbo inglese *to crawl*, 'avanzare lentamente' e 'progredire a lento ritmo, strisciando o gattonando'; nella Web Linguistics acquisisce il significato di 'scandagliare il web', mediante un apposito programma chiamato *crawler*. Il *crawler* scandaglia una lista di URL fornita dal linguista[138] e, mentre analizza ogni singola pagina, identifica i vari collegamenti ipertestuali contenuti al suo interno, aggiungendoli alla lista iniziale di URL da visitare.

I *web crawler* sono solitamente automatici e ciò rende questa procedura piuttosto veloce e precisa. Tra i programmi di *crawling* più diffusi possiamo citare *SpiderLing*[139],

---

[135] Per un quadro completo e specifico si veda Benko (2014), Benko, Zacharov (2016) e l'esperienza di Baroni, Bernardini, Ferraresi, Zanchetta (2009). Tutti questi autori descrivono nel dettaglio le procedure e gli strumenti per creare i Web corpora.

[136] Per maggiori informazioni su BootCaT o per scaricarlo gratuitamente, si consulti il link: https://bootcat.dipintra.it/.

[137] Si tratta di un *web concordancer* che individua le parole chiave in un determinato contesto. Per maggiori informazioni si veda il sito: https://www.kwicfinder.com/KWiCFinder.html.

[138] Il nome tecnico è *seed URL*, dalla parola inglese *seed*, ossia 'seme' o 'germoglio'. Per una descrizione accurata al riguardo si veda Baroni, Bernardini (2006: 16). La lista di *seed URL* può anche essere generata automaticamente mediante specifici programmi, come il già citato *BootCaT*.

[139] Per maggiori informazioni sul programma di web crawling *SpiderLing* si veda il link: http://corpus.tools/wiki/SpiderLing.

presente in Suchomel e Pomikálek (2012), due degli sviluppatori del programma e utilizzato da Jakubíček, Kilgarriff, Kovář, Rychlý, Suchomel (2013) per la costruzione dei corpora appartenenti alla cosiddetta *TenTen Corpus Family*[140], e da Benko (2014) per la creazione dei corpora *Aranea*[141].

Un altro *web crawler* molto popolare è *Heritrix*[142], citato da Baroni e Bernardini (2006) e Hundt, Nesselhauf, Biewer (2007) e utilizzato da Baroni e Kilgarriff (2006) e successivamente da Baroni, Bernardini, Ferraresi, Zanchetta (2009) per la creazione dei corpora *ukWaC*, *deWaC* e *itWaC*[143].

Dopo aver raccolto i dati necessari, occorre attuare procedure di *deduplication*, *cleaning* e *filtering*: i testi vanno innanzitutto de-duplicati, per rimuovere le ripetizioni che inficerebbero sui risultati finali di ricerca, come, ad esempio, una semplice indagine riguardante la frequenza delle parole. Attualmente esistono specifici programmi automatici per la de-duplicazione, tra i quali *Onion* (*ONe Instance ONly*)[144] e *FindDuplicates*[145].

Occorre poi pulire i testi dai cosiddetti *boilerplate* presenti nelle web page, ossia tutte quelle risorse non-testuali che non hanno carattere informativo, come i menù di navigazione, i disclaimer, la pubblicità e i web spam, i link, le informazioni sul copyright, le intestazioni, i piè di pagina, ecc.

---

[140] Il nome *TenTen Corpus Family* deriva dall'iniziale obiettivo dei creatori di questa famiglia di corpora: raggiungere i dieci (*ten*, in inglese) miliardi di *token*. Attualmente questi corpora comparabili sono disponibili in oltre quaranta lingue. Della versione russa, chiamata *ruTenTen*, parleremo dettagliatamente nel prossimo capitolo, nel paragrafo intitolato "Gli strumenti di indagine" (§ 4.2). Per maggiori informazioni si veda il link: https://www.sketchengine.eu/documentation/tenten-corpora/.

[141] La famiglia di corpora Aranea rappresenta oltre 23 lingue e varietà linguistiche. I nomi dei singoli corpora, in latino, indicano la varietà linguistica contenuta nel corpus (ad esempio: Anglicum, Italicum, Russicum, ecc.), inoltre la denominazione è seguita dalla grandezza (Minus, Medium, Maius, Maximum). Ad esempio, per il russo esistono i seguenti corpora Aranea: *Russicum* Minus, Maius e Maximum, *Russicum Russicum* Minus e Maius; *Russicum Externum* Minus e Maius. Per maggiori informazioni si consulti il link: http://ucts.uniba.sk/aranea_about/index.html.

[142] Per maggiori informazioni sul programma *Heritrix* si veda il link: https://webarchive.jira.com/wiki/spaces/Heritrix/overview.

[143] I corpora *ukWaC*, *deWaC* e *itWaC* fanno parte del progetto *Wacky*, acronimo di *Web-As-Corpus Kool Yinitiative*, creato da una comunità di ricercatori, linguisti e specialisti nel campo dell'IT. Per maggiori informazioni si consulti il sito ufficiale: https://wacky.sslmit.unibo.it/doku.php?id=start.

[144] *Onion* vine utilizzato e citato da Jakubíček, Kilgarriff, Kovář, Rychlý, Suchomel (2013) e da Benko (2014). Per maggiori informazioni sulla sua creazione e funzionamento si veda Pomikalek (2011) e il link: http://corpus.tools/wiki/Onion.

[145] Il programma *FindDuplicates* viene citato ed utilizzato da Fletcher (2004)

È possibile farlo mediante appositi programmi, come *jusText*[146], basato su un algoritmo che rimuove tutto il materiale contenente basse percentuali di parole grammaticali. Altre risorse gratuite per la rimozione dei boilerplate sono *Body Text Extraction* (*BTE*), *Boilerpipe*, *CleanEval*[147].

Un'ulteriore procedura di *cleaning* riguarda il materiale pornografico contenuto nelle pagine web. Un metodo molto semplice, ma incredibilmente efficace, è quello utilizzato da Baroni e Kilgarriff: "We filter out documents that have at least three types or ten tokens from a list of words highly used in pornography" (Baroni, Kilgarriff, 2006: 88). Lo stesso metodo è stato utilizzato da altri linguisti, come Baroni, Bernardini, Ferraresi, Zanchetta (2009).

Dopo queste procedure, i dati testuali risultano idonei e pronti per la tokenizzazione, in altre parole i testi vengono segmentati in *token*, le minime unità linguistiche. Questo passaggio preliminare è essenziale per le successive operazioni di annotazione, come la lemmatizzazione, il *POS-tagging*, il *parsing*, ecc. (§ 2.3.1). Anche in questo caso esistono programmi appositi, come *Unitok*[148], per la tokenizzazione, mentre per l'annotazione morfosintattica è molto popolare *TreeTagger*[149], in quanto può essere utilizzato con numerose lingue. Solo in seguito a tutte queste operazioni sarà possibile caricare il Web corpus online e renderlo fruibile per gli utenti.

Questa breve disamina voleva fornire un quadro teorico generale circa il complesso processo di creazione dei Web corpora; nel prossimo paragrafo verranno trattate due importanti problematiche relative ai web corpora: il copyright e il *netspeak*.

3.4 Problemi relativi alla compilazione dei web corpora

Quando si utilizza il Web come corpus linguistico o i dati testuali presenti in Internet per creare i Web corpora, occorre riflettere su quali siano le leggi in materia di

[146] *JustText* viene citato ed utilizzato da utilizzato da Jakubíček, Kilgarriff, Kovář, Rychlý, Suchomel (2013) e da Benko (2014). Per maggiori informazioni sulla sua creazione e funzionamento si veda Pomikalek (2011).

[147] Per una rassegna di tutti questi algoritmi e una valutazione sulle loro performance si veda Endredy, Novak (2013).

[148] *Unitok* è stato utilizzato anche da Benko, Zacharov (2016) per la creazione del web corpus russo *Araneum Russicum Maximum*. Per maggiori informazioni si veda: http://corpus.tools/wiki/Unitok.

[149] *TreeTagger* è stato utilizzato da Jakubíček, Kilgarriff, Kovář, Rychlý, Suchomel (2013) per la creazione dei corpora appartenenti alla cosiddetta *TenTen Corpus Family*, da Baroni, Kilgarriff (2006) e, infine, da Benko, Zacharov (2016) per la creazione del web corpus russo *Araneum Russicum Maximum*. Per maggiori informazioni si veda il link: https://www.cis.lmu.de/~schmid/tools/TreeTagger/.

copyright. Un'altra problematica relativa alla compilazione dei web corpora, meno ovvia o evidente, è rappresentata dal cosiddetto *netspeak*.

In questo paragrafo verranno affrontate queste due tematiche, che si pongono come nuove sfide per i linguisti e per la creazione dei corpora di nuova generazione.

3.4.1 Il copyright nel Web

Nel 1986, in relazione alla futura digitalizzazione, J. Clear scrisse:

> In the 1990s the information technology boom will certainly ensure that documents are stored and transmitted in digital form. Unfortunately, the speed of technological advance has left us with an ethical and legal confusion over the ownership of information, which is hindering the acquisition of text. (Clear, 1986: 385)

Sicuramente anticipò con lucidità la futura incertezza, relativa ai problemi legali e al copyright[150], che avrebbe caratterizzato gli anni '90.

Nel 1991, Atkins, Clear e Osler affermarono che, in risposta ai rapidi sviluppi tecnologici, le leggi relative al copyright erano state estese o revisionate e di conseguenza "the effect for the corpus builder is that it is quite likely that any text (or sample of text) which is to be computerised and included in a corpus will be under copyright protection and that permission will have to be obtained for its use" (Atkins, Clear, Osler, 1991: 6). Successivamente aggiunsero che non solo le fonti scritte, ma anche la digitalizzazione delle trascrizioni audio (conversazioni, programmi alla radio o alla TV, ecc.) erano tutelatate dal copyright e il loro utilizzo richiedeva autorizzazioni specifiche.

Tuttavia, il periodo di indeterminatezza riguardante la legge sui diritti d'autore nel mondo del Web, anticipato da J. Clear, si sarebbe protratto sino agli anni 2000; ne è testimone l'articolo di Cavaglià e Kilgarriff in relazione al Web e ai materiali testuali in esso contenuti, pubblicato proprio nel 2000, in cui gli autori asserivano: "Moreover text is available immediately, for free, and can be downloaded without concern for copyright" (Cavaglià, Kilgarriff, 2000: 1). È chiaro che anche i testi digitalizzati presenti nel Web sono protetti dai diritti d'autore al pari di quelli in formato cartaceo,

[150] Il termine copyright è utilizzato per indicare i diritti d'autore nei paesi del *common law*, ossia afferenti al sistema giuridico dei paesi anglo-americani basato sul precedente giurisprudenziale. Tuttavia, attualmente il termine copyright viene utilizzato indistintamente, anche nei paesi di *civil law*. In russo, invece, le due parole hanno due significati differenti: con *kopirajt* spesso si indica la scrittura di un testo a scopo pubblicitario o altre forme di marketing, mentre l'*avtorskoe pravo* (in italiano 'legge sul diritto d'autore') riguarda l'insieme delle leggi relative alla creazione e all'uso di opere scientifiche, letterarie, artistiche, ecc.

pertanto non possono essere utilizzati indiscriminatamente, eccetto qualche rara eccezione che vedremo a breve.

Nel corso degli anni 2000 altri autori si sono espressi a riguardo e per fornire un quadro più completo e ripercorrere la storia della disciplina in relazione al copyright, verranno riportate le voci principali.

Nel 2005 Sinclair affermava:

> Another tricky question is that of copyright — not the familiar copyright of publications, but the more nebulous issue of electronic copyright. In principle, under UK law, publication on the internet confers the rights on the author whether or not there is an explicit copyright statement. (Wynne, 2005: 98)

Successivamente, nel 2009, Baroni, Bernardini, Ferraresi e Zanchetta in relazione alla creazione dei *WaCky Corpora* affermavano che "the copyright issue remains a thorny one: there is no easy way of determining whether the content of a particular page is copyrighted, nor is it feasible to ask millions of potential copyright holders for usage permission" (Baroni, Bernardini, Ferraresi, Zanchetta, 2009: 18) ed aggiungevano che nel loro sito ufficiale erano presenti tutte le informazioni necessarie per chiedere la cancellazione di specifici documenti dai loro corpora.

È importante sottolineare che quando si parla di copyright in relazione alla linguistica dei corpora, non si fa riferimento solo a quello dei singoli materiali testuali contenuti in un corpus, ma anche ai diritti d'autore che tutelano l'intero corpus linguistico. In altre parole, ci sono almeno due differenti livelli di copyright che i linguisti devono considerare sin dal momento della creazione del corpus.

Secondo McEnery e Hardie esistono quattro differenti approcci alla questione, ripresi successivamente da Gatto[151]: il primo consiste nel contattare i possessori dei diritti d'autore sul materiale testuale e chiedere il permesso di utilizzare e ri-distribuire i testi; questo metodo è stato utilizzato dai creatori di corpora di piccole dimensioni, come il *BNC* e il *LOB Corpus*. Una seconda soluzione è utilizzare materiale testuale esplicitamente di pubblico dominio, come i materiali *Creative Commons*[152] oppure *free-*

[151] Di seguito vengono brevemente riassunti i quattro approcci proposti dagli autori; per una lettura più approfondita a riguardo si veda McEnery, Hardie (2012: 57) e Gatto (2014: 64).

[152] *Creative Commons* è un'organizzazione senza scopo di lucro che da vent'anni fornisce licenze di vario tipo relative al copyright, al fine di rendere fruibili le risorse del Web. Il Web corpus dell'italiano PAISÀ è stato creato esclusivamente con testi Creative Commons, come affermano gli autori: "The main novelty of the PAIS`A web corpus is that it exclusively draws on Creative Commons licensed data […]" (Lyding *et alii*, 2014: 36). Per maggiori informazioni si veda il sito ufficiale di *Creative Commons*: https://creativecommons.org/.

software, come quelli contenuti in *GNU*[153]. Il terzo approccio consiste nel raccogliere i dati testuali senza tener conto del copyright e senza distribuire il corpus finale, oppure renderlo disponibile ad altri studiosi attraverso strumenti che non infrangono le leggi sul copyright[154]; questo approccio, il più diffuso al momento, è quello utilizzato per la creazione dei già citati *WaCky corpora*. Il quarto ed ultimo approccio proposto da McEnery e Hardie è quello di ridistribuire non il file di testo, ma piuttosto l'elenco dei siti web utilizzati per la creazione del corpus. Quest'ultima soluzione non viola il copyright e permette a qualsiasi ricercatore di ricostruire una copia personale del corpus.

In questo paragrafo è stata introdotta la questione del copyright, sono stati riportati alcuni punti di vista dei principali autori a riguardo e successivamente sono stati illustrati quattro differenti approcci per creare e distribuire i Web corpora, nel rispetto del copyright. Nel prossimo paragrafo verrà trattata un'altra problematica, meno evidente, legata all'utilizzo dei dati testuali del web: il *netspeak* e le sue caratteristiche.

3.4.2 Il *netspeak*

Lo sviluppo di Internet, inteso come mezzo di comunicazione virtuale e globale, ha portato alla nascita di nuovi generi testuali, nonché a dei mutamenti della lingua stessa.

A partire dagli anni 2000 sono state introdotte innumerevoli definizioni, come *netspeak*, *netlingo*, *weblish*, *globespeak*, *cyber-slang*, *digispeak*, *chatspeak* e ancora *Internet language*, *cyberspeak*, *electronic discourse*, *electronic language*, *interactive written discourse* o *computer-mediated communication*[155], tante denominazioni per indicare una sola lingua, quella utilizzata nel Web.

Per certi versi si può parlare di una vera e propria varietà linguistica a sé stante, con caratteristiche così uniche da essere diventata un interessante oggetto di studio;

[153] *GNU*, acronimo di "*GNU's Not Unix*", è un sistema operativo costituito da applicazioni, librerie, strumenti di sviluppo, giochi, ecc. Secondo quanto riportato nel sito ufficiale, gli utenti godono di quattro libertà: "Libertà di eseguire il programma come si desidera, per qualsiasi scopo. Libertà di studiare come funziona il programma e modificarlo in modo che funzioni a piacimento. Libertà di ridistribuire copie in modo da aiutare gli altri. Libertà di distribuire pubblicamente le proprie versioni modificate agli altri". Per altre informazioni si veda il sito ufficiale: https://www.gnu.org/home.it.html.

[154] Un esempio al riguardo consiste nel distribuire questi corpora mediante *concordancer* che mostrano solo brevi stringhe di parole e non forniscono molto contesto. In questo modo è impossibile ricostruire l'intero testo e il copyright non viene infranto.

[155] Esistono anche le varianti *netlish* e *weblish*, ma sono utilizzate in riferimento alla lingua inglese (infatti, il suffisso *–lish* deriva da *English*). Per una trattazione approfondita delle denominazioni riportate si veda Crystal (2004: 17).

infatti, come afferma Corino, l'uso di Internet "ha aperto nuove prospettive e interessi di ricerca in campo linguistico, dallo studio delle email alla chat, dai newsgroup a Twitter, dai blog a Facebook, si sono moltiplicati gli studi di pragmatica e testualità" (Corino, 2014: 231).

Anche nell'ambito della corpus linguistics sono state condotte delle ricerche a riguardo, ma se da una parte il *netspeak* costituisce un interessante oggetto di studio, nonché materiale per la creazione dei web corpora, dall'altro possono emergere una serie di problematiche, legate alle sue peculiarità intrinseche.

Innanzitutto, il *netspeak* utilizzato nelle chat, nei blog, nei forum, nei social network e talvolta anche nelle e-mail, presenta caratteristiche ibride tipiche sia della lingua parlata, sia di quella scritta; il noto linguista David Crystal si è occupato di questo sin dai primi anni 2000, sintomo che già all'epoca la lingua del Web era sensibilmente mutata, osservando che negli anni '90 alcuni studiosi avevano definito la lingua di Internet un "written speech", o ancora "write the way people talk" (Crystal, 2004: 25).

Secondo Crystal, ciò che differenzia la lingua parlata dal *netspeak* è in primis l'assenza di un feedback simultaneo, tipico dell'interazione tra due o più individui: infatti, il messaggio deve essere necessariamente inviato, prima di raggiungere uno o più destinatari. L'assenza di feedback simultaneo determina la mancanza di un'altra caratteristica tipica del parlato, ossia delle reazioni simultanee dell'ascoltatore, come cenni del capo, interiezioni, linguaggio non verbale di approvazione o dissenso; inoltre, nel parlato la conversazione può sovrapporsi o essere interrotta, mentre ciò non accade nei messaggi, dove il ritmo della conversazione è generalmente più lento.

Similmente alla lingua scritta, per sopperire alla mancanza di intonazione, tono di voce e pause, vengono utilizzate la punteggiatura, le lettere maiuscole, i colori, gli spazi e così via. Come riporta Crystal:

> Examples include repeated letters (aaaaahhhhh, hiiiiiii, ooops, soooo), repeated punctuation marks (no more!!!!!, whohe????, hey!!!!!!!!!, see what you started?????????????????) […] all capitals for 'shouting': I SAID NO; letter spacing for 'loud and clear': WH Y N O T, w h y n o t. (Crystal, 2004: 34)

Un'altra caratteristica unica del *netspeak* è l'utilizzo di emoticon, smile, gif o disegni[156] per manifestare emozioni, comunicare espressioni facciali, gesti o addirittura per parlare di attività, oggetti o cibo, che lo rendono più simile al parlato, rispetto alla tradizionale lingua scritta.

Questi contenuti, che possono rendere un messaggio più amichevole, con il tempo hanno determinato un mutamento nel modo in cui i messaggi altrui vengono percepiti o interpretati: infatti, senza alcun tipo di emoticon i messaggi possono risultare freddi, distaccati o addirittura aggressivi. Inoltre, il *netspeak* e l'utilizzo diffuso degli emoticon ha cambiato il modo di percepire le tradizionali forme di punteggiatura: ad esempio, secondo recenti studi il punto alla fine di un messaggio viene percepito come segnale di insincerità dell'interlocutore, oppure di un atteggiamento passivo-aggressivo[157].

Una caratteristica che rende il *netspeak* più simile alla comunicazione scritta è la possibilità di pensare ad un discorso strutturato, cancellare, riscrivere e rileggere quanto scritto; in questo caso è stato volutamente utilizzato il sostantivo "possibilità", in quanto talvolta i messaggi online non vengono ricontrollati o corretti, ma semplicemente inviati così come sono stati scritti, esattamente come se l'utente stesse parlando. Questo determina un alto numero di messaggi contenenti errori o assenza di punteggiatura, errori di ortografia o di battitura, ripetizioni, sintassi sconnessa e così via.

D. Crystal, nel condurre la sua analisi, ha individuato una serie di caratteristiche tipiche della comunicazione parlata e scritta, stabilendo successivamente se queste caratteristiche siano presenti anche nelle chat, nelle e-mail ecc. I risultati della sua ricerca sono riassunti nella *Figura 4* (Crystal, 2004: 42) e nella *Figura 5* (Crystal, 2004: 43), qui riportate.

---

[156] In questo contesto si fa riferimento a quei disegni creati tramite segni di punteggiatura o simboli. Un esempio al riguardo è il seguente, intitolato "Bambino con il palloncino":

Esiste, inoltre, l'*ASCII art*, un tipo di arte che utilizza i 95 caratteri ASCII per produrre immagini realistiche.

[157] In particolare, lo studio condotto dall'Università di Binghamton, i cui risultati sono stati presentati nell'articolo *Texting insincerely: The role of the period in text messaging* (Gunraj *et alii*, 2016). Numerosi studi sono stati condotti dalla specialista di Internet Linguistics Gretchen McCulloch, consultabili nel suo sito ufficiale https://gretchenmcculloch.com/. Infine, un'altra studiosa molto produttiva è Erika Darics, docente presso l'Università di Groningen; si veda il link: https://scholar.google.co.uk/citations?user=M1YG-2kAAAAJ&hl=en.

	Web	e-mail	Chatgroups	Virtual worlds
1 time-bound	no	yes, but in different ways	yes, but in different ways	yes, but in different ways
2 spontaneous	no	variable	yes, but with restrictions	yes, but with restrictions
3 face-to-face	no	no	no	no
4 loosely structured	variable	variable	yes	yes
5 socially interactive	no, with increasing options	variable	yes, but with restrictions	yes, but with restrictions
6 immediately revisable	no	no	no	no
7 prosodically rich	no	no	no	no

Figura 4. Le caratteristiche della comunicazione parlata applicate al *netspeak* (Crystal, 2004: 42).

	Web	e-mail	Chatgroups	Virtual worlds
1 space-bound	yes, with extra options	yes, but routinely deleted	yes, but with restrictions	yes, but with restrictions
2 contrived	yes	variable	no, but with some adaptation	no, but with some adaptation
3 visually decontextualized	yes, but with considerable adaptation	yes	yes	yes, but with some adaptation
4 elaborately structured	yes	variable	no	no
5 factually communicative	yes	yes	variable	yes, but with some adaptation
6 repeatedly revisable	yes	variable	no	no
7 graphically rich	yes, but in different ways	no	no	yes, but in different ways

Figura 5. Le caratteristiche della comunicazione scritta applicate al *netspeak* (Crystal, 2004: 43).

Sulla base dei risultati ottenuti, Crystal conclude che il *netspeak* è più simile alla comunicazione scritta ed aggiunge:

> In my estimation the actual amount that Netspeak has in common with speech is very limited. The Web is furthest away from it; chatgroup and virtual world interactions are somewhat closer to it; and e-mails sit uncertainly in the middle. (Crystal, 2004: 40)

Le sue osservazioni sono state confermate da uno studio sulla lingua norvegese, condotto nel 2011 dai linguisti Johannessen e Guevara dell'Università di Oslo. I due linguisti hanno comparato un Web corpus, il *Norwegian NoWaC corpus*, con un corpus di testi scritti, l'*Oslo Corpus of Written Norwegian Bokmål Texts*, ed una combinazione di due corpora di fonti orali, il *Nordic Dialect Corpus* e il *NoTa-Oslo* (*Norwegian Speech Corpus - Oslo part*); l'obiettivo di questa ricerca era stabilire se il Web corpus fosse più simile al corpus scritto o a quello orale.

A seguito della loro indagine i due linguisti hanno concluso:

> For typically written language variables, such as formal subordinators, NoWaC is like a written corpus. Looking at variables that will say something about the extent to which spoken topics are concerned, such as subordination, NoWaC is still like a written corpus,

> although by a small margin. Checking for spoken version words of those that have several variants, NoWaC is still also like a written corpus. (Johannessen, Guevara 2011: 125)

E successivamente hanno aggiunto:

> However, NoWaC does have interjections that are typical of dialogue, revealing this way that it does have some qualities shared with the spoken corpus. […] NoWaC shows a relatively stronger correlation to the written reference corpus, although it is also correlated significantly to spoken data. (Johannessen, Guevara 2011: 127)

Pertanto, secondo quanto hanno rilevato Johannessen e Guevara, il *netspeak* presenta caratteristiche tipiche sia del parlato che dello scritto, tuttavia, come ha notato Crystal prima di loro, il *netspeak* mostra una maggiore affinità con la lingua scritta.

Infine, occorre citare altre caratteristiche o fenomeni tipici del *netspeak*: utilizzo di abbreviazioni e acronimi, reduplicazione delle lettere e messaggi ripetuti, citazioni di altri commenti o di altri utenti.

Una interessante osservazione è quella fatta da Paracchini, in relazione al *netspeak* russo: la linguista parla di "adeguamento fonetico", ossia della tendenza ad avvicinare il linguaggio scritto a quello parlato, traslando in forma scritta i suoni percepiti, ad esempio, scrivendo *zdras'te* al posto di *zdravstvujte* (Paracchini, 2017: 51).

Un'altra peculiarità del russo, emersa dalla ricerca condotta personalmente da chi scrive e i cui risultati verranno presentati nel prossimo capitolo, è la tendenza a traslitterare in russo parole inglesi utilizzando trascrizioni differenti; tra quelle analizzate nella mia ricerca cito блогер e блоггер, che potremmo interpretare come "bloger" e "blogger", дислайк e дизлайк, ossia "dislike" e "dizlike". Infine, инфлюенсер e инфлюэнсер, ossia *influencer* con due varianti della vocale "e" dolce e dura; quest'ultimo fenomeno, peraltro, risulta molto diffuso.

In *Figura 6* è riportata una schermata tratta da Sketch Engine, relativa ai risultati della ricerca per l'acronimo in russo *LOL* (*Lot Of Laughts* oppure *Laughing Out Loud*) nel corpus *ruTenTen17*.

Figura 6. Risultati di ricerca per l'acronimo *LOL* in *ruTenTen17*.

Come si può osservare, nella schermata si vedono diverse emoticon, create mediante segni di punteggiatura e evidenziate in *Figura 6* mediante un riquadro verde, vi è la compresenza di caratteri latini e cirillici e l'acronimo viene scritto con combinazioni differenti di maiuscole e minuscole, ad esempio, *LOL*, *Lol* e *lol*. Inoltre, sono presenti errori di battitura: ad esempio, nella frase 5373 lo strumentale plurale di *telefon* è scritto erroneamente *s telefonmi* e non *s telefonami*. Infine, ci sono diverse parole inglesi traslitterate in cirillico: nella frase 5371 viene utilizzata la parola *pikrilejted*, che è un prestito traslitterato dall'inglese *pic related*, ossia 'immagine relativa all'argomento trattato'; nella frase 5374, invece, compare l'espressione *do reliza*, in cui *reliz* è il prestito traslitterato della parola inglese *release*.

Tutte queste osservazioni sono emerse in relazione ad una decina di frasi e questo rende perfettamente l'idea delle problematiche che devono quotidianamente affrontare i linguisti al momento della creazione di un Web corpus: si tratta del cosiddetto "noise", che va filtrato ed omesso oppure deve essere conservato? Cosa deve essere ritenuto sbagliato e come bisogna comportarsi dinnanzi all'errore o nei confronti dell'utilizzo non normativo della lingua? Il linguista dovrà necessariamente prendere decisioni a tal proposito e, soprattutto, comunicarle e giustificarle ai fruitori del Web corpus.

3.5 Conclusioni del capitolo

In questo capitolo si è ripercorsa la storia della Web Linguistics, dalla sua nascita ad oggi, ed i vari approcci alla disciplina, i cui due principali si riassumono nelle espressioni *Web as Corpus* e *Web for Corpus*. Successivamente, si è passati alla rassegna delle principali critiche mosse nei confronti dell'approccio *Web as Corpus*, riguardanti l'autenticità ed autorevolezza dei dati linguistici, le dimensioni, l'assenza di finitezza e bilanciamento del Web, l'impossibilità di replicare o riprodurre i risultati della ricerca. A conclusione del paragrafo sono stati riportati i pareri dei principali studiosi nei riguardi dell'approccio *WaC*.

Nella terza parte del capitolo sono state illustrate le numerose e complesse procedure per la creazione dei Web corpora, secondo l'approccio *Web for Corpus*, ed infine sono state analizzate le principali problematiche relative alla loro compilazione, ovvero il copyright e la lingua del Web, il *netspeak*.

Sulla base di quanto è emerso nel corso di questo capitolo, ritengo che il Web sia un'importante fonte di materiale testuale di ogni genere e lingua, tuttavia è ancora piuttosto sconosciuto, incontrollato e incontrollabile per poter essere uno strumento affidabile d'indagine. La soluzione, a mio parere, dev'essere individuata nella creazione di Web corpora basati sul materiale tratto da Internet, scaricato e successivamente filtrato, ripulito ed organizzato.

Innegabilmente tutte queste operazioni possono influire sulla autenticità dei dati testuali, ma è altrettanto innegabile che in questo modo sarà possibile ottenere un corpus propriamente detto, con tutte le caratteristiche dei corpora linguistici tradizionali: autentico, di dimensioni finite, bilanciato e rappresentativo, i cui risultati di ricerca siano replicabili e riproducibili.

Nel prossimo ed ultimo capitolo verrà illustrata nel dettaglio la ricerca svolta mediante tre differenti corpora, il corpus *NKRJa* e i Web corpora *ruTenTen11* e *ruTenTen17*, volta ad analizzare alcune caratteristiche del *netspeak* russo contemporaneo.

4. I corpora linguistici come strumenti d'indagine del *netspeak* russo

4.1 Presentazione ed obiettivi della ricerca

In questo capitolo verrà illustrata la ricerca effettuata al fine di indagare il *netspeak* russo contemporaneo, ovvero il linguaggio giovanile, prevalentemente di origine anglofona, utilizzato nel World Wide Web.

Al fine di condurre un'analisi di tipo scientifico nell'indagare tale varietà linguistica, sono stati utilizzati diversi strumenti d'indagine che garantiscono risultati matematicamente quantificabili, affidabili e ripetibili: *NKRJa* e due Web corpora, il *ruTenTen11* e il *ruTenTen17*, disponibili nella piattaforma Sketch Engine. Nei casi in cui dall'analisi emergevano risultati ambigui o in contrapposizione tra loro, è stato utilizzato anche un terzo web corpus: il *Timestamped JSI web corpus 2014-2021 Russian*, anch'esso consultabile tramite Sketch Engine.

L'utilizzo congiunto di un corpus "tradizionale" e di più Web corpora ha permesso di realizzare un'analisi comparata tra questi strumenti di ricerca: infatti, è stato possibile stabilire quale strumento sia il più adeguato ed aggiornato per indagare la lingua russa nella sua contemporaneità. Allo stesso tempo, l'utilizzo di due Web corpora, realizzati in anni differenti, permette di analizzare l'evoluzione diacronica della lingua russa nel Web, nel 2011 (*ruTenTen11*) e nel 2017 (*ruTenTen17*).

Gli altri due strumenti d'indagine utilizzati sono il dizionario cartaceo Zanichelli *il Kovalev*, e due risorse online, *Vikislovar'* e *Academic.ru*. Anche in questo caso è stata condotta un'indagine comparativa tra i dizionari cartacei ed online.

Nel prossimo paragrafo saranno illustrati nel dettaglio questi strumenti di ricerca. In seguito si passerà alla descrizione della modalità di selezione e raccolta dei dati linguistici, ed infine il capitolo si focalizzerà sull'analisi dei dati e sulle conclusioni alle quali è giunta questa indagine linguistica.

4.2 Gli strumenti d'indagine utilizzati per la ricerca

Come accennato nel precedente paragrafo, i principali strumenti di analisi sono stati i corpora linguistici della lingua russa ed alcuni dizionari russi, cartacei ed online; in questo paragrafo verrà esaminata nel dettaglio ognuna di queste risorse.

4.2.1 Il *Nacional'nyj Korpus Russkogo Jazyka*

I lavori per la creazione di questo importante corpus nazionale sono iniziati nel 2003, grazie ad un nutrito team di esperti e linguisti provenienti dalle università di Mosca, San Pietroburgo, Kazan', Voronež, Saratov ed altri centri di ricerca russi.

Attualmente il corpus ha superato il miliardo di token, distribuiti in 17 sotto-corpora, il più grande dei quali è l'*Osnovnoj korpus*, il 'Corpus principale', contenente oltre 337 milioni di token, seguito dal *Gazetnyj korpus SMI 2000-ch godov*, il 'Corpus giornalistico dei mass media degli anni 2000' contenente quasi 333 milioni token.

Una funzione interessante di *NKRJa*, che vale la pena menzionare per la sua utilità, è la possibilità di verificare l'anno d'ingresso di una parola nella lingua russa e la sua frequenza d'uso, relativa ad un determinato intervallo temporale. I dati non solo vengono presentati in un grafico sotto forma di diagramma cartesiano, ma sono anche riassunti in una specifica tabella, grazie alla quale è possibile esaminare la frequenza d'uso della parola selezionata per ogni anno dell'intervallo ricercato e leggere gli esempi d'uso registrati nel corpus.

La grande varietà di generi testuali e periodi storici raccolti nel Corpus rendono *NKRJa* uno strumento idoneo ed indispensabile per le ricerche linguistiche, ma in questo capitolo si tenterà di comprendere se sia uno strumento altrettanto adeguato per indagare la contemporaneità della lingua russa.

I sotto-corpora utilizzati per la ricerca sono il già citato *Osnovnoj korpus* e l'*Ustnyj korpus*, il 'Corpus della lingua parlata'. Per quanto riguarda il primo, si è detto essere il più grande sotto-corpus di *NKRJa* per le sue dimensioni: esso, infatti, costituisce il 33,3% dell'intero corpus nazionale e contiene testi della contemporaneità, della metà del XX ed inizio del XXI secolo, nonché testi più antichi, scritti tra la metà del XVIII secolo e il XX secolo.

L'*Osnovnoj korpus* presenta un'annotazione metalinguistica: in particolare, viene indicato il nome, il sesso, la data di nascita o l'età approssimativa dell'autore e la data di creazione, il genere, il cronotopo (ossia le informazioni riguardanti il luogo o il periodo degli eventi descritti nel testo), nonché l'argomento del testo.

Per quanto concerne l'annotazione morfologica, sono indicati il genere, il caso, il numero, l'animatezza, la parte del discorso, se è una forma breve o lunga nel caso di participi e aggettivi, l'aspetto, il modo, il tempo, la diatesi verbale, la persona, il numero

del verbo, ecc. Infine, secondo quanto riportato nel sito di *NKRJa*[158], in una piccola parte dell'*Osnovnoj korpus* sono state eseguite manualmente specifiche procedure di rimozione dell'omonimia e la correzione dei risultati di analisi morfologica automatica.

L'*Ustnyj korpus*, invece, è stato creato nel 2007 e, secondo le informazioni presenti nella sezione *Novosti proekta* ('Ultime notizie sul progetto') relative all'1 giugno 2021, attualmente consta di 13,4 milioni di token, una cifra senz'altro ragguardevole, nonostante rappresenti solamente l'1,3% dell'intero corpus[159].

Il Corpus comprende trascrizioni di conversazioni, monologhi spontanei, interviste, racconti riguardanti una specifica tematica, riassunti orali di un testo scritto e letture ad alta voce di un testo noto o sconosciuto al lettore. Esiste poi una sezione cinematografica che raccoglie le trascrizioni di alcuni film prodotti a partire dagli anni '30 del secolo scorso. In generale, il Corpus copre un arco temporale che va dal XX secolo all'inizio del XXI; inoltre, la provenienza geografica delle registrazioni è piuttosto varia: tra le città più significative si citano Mosca, San Pietroburgo, Saratov, Ekaterinburg, Voronež e Novosibirsk[160].

Anche l'*Ustnyj korpus* presenta annotazione metalinguistica, indicante la sfera sociale (pubblica, privata, cinematografica), il tipo di discorso (intervista, colloquio, conversazione, dialogo breve), la tematica (vita privata, medicina e salute, politica, vita pubblica, ecc.), lo stile (neutrale, ufficiale, ecc.), il luogo, l'anno di creazione del discorso e le caratteristiche del pubblico (numero di spettatori, età, livello di preparazione, ecc.). Infine, questo sotto-corpus è annotato morfologicamente e semanticamente, così come tutti gli altri sotto-corpora di *NKRJa*.

Nella ricerca condotta è stato utilizzato non solo un corpus di lingua scritta, ma anche uno di lingua parlata, l'*Ustnyj korpus*, poiché il *netspeak* presenta diverse caratteristiche che lo rendono simile alla lingua orale, pertanto uno strumento costituito da trascrizioni di parlato potrebbe dare interessanti risultati.

4.2.2 Il *Russian Web Corpus*, il *Timestamped JSI web corpus* e *Sketch Engine*

[158] Si veda la sezione *Osnovnoj korpus tekstov* al link: https://ruscorpora.ru/new/corpora-structure.html.
[159] Questi sono i dati riportati nel sito di *NKRJa*, aggiornati al 23 agosto 2021. Per maggiori informazioni si veda la sezione *Statistika korpusa*: https://ruscorpora.ru/new/corpora-stat.html.
[160] Per informazioni più specifiche riguardanti questo corpus si veda (Grišina, Savčuk, 2009).

Il *Russian Web Corpus*, comunemente abbreviato *ruTenTen*, è un web corpus di lingua russa, appartenente alla famiglia dei corpus *TenTen* e disponibile nel sito di Sketch Engine. Attualmente i corpora *TenTen* comprendono oltre 40 lingue[161] e per quanto concerne le lingue slave esistono di bulgaro, ceco, polacco, russo, slovacco, sloveno ed ucraino; per la ricerca sono stati adoperati il *ruTenTen11* del 2011, contenente 14,5 miliardi di token, e il *ruTenTen17* del 2017, contenente oltre 9 miliardi di token.

Questi corpora sono creati utilizzando i medesimi criteri e strumenti: in particolare, i dati testuali vengono raccolti mediante il già citato web crawler *SpiderLing* (§ 3.3), successivamente ripuliti e filtrati con il software *jusText*, tokenizzati, dedupicati con *Onion* e infine lemmatizzati ed annotati per parti del discorso (*POS-tagging*).

L'altro corpus utilizzato per la ricerca, il *Timestamped JSI web corpus 2014-2021 Russian*, appartiene alla famiglia di corpora *Timestamped JSI web corpus*[162], nata grazie al contributo del *Jozef Stefan Institute* in Slovenia. Al momento ricomprende 18 lingue tra cui alcune lingue slave, come il ceco, il polacco, il russo e il serbo; inoltre, i corpora presentano annotazione morfosintattica e diacronica.

Mediante Sketch Engine è possibile consultare varie versioni del *Timestamped JSI web corpus* di lingua russa; quella utilizzata per la ricerca è il già citato *Timestamped JSI web corpus 2014-2021 Russian*. Secondo le informazioni riportate nel sito di Sketch Engine[163], questo corpus, che ad oggi consta di 5,7 miliardi di token, è costantemente arricchito di nuovi testi, circa 100 milioni di parole ogni mese, e viene annotato per parti del discorso.

Sketch Engine offre numerose risorse, tra cui la possibilità di ricercare combinazioni di parole o di frasi e le loro corrispettive collocazioni, grazie ad uno strumento chiamato *Word Sketch*[164]. Un esempio a tal proposito è rappresentato dalla *Figura 7*, che illustra il Word Sketch del sostantivo russo селфи (*selfie*) in *ruTenTen17*.

[161] Per vedere la lista completa delle lingue dei *TenTen* corpora presenti su Sketch Engine si veda al link: https://www.sketchengine.eu/documentation/tenten-corpora/.
[162] Per maggiori informazioni si veda al link: https://www.sketchengine.eu/jozef-stefan-institute-newsfeed-corpus/.
[163] Per maggiori informazioni sul Timestamped Russian corpus si veda al link: https://www.sketchengine.eu/timestamped-russian-corpus/.
[164] Per maggiori informazioni si veda al link: https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/.

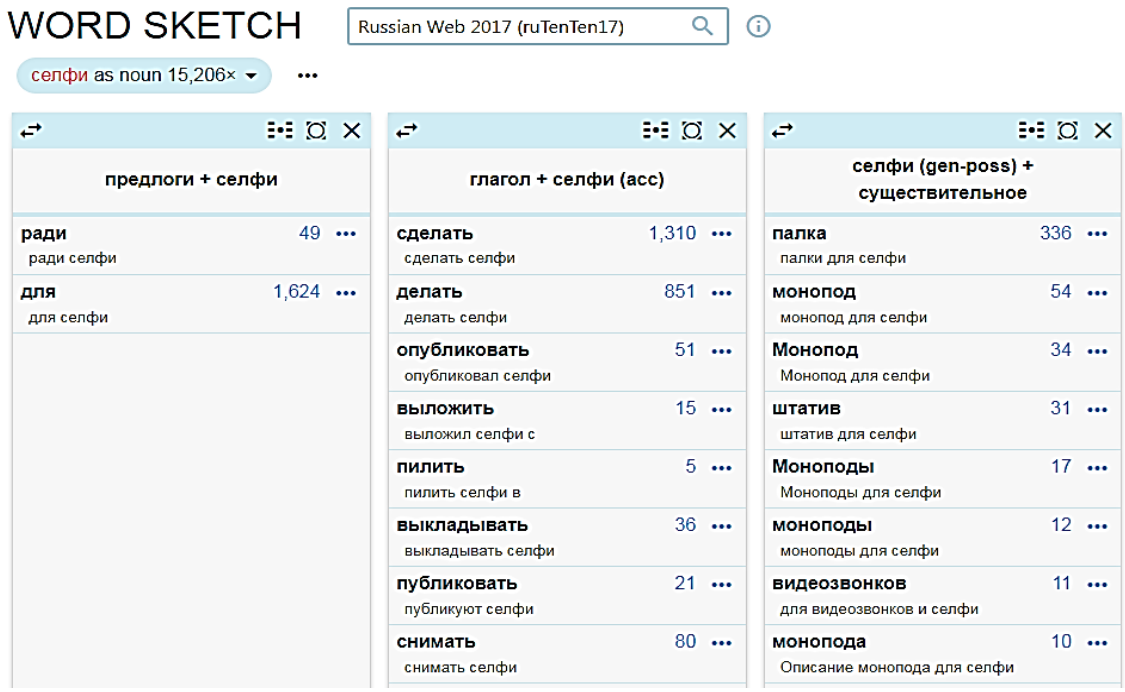

Figura 7. Il Word Sketch della parola *selfie* in *ruTenTen17*.

Com'è possibile osservare, i risultati della ricerca sono suddivisi in differenti sezioni, che in questo caso rappresentano le combinazioni "Preposizione + *selfie*", "Verbo + *selfie* (al caso accusativo)" e "*selfie* (al caso genitivo) + sostantivo"; inoltre, le cifre in blu indicano la frequenza d'uso registrata nel corpus.

Tuttavia, questa funzione per la lingua russa presenta dei problemi: *selfie* in russo è un sostantivo indeclinabile, pertanto i risultati della ricerca mostrano implicitamente tutti i casi grammaticali del sostantivo. Al contrario, se viene inserito nella *query* di ricerca il lemma declinabile апгрейд ('upgrade'), i risultati ottenuti riguardano solo il nominativo singolare. Facendo il medesimo esperimento, utilizzando però un corpus inglese *enTenTen20* e inserendo un sostantivo con il plurale irregolare, come la parola inglese "mouse", i risultati riportano i *Word Sketch* sia del singolare "mouse", sia del plurale "mice". Nella *Figura 8* sono illustrati i due esempi appena riportati:

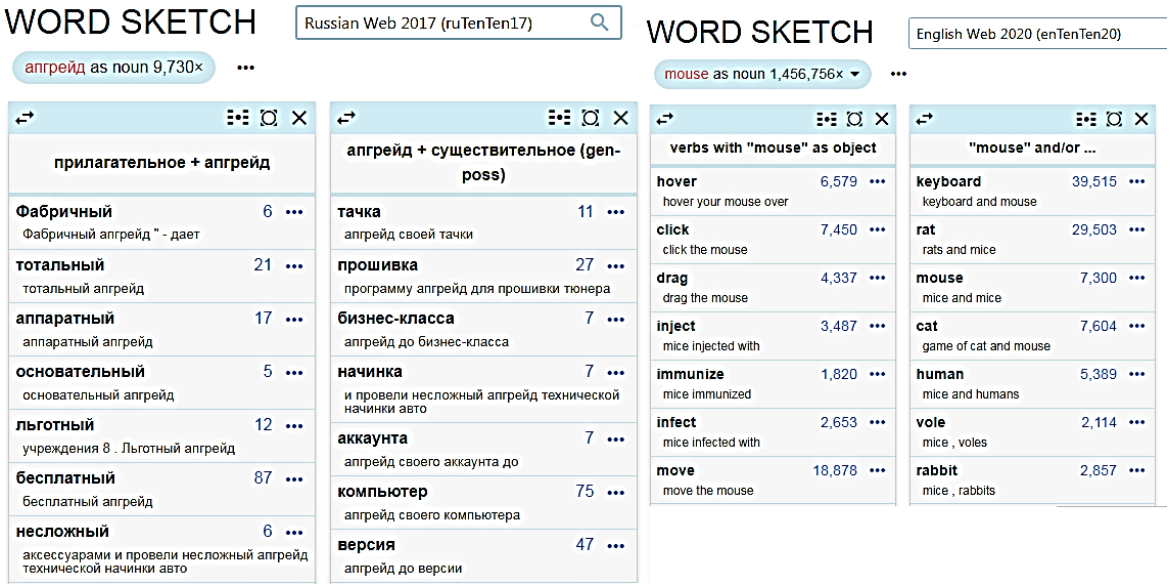

Figura 8. I risultati della ricerca del sostantivo declinabile апгрейд (‘upgrade’) e del sostantivo irregolare inglese *mouse*.

Si potrebbe ipotizzare che il sostantivo russo апгрейд sia presente nel corpus solo al nominativo singolare; tuttavia, questa possibilità è presto scartata con una semplice ricerca del sostantivo nella sezione *Wordlist* di Sketch Engine che mostra la frequenza del sostantivo nei vari casi, così come illustrato nella sottostante *Figura 9*.

Per un’ulteriore verifica, sono state effettuate altre due ricerche su *Word Sketch*: il sostantivo любовь (‘amore’) ed il verbo принимать (tra i principali significati ‘prendere, accettare, assumere’). In questo caso i risultati proponevano sia esempi con i diversi casi grammaticali del sostantivo, sia con varie forme coniugate del verbo e addirittura forme del suo correlato perfettivo.

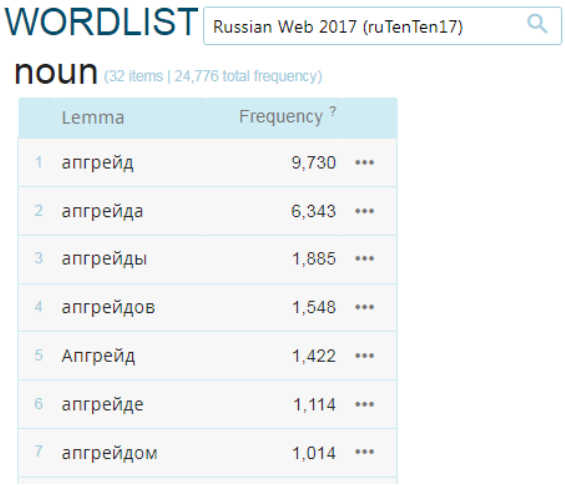

Figura 9. La frequenza del sostantivo апгрейд (*upgrade*) nella sezione *Wordlist* di Sketch Engine.

È chiaro che, sebbene sia una risorsa molto utile, per quanto concerne la lingua russa sussistono ancora dei problemi legati, forse, al carattere flessivo della lingua.

Altre interessanti funzioni offerte da Sketch Engine sono: la ricerca di sinonimi, antonimi e di parole simili mediante *Thesaurus*[165]; la possibilità di confrontare due parole simili attraverso la loro collocazione d'uso con *Word Sketch Differences*[166], di verificare la traduzione di alcune parole attraverso i corpora paralleli[167], di generare liste di parole o espressioni (le cosiddette *multi-word expressions*) sulla base della loro frequenza d'uso con *Wordlist*[168] e *N-grams*[169], di estrarre le parole chiave e la terminologia tipica di un determinato settore o ambito d'interesse[170] e, infine, di identificare i neologismi e condurre un'analisi diacronica della lingua[171].

Tutte queste funzioni rendono Sketch Engine uno strumento utile non solo per i linguisti e i lessicografi, ma anche per gli studenti e gli insegnanti, i traduttori o gli storici della lingua.

4.2.3 I dizionari cartacei ed online

I dizionari sono uno strumento indispensabile per indagare il significato dei lemmi di una lingua, soprattutto straniera. Per condurre la ricerca sono state utilizzate due differenti tipologie di dizionario: la quarta edizione del dizionario cartaceo russo-italiano Zanichelli *il Kovalev* e due risorse online, *Vikislovar'*[172] e *Academic.ru*[173].

Vikislovar' è un sito che permette di consultare dizionari e tesauri in più di 600 lingue, alla cui creazione può potenzialmente partecipare qualsiasi utente, al pari di *Wikipedia*. Nato nel 2004, ad oggi include più di 1 milione di parole. Digitando come *query* di ricerca la radice di una parola o un lemma, appaiono come risultati non solo le

[165] La funzione *Thesaurus* di Sketch Engine è disponibile al link: https://www.sketchengine.eu/guide/thesaurus-synonyms-antonyms-similar-words/.
[166] La funzione *Word Sketch Differences* di Sketch Engine è disponibile al link: https://www.sketchengine.eu/guide/word-sketch-difference-compare-words/.
[167] La funzione *Parallel concordance* di Sketch Engine è disponibile al link: https://www.sketchengine.eu/guide/parallel-concordance-searching-translations/.
[168] La funzione *Wordlist* di Sketch Engine è disponibile al link: https://www.sketchengine.eu/guide/wordlist-frequency-lists/.
[169] La funzione *N-grams* di Sketch Engine è disponibile al link: https://www.sketchengine.eu/guide/n-grams-multiword-expressions/.
[170] Per maggiori informazioni sulla possibilità di ricercare parole chiave o estrarre la terminologia tipica: https://www.sketchengine.eu/guide/keywords-and-term-extraction/.
[171] La funzione per ricercare neologismi e fare analisi diacroniche su Sketch Engine è disponibile al link: https://www.sketchengine.eu/guide/trends/.
[172] Sito ufficiale di *Vikislovar'*: https://ru.wiktionary.org/wiki.
[173] Sito ufficiale di *Academic.ru*: https://academic.ru/.

parole derivate relative al lemma ricercato, ma anche quelle composte. Inoltre, è solitamente indicata l'intera declinazione dei sostantivi e degli aggettivi, nonché la coniugazione dei verbi.

Academic.ru è un sito che raccoglie enciclopedie e dizionari riguardanti specifiche discipline o ambiti, come la medicina, la biologia, la geografia, la geologia, la filosofia, la sociologia, la religione, la storia, le scienze naturali, la letteratura, la musica, la cucina, l'economia, la tecnologia e l'informatica. Inoltre, contiene dizionari etimologici e bilingui.

Sono state selezionate queste due differenti tipologie di dizionario per condurre un'ulteriore analisi comparativa tra una risorsa tradizionale, come il dizionario cartaceo, e due risorse online. In questa ricerca si tenterà di stabilire quale delle due sia più idonea per indagare lo slang giovanile ed il linguaggio giovanile contemporaneo.

4.3 Le modalità d'analisi

In prima istanza è stato necessario individuare e selezionare i lessemi idonei all'analisi: nel motore di ricerca russo *Yandex*[174] sono state inserite le seguenti domande di ricerca: *Sleng molodëži 2021*, *Molodëžnij sleng 2021*, *Molodëžnij žargon*, *Russkij žargon molodëži*, *Modnye slova 2021*, *Samye populjarnye slova 2021*[175].

Sulla base dei risultati di ricerca sono stati consultati numerosi siti, blog e dizionari online; è possibile prendere visione della lista completa delle risorse consultate, nella sezione *Lista di dizionari e siti consultati per l'individuazione dei lessemi da analizzare*, disponibile in Appendice a pagina 152 del presente lavoro.

Incrociando la frequenza dei risultati di ricerca è stata stilata una lista di oltre 150 parole, alcune delle quali successivamente scartate perché ritenute inappropriate, volgari o per altri motivi; in seguito a questa ulteriore selezione, è stata stilata una seconda lista, quella definitiva, intitolata *Indice dei lessemi analizzati*, consultabile anch'essa in Appendice a pagina 154.

Dopo aver selezionato i lessemi, è stato necessario creare una tabella di analisi, per riportare in modo schematico, e al tempo stesso dettagliato, i dati ottenuti dall'analisi dei lessemi. In questo paragrafo viene proposto un esempio in *Tabella 4*,

[174] Il link del motore di ricerca russo *Yandex*: https://yandex.com/.
[175] Rispettivamente 'Lo slang dei giovani 2021', 'Slang giovanile 2021', 'Gergo giovanile', 'Gergo russo dei giovani', 'Le parole alla moda 2021', 'Le parole più popolari del 2021'.

con i risultati della ricerca relativa alla parola трэш e la sua variante треш (*trash* ‘spazzatura’, ‘qualcosa di volgare’; inoltre, sono stati analizzati gli aggettivi derivati трешевый, трешовый, трэшевый, трэшовый e i sostantivi трэшер, трешер.

Come si può osservare, nella *Tabella 4* viene indicato il genere e, nel caso si tratti di un sostantivo, se è declinabile o invariato; vengono poi elencate le eventuali parole derivate e composte, l’etimologia del termine, il tipo di interferenza linguistica (prestito di necessità, di lusso, calco, calco parziale, ecc.) e se esiste o meno un termine corrispondente in russo.

La sezione successiva, evidenziata con un colore differente, è dedicata all’analisi dei dati per mezzo di dizionari e corpora: viene indicato se il lemma è presente nel dizionario cartaceo e nei due dizionari online, se è presente in *NKRJa* nella sezione *Osnovnoj korpus* e nell’*Ustnyj korpus*; infine, sono riportati i dati relativi ai corpora *ruTenTen11* e *ruTenTen17*, le concordanze e gli esempi d’uso. La parte più importante è quella dedicata alle annotazioni personali, in cui sono riportate le osservazioni significative emerse nel corso dell’analisi.

Infine, per ogni analisi è indicato il giorno in cui questa è stata condotta; nell’esempio riportato l’analisi è datata 11 Marzo 2021.

Треш, трэш		Dati aggiornati all’11/03/2021
Significato e definizione in russo	Trash, spazzatura, genere trash	1) Жанр кино, относящийся к фильмам ужасов, в котором смакование сцен насилия и страха достигает гиперболических размеров, доходя до пародии на сам стиль. 2) Стиль в рок-музыке, разновидность хэви-металл, характерен агрессивными гитарными соло и неистовым хрипящим вокалом. Fonte: *Academic.ru.*
Genere e declinazione	Sostantivo Maschile Forte, I declinazione	
Parole derivate o composte	Трешевый, трешовый, трэшевый, трэшовый, трэшер, трешер	
Etimologia del termine	Inglese: *Trash*	**Trash (nome):** tardo XIV sec., “oggetto di scarsa utilità o valore, rifiuti, scorie”, forse da una fonte scandinava (confronta *tros* “spazzatura, foglie cadute e rametti”, dialetto norvegese *trask* “legname, spazzatura, bagagli”, svedese *trasa* “stracci, brandelli”), di origine sconosciuta.

		Fonte: www.etymonline.com[176]
Tipo di interferenza linguistica	Prestito di lusso	
Corrispettivo in russo	Мусор, отбросы, грубый, вульгарный,	
Dizionario cartaceo/online	Cartaceo: No	Online: Sì (in entrambi i dizionari).
NKRJa Osnovnoj korpus	**Треш**: 10 volte. **Трешер**: No. **Трешевый**: 2 volte. **Трешовый**: 1 volta.	**Трэш**: 72 volte. **Трэшер**: No. **Трэшевый**: 8 volte. **Трэшовый**: 1 volta.
NKRJa Ustnyj korpus	**Треш**: 2 volte. **Трешер**: No. **Трешевый**: No. **Трешовый**: 1 volta.	**Трэш**: 2 volte. **Трэшер**: No. **Трэшевый**: No. **Трэшовый**: No.
ruTenTen11	**Треш**: 7.744 volte. **Трешер**: 471 volte. **Трешевый**: 642 volte. **Трешовый**: 450 volte.	**Трэш**: 10.425 volte. **Трэшер**: 379 volte. **Трэшевый**: 1.118 volte. **Трэшовый**: 380 volte.
ruTenTen17	**Треш**: 5.097 volte. **Трешер**: 276 volte. **Трешевый**: 365volte. **Трешовый**: 496volte.	**Трэш**: 6.584 volte. **Трэшер**: 378 volte. **Трэшевый**: 720 volte. **Трэшовый**: 245 volte.
Concordanze	Откровенный трэш: 67 volte. Настоящий трэш: 66 volte. Verbo играть + трэш: 50 volte.	Трэш прически: 339 volte. Трэш одежда: 222 volte. Трэш метал(л): 60 volte.
Frase come esempio d'uso *Traduzione*	В 1999 на экраны вышел научно-фантастический **трэш** «Beowulf», не без остроумия озаглавленный нашими прокатчиками как «Биоволк». Это совсем не **трэш**, и не модная тематика, это сами молодые писатели навязали своё русское, национальное видение мира. Фильм полный **трэш**. Если у вас есть свободные полтора часа, и вы не знаете чем себя занять, то думаю можно посмотреть этот фильм.	Nel 1999 uscì sugli schermi il film **trash** fantascientifico "Beowulf", ma con poca arguzia fu intitolato "Biovolk" dai nostri distributori cinematografici. Questa non è spazzatura, e non è una tematica alla moda, gli stessi giovani scrittori hanno imposto la propria visione nazionale russa del mondo. Il film è un completo trash. Se avete un'ora e mezza libera e non sapete con che cosa tenervi occupati, allora penso che questo film si possa guardare.

[176] Tutte le etimologie dei vocaboli inglesi sono tratte dal sito www.etimonline.com. Per maggiori informazioni sull'etimologia della parola *trash* si veda al link: https://www.etymonline.com/word/trash.

<table>
<tr>
<td>Annotazioni personali</td>
<td>1) Secondo quanto riporta Vikislovar’ esistono due differenti declinazioni del sostantivo трэш con differente accento e differente desinenza del caso strumentale singolare; inoltre, entrambe hanno solo la forma del singolare (Tabella A e B a destra).
Al contrario, secondo quanto riportato da Vikislovar’, треш ha una sola declinazione, che in questo caso presenta anche il plurale (si veda la sottostante Tabella C).
<table>
<tr><td colspan="3">Tabella C</td></tr>
<tr><td>Падеж</td><td>ед. ч.</td><td>мн. ч.</td></tr>
<tr><td>Им.</td><td>тре́ш</td><td>тре́ши</td></tr>
<tr><td>Р.</td><td>тре́ша</td><td>тре́шей</td></tr>
<tr><td>Д.</td><td>тре́шу</td><td>тре́шам</td></tr>
<tr><td>В.</td><td>тре́ш</td><td>тре́ши</td></tr>
<tr><td>Тв.</td><td>тре́шем</td><td>тре́шами</td></tr>
<tr><td>Пр.</td><td>тре́ше</td><td>тре́шах</td></tr>
</table>
</td>
<td>
<table>
<tr><td colspan="3">Tabella A</td></tr>
<tr><td>Падеж</td><td>ед. ч.</td><td>мн. ч.</td></tr>
<tr><td>Им.</td><td>Трэ́ш</td><td>—</td></tr>
<tr><td>Р.</td><td>трэ́ша</td><td>—</td></tr>
<tr><td>Д.</td><td>трэ́шу</td><td>—</td></tr>
<tr><td>В.</td><td>трэ́ш</td><td>—</td></tr>
<tr><td>Тв.</td><td>трэ́шем</td><td>—</td></tr>
<tr><td>Пр.</td><td>трэ́ше</td><td>—</td></tr>
</table>
<table>
<tr><td colspan="3">Tabella B</td></tr>
<tr><td>Падеж</td><td>ед. ч.</td><td>мн. ч.</td></tr>
<tr><td>Им.</td><td>Трэш</td><td>—</td></tr>
<tr><td>Р.</td><td>трэша́</td><td>—</td></tr>
<tr><td>Д.</td><td>трэшу́</td><td>—</td></tr>
<tr><td>В.</td><td>трэш</td><td>—</td></tr>
<tr><td>Тв.</td><td>трэшо́м</td><td>—</td></tr>
<tr><td>Пр.</td><td>трэше́</td><td>—</td></tr>
</table>
</td>
</tr>
</table>

Tabella 4. Analisi delle parole трэш, треш e i loro derivati.

Nei prossimi paragrafi saranno illustrate le principali osservazioni e tendenze emerse nel corso della ricerca. I risultati saranno presentati nel seguente modo: in prima istanza verranno riportate le peculiarità riguardanti la trascrizione dei vocaboli selezionati e, in particolare, l’esistenza di differenti trascrizioni dello stesso lessema; la seconda parte dell’analisi è di natura morfologica e riguarda la russificazione dei prestiti stranieri, i processi di creazione di nuove parole ed i processi derivazionali dei prestiti.

4.4 Le modalità di trascrizione dei prestiti stranieri

I prestiti stranieri, in questo caso di origine inglese, vengono trascritti dai caratteri latini a quelli cirillici; solitamente, le parole sono trascritte così come vengono pronunciate e non come vengono scritte, in una sorta di trascrizione fonetica della lingua. La principale problematica per i russi sta nel fatto che la lingua inglese, a differenza del russo, non è una lingua fonetica: le parole, infatti, non si pronunciano così come sono

scritte[177], tant'è che gli stessi madrelingua inglesi riscontrano numerose difficoltà nell'apprendere a scrivere correttamente in inglese.

Ad esempio, in inglese sono omofone le seguenti coppie di sostantivi: *chute – shoot* (/ˈʃuːt/), *beauty – booty* (/ˈbuːti/), *queue – coo* (/ˈkuː/) e *dual – jewel* (/ˈdʒuːəl/). Inoltre, esistono parole con il medesimo nesso consonantico, pronunciate in maniera differente, come quelle con il nesso consonantico "ough" che viene pronunciato: *though* (/ðəʊ/), *tough* (/tʌf/), *cough* (/kɒf/), *through* (/θruː/), *nought* (/nɔːt/) e così via.

Tuttavia, a livello di terminologia e definizione non si può parlare propriamente di "trascrizione fonetica", descritta in Treccani come "un sistema di rappresentazione grafica dei foni di una lingua realizzata attraverso specifici alfabeti, elaborati appositamente, solo in parte coincidenti con i simboli della scrittura corrente, utilizzati prevalentemente in seno alle scienze linguistiche"[178], in quanto, come vedremo più avanti, la trascrizione in cirillico dei lessemi dall'inglese non si avvale dell'alfabeto IPA (International Phonetic Alphabet), ed è ben lungi dall'essere un metodo di trascrizione scientifico. Al tempo stesso non si può nemmeno parlare di traslitterazione vera e propria, se per traslitterazione si intende una "trascrizione di un testo secondo un sistema alfabetico diverso dall'originale. La traslitterazione non mira tanto a dare un'interpretazione fonetica di un testo o a facilitarne la lettura quanto a riprodurre l'originale, lettera per lettera"[179].

Sulla base dei miei studi sono propensa a proporre una nuova denominazione, in grado di raccogliere e coniugare le caratteristiche di entrambi i termini, ossia "traslitterazione fonetica". In relazione alla lingua russa questa definizione, a mio avviso, è la più adatta, in quanto i termini inglesi vengono sì traslitterati, ma sulla base della loro pronuncia.

Dopo questa precisazione terminologica, verranno qui a seguito proposte le osservazioni emerse nel corso della ricerca: in particolare, le modalità di trascrizione in

[177] Per ulteriori informazioni si consultino degli interessanti articoli a riguardo, reperibili ai link: https://officinamagazine.it/perche-linglese-non-si-pronuncia-come-si-scrive/ e https://www.ilpost.it/2021/09/09/inglese-pronuncia-trascrizione/.

[178] Definizione di trascrizione fonetica nel sito ufficiale dell'Enciclopedia Treccani: https://www.treccani.it/enciclopedia/trascrizione-fonetica_(Enciclopedia-dell'Italiano)/.

[179] Definizione di traslitterazione nel sito ufficiale dell'Enciclopedia Treccani: https://www.treccani.it/enciclopedia/traslitterazione/. Nella medesima pagina viene successivamente precisato: "In questo la traslitterazione si differenzia dalla trascrizione fonetica, che serve a rappresentare, con maggiore o minore precisione secondo i casi, una pronuncia, i suoni dunque e non le lettere alfabetiche".

russo dei prestiti inglesi e la presenza di differenti modalità di trascrizione dello stesso lemma.

4.4.1 Utilizzo di Э (*e* dura) e di E (*e* dolce)

In russo esistono due differenti grafemi per rappresentare la vocale *e*; in particolare, Э (/ɛ/) indica una vocale dura, pronunciata aperta, mentre E (/je/ o /ʲe/), indica una "vocale debole, che addolcisce la consonante che la precede" (Cevese, Dobrovolskaja, Magnanini, 2000: 3).

La presenza di queste due vocali nella lingua russa si traduce nella coesistenza di trascrizioni differenti per la medesima parola inglese e questo, secondo le analisi condotte, è uno dei fenomeni più frequenti, individuato nelle seguenti parole:

1	Апгрейд, апгрэйд; апгрейт, апгрэйт; апгрейдер, апгрэйдер; апгрейдинг, апгрэйдинг; апгрейдный, апгрэйдный; апгрейдовый, апгрэйдовый; апгрейденный, апгрэйденный.
2	Апгрейдить, апгрэйдить; апгрейдиться, апгрэйдиться; апгрейднуть, апгрэйднуть; апгрейтить, апгрэйтить; апгрейдировать, апгрейдировать.
3	Гейм, гэйм; гейминг, гэйминг; геймингoвый, гэймингoвый; геймер, гэймер; Геймерский, гэймерский; геймить, гэймить; геймиться, гэймиться; геймплей, гэймплей; геймплэй, гэймплэй; геймплейный, гэймплейный; геймификация, гэймификация.
4	Инфлюенсер, инфлюэнсер.
5	Крейз, крэйз; крейзи, крэйзи; крейзер, крэйзер; крейзинг, крэйзинг; крейзанутый, крэйзанутый; крейзовый, крэйзовый.
6	Мейнстрим, мэйнстрим; мейнстримовый, мэйнстримовый; мейнстримный, мэйнстримный; мейнстримовский, мэйнстримовский; мейнстриминг, мэйнстриминг.
7	Нетикет, нэтикет.
8	Селфи, сэлфи; селфить, сэлфить; селфиться, сэлфиться; селфи-камера, сэлфи-камера; селфи-палка, сэлфи-палка; селфи-стик, сэлфи-стик.
9	Треш, трэш; трешер, трэшер; трешевый, трэшевый; трешовый, трэшовый.
10	Фейк, фэйк; фейковый, фэйковый; фейк-новости, фэйк-новости; фейк-ньюс, фэйк-ньюс; фейкньюс, фэйкньюс.
11	Фейспалм, фэйспалм; фейспалмить, фэйспалмить.
12	Флешмоб, флэшмоб; флеш-моб, флэш-моб; флешмобер, флэшмобер; флешмоббер,

	флэшмоббер; флеш-мобер, флэш-мобер; флеш-мобер, флэш-моббер.
13	Хейтер, хэйтер; хейтерить, хэйтерить; хейтить, хейтерс; хейтерство, хэйтерство; хейтинг, хэйтинг; хейтерский, хэйтерский.
14	Хештег, хэштэг, хэштег.
15	Шейм, шэим; шейминг, шэйминг.

Tabella 5. Parole che presentano doppia trascrizione con vocale ʻe' dura e dolce.

È possibile consultare nel dettaglio i dati delle singole analisi nella sezione *Analisi dei dati* a pagina 158; in questa sede ci limiteremo a riassumere le principali tendenze riscontrate.

L'ipotesi iniziale è che venga utilizzata una vocale, piuttosto che l'altra, sulla base del fono[180] di partenza della lingua inglese; a tal fine, oltre all'analisi dei dati, è stata condotta un'analisi comparativa tra la traslitterazione fonetica delle parole inglesi e quella dei prestiti russi.

Dall'analisi è emerso che 12 volte su 15 (corrispondente all'80%) le parole sono trascritte utilizzando ʻe' dolce: in particolare, si fa riferimento alle parole инфлюенсер (*influencer*) e нетикет (*netiquette*), al sostantivo апгрейд e ai relativi verbi derivati, ai lessemi гейм (*game*), крейз (*craze*), мейнстрим (*mainstream*), селфи (*selfie*), фейспалм (*facepalm*), фейк (*fake*), флешмоб (*flash mob*), хейтер (*hater*) e ai corrispettivi derivati e composti. Al contrario, vengono prevalentemente scritte con ʻe' dura le parole трэш (*trash*), хэштег (*hashtag*) e шэйм (*shame*); tuttavia, com'è possibile osservare dai rilevamenti fatti, non si tratta di una maggioranza schiacciante, ma di una predominanza di questa tendenza.

Come anticipato in precedenza, l'ipotesi di partenza era che esistesse una correlazione tra il fono della parola inglese e l'utilizzo di una determinata vocale nel prestito russo; per confermare o confutare tale ipotesi è stata creata una tabella, riportata

[180] Nel dizionario Treccani il *fono* viene definito come "ogni suono concreto adoperato nel linguaggio, indipendentemente dal suo valore distintivo" (si veda al link: https://www.treccani.it/vocabolario/fono_res-03c72dc1-001d-11de-9d89-0016357eee51/) e non va confuso con il *fonema*, ossia "l'unità fonologica minima di un sistema linguistico, dotata di capacità distintiva" (si veda al link: https://www.treccani.it/vocabolario/fonema/). In altre parole si tratta di fono quando, pur cambiando pronuncia di una parola, sulla base di una pronuncia regionale, dialettale, a causa della cosiddetta "erre moscia", ecc., il significato della parola rimane il medesimo (ad esempio, la consonante "s" in Veneto, realizzata come /s/ oppure come /z/). Al contrario, si parla di fonema quando, cambiando il suono della parola, muta anche il significato della parola stessa (ad esempio, nelle parole *caso* e *vaso* il differente fonema, con valore distintivo, determina una differenza di significato).

sotto, nella quale vengono illustrate le trascrizioni fonetiche dei lessemi oggetto d'indagine. In grigio sono evidenziate le tre parole scritte con ʿe' dura.

Fonema inglese (nell'alfabeto IPA)	**Lemma in inglese e trascrizione fonetica**	**Prestito in russo**
/eɪ/	Craze: /kr**eɪ**z/	Кр**е**йз
	Fake: /f**eɪ**k/	Ф**е**йк
	Facepalm: /ˈf**eɪ**spɑːm/	Ф**е**йспалм
	Game: /g**eɪ**m/	Г**е**йм
	Hater: /ˈh**eɪ**tər/	Х**е**йтер
	Mainstream: /ˈm**eɪ**nstriːm/	М**е**йнстрим
	Shame: /ʃ**eɪ**m/	Ш**э**йм
	Upgrade: /ˌʌpˈgr**eɪ**d/	Апгр**е**йд
/e/	Netiquette: /ˈn**e**tɪket/	Н**е**тикет
	Selfie: /ˈs**e**lfi/	С**е**лфи
/æ/	Flash mob: /ˈfl**æ**ʃ mɒb/	Фл**е**шмоб
	Hashtag: /ˈh**æ**ʃtæg/	Х**э**штег
	Trash: /tr**æ**ʃ/	Тр**э**ш
/ə/	Influencer: /ˈɪnfluənsər/	Инфлю**е**нсер

Tabella 6. Trascrizione fonetica dei lessemi inglesi e la loro realizzazione nella lingua russa.

Come è possibile osservare nella *Tabella 6*, il dittongo /eɪ/ nella maggior parte dei casi viene traslitterato in russo con ʿe' dolce, ma la parola inglese *shame* confuta l'ipotesi iniziale, in quanto viene trascritta utilizzando ʿe' dura.

Inoltre, va scartata anche l'ipotesi secondo cui l'utilizzo di questa vocale dipenda dal numero di sillabe della parola e dalla posizione dell'accento, in quanto il numero di sillabe e la posizione dell'accento nella parola *shame* sono le medesime delle parole *craze*, *game* e *fake*, che tuttavia vengono trascritte con ʿe' dolce. Infine, in russo non sono frequenti parole con la consonante fricativa sorda "ш" seguita da "э": ad esempio, nel dizionario *il Kovalev* non è presente alcuna parola che inizi per "шэ-", mentre sono numerose le parole che iniziano per "ше-".

La medesima cosa avviene per la vocale anteriore semi-aperta e non arrotondata /æ/ che 2 volte su 3, ossia nelle parole *hashtag* e *trash*, viene realizzata in russo mediante ʿe' dura, ma la parola *flash mob* viene trascritta con ʿe' dolce.

Il manuale *Rozental'* in riferimento alla lettera "Э" riporta la seguente regola ortografica:

> 1. После согласных в корне пишется буква *е*: тендер, кафе. Исключения: мэр, пэр, сэр (и производные) и некоторые имена собственные (Мэри, Улан-Удэ и др.).
> 2. После *и* в корне пишется буква е: диета, пиетет.
> 3. После гласных (кроме *и*) в корне пишется преимущественно буква э: поэт, дуэль, маэстро. Но возможно и написание буквы е: проект, реестр[181]. (Rozental', 2010: 32).

Anche in questo caso, però, la regola riportata non è pienamente rispettata negli esempi analizzati.

Riassumendo, l'analisi ha mostrato una netta preferenza per 'e' dolce nelle trascrizioni dei prestiti stranieri, tendenza che va sempre più delineandosi con il passare del tempo. Inoltre, risulta confutata l'ipotesi iniziale di una correlazione tra foni inglesi e grafemi russi, che giustifichi l'utilizzo di una vocale piuttosto che di un'altra.

4.4.2 Utilizzo di *mjagkij snak* (ь) e *tvërdyj snak* (ъ), o omissione del segno

Questo fenomeno è stato osservato una sola volta, nella parola абьюз ("abuso") e nelle varianti абъюз e абюз. Com'è possibile osservare, la prima variante presenta *mjagkij snak* ('segno dolce'), la seconda *tvërdyj snak* ('segno duro), mentre l'ultima non presenta alcun segno.

Questa parola non è presente nel dizionario cartaceo *il Kovalev*, ma è presente nel dizionario online *Vikislovar'* nella variante con *mjagkij snak*; inoltre, nonostante non vi sia attestazione di questa parola in *NKRJa*, è presente in *ruTenTen11*, in *ruTenTen17* e in *Timestamped JSI web corpus 2014-2021 Russian*.

I dati rilevati, consultabili nella sezione *Analisi dei dati* a pagina 166, con particolare riferimento all'analisi 2.1, possono aiutare a far luce circa la corretta trascrizione di questo lessema. Tutti e tre i web corpora confermano la versione fornita dal dizionario *Vikislovar'*: infatti, in *ruTenTen11, ruTenTen17,* e in *Timestamped JSI web corpus 2014-2021 Russian* la variante più frequente è абьюз con *mjagkij snak.* Inoltre, dai dati risulta evidente che questa trascrizione si sia consolidata negli anni: se nel 2011 era ancora piuttosto frequente l'utilizzo di *tvërdyj snak* e in soli 6 casi vi era la

[181] "1. Dopo le consonanti nella radice si scrive la lettera *e*: *tender, kafe*. Eccezioni: *mer*, *per*, *ser* (e derivati) e alcuni nomi propri (*Meri, Ulan-Ude* e altri).
2. Dopo *i* nella radice si scrive la lettera *e*: *dieta*, *pietet*.
3. Dopo le vocali (eccetto *i*) nella radice si scrive prevalentemente la lettera э: *poet*, *duel'*, *maestro*. Ma è possibile anche la scrittura della lettera *e*: *proekt*, *reestr*." La traduzione è mia.

totale omissione di segno, nel 2017 l'utilizzo di *tvërdyj snak* si era sensibilmente abbassato, mentre l'omissione del segno era scomparsa.

Un'ulteriore conferma dell'affermarsi della variante con *mjagkij snak* è riscontrabile nel sostantivo derivato абьюзер (dall'inglese *abuser*, "colui che abusa"), nel verbo абьюзить ("abusare") e nell'aggettivo абьюзный ("relativo all'abuso"); infatti, con il passare del tempo la parola абьюз si è consolidata nella lingua russa, dando origine a parole derivate, scritte in prevalenza quasi assoluta con *mjagkij snak*.

Ritengo che si sia affermata la traslitterazione fonetica con *mjagkij snak* perché in russo sono numerose le parole con la sequenza -бью-[182], quindi è un modello già presente e affermato nella lingua, a differenza di -бъю- che è molto raro. Infatti, l'utilizzo del *mjagkij snak* è regolato da norme precise a livello ortografico; il manuale *Vse pravila russkoj orfografii i punktuacii* riporta: "Ь пишется: 1) в корне слова перед буквами Е, Ё, Ю, Я, И (пьеса, льет, пьющий, рьяный, соловьи)[183]" (Baronova, 2013: 45). A mio avviso, sono queste le motivazioni che giustificano l'affermarsi della variante абьюз in lingua russa.

4.4.3 Nomi composti uniti o separati da un trattino

Un altro elemento emerso nel corso dell'indagine è la presenza di parole composte trascritte in modi differenti; in particolare, queste possono essere scritte attaccate oppure separate da un trattino, più raramente separate da un semplice spazio.

Tale fenomeno è stato riscontrato nelle seguenti parole:

1	Апгрейд, ап-грейд.
2	Бодиарт, боди-арт.
3	Бодипозитив, боди-позитив; бодипозитивный, боди-позитивный.
4	Бодишейминг, боди-шейминг; бодишеймер, боди-шеймер.
5	Кибербуллинг, кибер-буллинг; кибербуллер, кибер-буллер.
6	Офтоп, оф-топ; оффтоп, офф-топ; офтопик, оф-топик; оффтопик, офф-топик.
7	Флешмоб – флэшмоб; флеш-моб – флэш-моб; флешмобер – флэшмобер; флешмоббер – флэшмоббер; флеш-мобер – флэш-мобер; флеш-моббер – флэш-

[182] Una lista di parole contenenti -бью-, disponibile al link: https://wordhelp.ru/contains/.
[183] 'Si scrive *mjagkij snak*: 1) nella radice della parola, dinnanzi alle lettere E, JO, JU, JA, I (*p'esa* 'pièce teatrale', *l'ët* 'egli versa', *p'jušij* 'bevitore', *r'janyj* 'zelante', *solov'i* 'usignoli').

	моббер.
8	Френдзона, френд-зона, френд зона.
9	Чайлдфри, чайлд-фри, чайлд фри.
10	Чилаут , чил-аут, чиллаут, чилл-аут

Tabella 7. Parole scritte attaccate o separate da un trattino.

Le singole analisi si possono consultare nella sezione *Analisi dei dati* a pagina 167; in questa sede ci limiteremo a riportare le principali osservazioni emerse.

La parola inglese *upgrade* viene trascritta in russo come una parola unica, così come accade nella lingua inglese. I dati confermano una schiacciante maggioranza della variante апгрейд, mentre per quanto riguarda la versione ап-грейд, che rappresenta solo lo 0,16% dei casi in *ruTenTen11*, e lo 0,07% in *ruTenTen17*, possiamo ipotizzare che sia un errore di trascrizione, commesso da russofoni che non conoscono o non padroneggiano correttamente la lingua inglese.

La parola боди- (*body*-) merita un discorso a sé, poiché è un interessante caso di studio. Come si può osservare dalla *Tabella 7* soprastante, sono state analizzate tre parole composte con radice боди-: бодиарт, con la variante боди-арт, бодипозитив e боди-позитив, infine бодишейминг e боди-шейминг.

Nonostante siano composti dagli stessi elementi, questi lessemi vengono trascritti in modi differenti: infatti, la parola боди-арт, a differenza della lingua inglese[184], viene trascritta utilizzando il trattino; questa tendenza, che nel 2011 rappresentava il 57,8% delle entrate totali, con il tempo si è consolidata nella lingua russa, tanto che nel 2017 rappresentava ben l'80,7%. Al contrario, i sostantivi бодипозитив[185] e бодишейминг [186] vengono trascritti come un'unica parola, a differenza di quanto accade nella lingua inglese.

[184] Secondo i dizionari consultati, la parola body art in inglese non viene scritta come un'unica parola, ma i due elementi che compongono la parola, body e art, vengono separati da uno spazio. Per maggiori informazioni si consulti il *Collins Dictionary* al link https://www.collinsdictionary.com/it/dizionario/inglese/body-art e il dizionario *Merriam-Webster* al link https://www.merriam-webster.com/dictionary/body%20art.

[185] La parola *body positive* non è presente né nell'*Oxford Learner's Dictionaries*, né nel *Merriam-Webster Dictionary*. Secondo il *Cambridge Dictionary* sono possibili sia la trascrizione *body positive*, sia *body-positive*. Si veda al link: https://dictionary.cambridge.org/it/dizionario/inglese/body-positive.

[186] Secondo il *Merriam-Webster Dictionary* e il *Cambridge Dictionary*, è corretto scrivere sia *body shaming*, sia *body-shaming* (consultabili ai link: https://www.merriam-webster.com/dictionary/body-shaming e https://dictionary.cambridge.org/it/dizionario/inglese/body-shaming).

Anche la parola оффтопик, che solitamente in inglese viene scritta con la preposizione *off* separata dal sostantivo *topic* mediante uno spazio o un trattino[187], in russo viene scritta tutta attaccata. Ciò che d'interessante è emerso dall'analisi è la predilezione per оффтоп[188], la forma abbreviata della parola, praticamente inesistente in inglese; in particolare, in *ruTenTen11* la parola оффтоп viene registrata 13.991 volte, contro le 4.242 volte di оффтопик, mentre in *ruTenTen17* la parola оффтоп viene registrata 8.896 volte, contro le 2.639 volte di оффтопик.

La parola кибербуллинг (*cyberbulling*) è trascritta attaccata, così come avviene in inglese[189], e la parola френдзона (*friend zone*), che in inglese viene trascritta o come un'unica parola, o separata da uno spazio[190], in russo viene scritta quasi con assoluta maggioranza attaccata, mentre la variante separata da uno spazio in russo è rarissima.

Per quanto riguarda la parola *child-free*, è interessante osservare come i *compounds* inglesi ('parole composte') contenenti l'aggettivo *-free* come secondo elemento del composto, sono solitamente scritti con i due elementi della parola separati mediante un trattino; esempi a tal proposito sono le parole *sugar-free*, *tax-free*, *fat-free*, *duty-free*, ecc. e la parola *child-free* non fa eccezione. Nella lingua russa, al contrario dello standard inglese e delle nostre aspettative, si è affermata la parola чайлдфри, scritta attaccata.

Similmente, anche la parola composta *flash mob*, che in inglese viene prevalentemente scritta separata da uno spazio[191], nella lingua russa si è affermata con la trascrizione флешмоб (9.866 occorrenze in *ruTenTen11* e 14.312 in *ruTenTen17*),

---

[187] Secondo il *Collins Dictionary* si scrive *off topic* oppure *off-topic* (disponibile al link: https://www.collinsdictionary.com/it/dizionario/inglese/off-topic).

[188] Questo procedimento in inglese viene definito *back clipping*, espressione che indica quel processo di omissione o troncamento della parte finale di un termine.

[189] La parola *cyberbulling* viene scritta tutta attaccata, secondo i dati riportati dall'*Oxford Learner's Dictionaries* (disponibili al link: https://www.oxfordlearnersdictionaries.com/definition/english/cyberbullying?q=cyberbullying) e dal *Collins Dictionary* (disponibile al link: https://www.collinsdictionary.com/it/dizionario/inglese/cyberbullying).

[190] Secondo il *Collins Dictionary*, si scrive *friend zone* (disponibile al link: https://www.collinsdictionary.com/it/dizionario/inglese/friend-zone), mentre secondo il *Cambridge Dictionary* la parola si può scrivere sia separata da uno spazio, sia attaccata (disponibile al link: https://dictionary.cambridge.org/it/dizionario/inglese/friendzone?q=friend+zone).

[191] Secondo 3 dizionari su 4, la parola si scrive *flash mob*; in particolare, secondo il *Merriam-Webster Dictionary* (disponibile al link: https://www.merriam-webster.com/dictionary/flash%20mob), il *Collins Dictionary* (disponibile al link: https://www.collinsdictionary.com/it/dizionario/inglese/flash-mob) e l'*Oxford Learner's Dictionaries* (disponibile al link: https://www.oxfordlearnersdictionaries.com/definition/english/flash-mob?q=flash+mob). Al contrario, secondo il *Cambridge Dictionary* si scrive tutto attaccato, ossia *flashmob* (disponibile al link: https://dictionary.cambridge.org/it/dizionario/inglese/flashmob).

seguita per frequenza dalla variante con ʿeʾ dura флэшмоб (4.936 occorrenze in *ruTenTen11* e 3.925 in *ruTenTen17*); al terzo e quarto posto per frequenza ci sono le varianti флеш-моб (4.106 occorrenze in *ruTenTen11* e 2.561 in *ruTenTen17*) e флэш-моб (3.638 occorrenze in *ruTenTen11* e 1.959 in *ruTenTen17*).

Un'altra parola, interessante per la sua particolarità, è *chill out*: in inglese questa parola può rappresentare un sostantivo[192], un aggettivo[193] o un *phrasal verb*[194] e in base al suo ruolo all'interno di una frase viene scritta in maniera differente: nel caso si tratti di un sostantivo o di un aggettivo, la parola *chill* e la preposizione *out* vengono separate mediante un trattino, mentre nel caso si tratti di un *phrasal verb* sono separate da uno spazio. A differenza di quanto avviene in inglese, in russo il sostantivo non solo viene prevalentemente scritto tutto attaccato, come un'unica parola, ma viene anche omessa una delle due consonanti liquide laterali "l" di *chill*, viene quindi trascritto чилаут.

Infatti, in *ruTenTen11* la variante чилаут appare più del doppio delle volte rispetto a чиллаут (1.432 volte, contro le 625 volte di чиллаут), mentre in *ruTenTen17* la variante чилаут appare 881 volte, contro le 358 volte di чиллаут. Esiste anche la trascrizione in cui *chill* e la preposizione *out* vengono separate mediante trattino, ma è meno diffusa: infatti, in *ruTenTen11* vengono registrate 827 entrate di чилл-аут, contro le 581 della variante чил-аут, mentre in *ruTenTen17* vengono registrate 280 entrate di чилл-аут, contro le 211 di чил-аут.

Ricapitolando, nella maggior parte dei casi i prestiti russi si differenziano sensibilmente nella grafia dalle parole inglesi; infatti, dall'analisi dei lessemi indicati nella *Tabella 7* è emerso che solo nel 30% dei casi viene rispettata totalmente la grafia

[192] Secondo quanto riportato dall'*Oxford Learner's Dictionaries*, il sostantivo *chill-out* indica "a style of electronic music that is not fast or lively and is intended to make you relaxed and calm" (definizione disponibile al link: https://www.oxfordlearnersdictionaries.com/definition/english/chill-out_1).

[193] Secondo quanto riportato dall'*Oxford Learner's Dictionaries*, l'aggettivo *chill-out* indica qualcosa "intended to make you feel relaxed and calm, especially in an area in a club where quiet music is played" (definizione disponibile al link: https://www.oxfordlearnersdictionaries.com/definition/english/chill-out_3).

[194] Secondo quanto riportato dall'*Oxford Learner's Dictionaries*, il *phrasal verb chill out* significa "to spend time relaxing; to relax and stop feeling angry or nervous about something" (la definizione è disponibile al link: https://www.oxfordlearnersdictionaries.com/definition/english/chill-out_2). Il *Collins Dictionary* riporta "To chill out means to relax after you have done something tiring or stressful" (la definizione è disponibile al link: https://www.collinsdictionary.com/it/dizionario/inglese/chill-out). Inoltre, secondo il *Cambridge Dictionary* un *phrasal verb* è definito come "a phrase that consists of a verb with a preposition or adverb or both, the meaning of which is different from the meaning of its separate parts"; questa definizione è consultabile al link: https://dictionary.cambridge.org/it/dizionario/inglese/phrasal-verb. In generale, quindi, i *phrasal verbs* sono costituiti da due parti, scritte separatamente.

della parola originale inglese. Inoltre, nel caso di оффтоп e чилаут la trascrizione si discosta sensibilmente dall'originale inglese.

4.4.4 Utilizzo della consonante doppia e singola

Nel corso delle analisi sono stati rilevati dei prestiti che in russo vengono traslitterati con la consonante doppia o con la consonante singola, laddove in inglese vi è il raddoppiamento consonantico o la consonante singola.

Tale fenomeno è stato riscontrato nelle seguenti parole:

1	Блогер, блоггер; блогерство, блоггерство; блогерский, блоггерский; блогерный, блоггерный.
2	Офтоп, оффтоп; оф-топ, офф-топ; офтопик, оффтопик; оф-топик, офф-топик.
3	Тролинг, троллинг; тролинговый, троллинговый; тролить, троллить; тролинговать, троллинговать; тролировать, троллировать.
4	Флешмобер, флешмоббер; флеш-мобер, флеш-моббер; флэшмобер, флэшмоббер; флэш-мобер; флэш-моббер.
5	Фоловер, фолловер; фоловинг, фолловинг; фоловить, фолловить.
6	Челендж, челлендж; челенж, челленж; челенджер, челленджер; челенжер, челленжер.
7	Чилаут, чиллаут; чил-аут, чилл-аут.
8	Чилить, чиллить; чилиться, чиллиться.

Tabella 8. Parole scritte con consonante doppia o singola.

Le singole analisi possono essere consultate nella sezione *Analisi dei dati* a pagina 170. Nella lingua inglese esiste una specifica regola ortografica riguardante il raddoppiamento consonantico, chiamata "The doubling rule" o "The 1-1-1 rule": se una parola, costituita da 1 sillaba, contiene 1 vocale e termina con 1 consonante, allora quella consonante in fine di sillaba subisce un raddoppiamento, prima che ad essa vengano aggiunti i suffissi *-ing*, *-ed*, *-er*, *-est*. È interessante osservare se tale modalità di scrittura venga o meno rispettata nei prestiti russi e la prossima sezione tenterà di far luce proprio su questo.

Per quanto concerne la parola inglese *blogger*, sia l'*Osnovnoj korpus* che l'*Ustnyj korpus* di *NKRJa* mostrano una prevalenza numerica della variante con una sola

consonante velare sonora “g”, ossia блогер, sebbene i dati non siano quantitativamente significativi (216 occorrenze nell’*Osnovnoj korpus* e 16 nell’*Ustnyj korpus*) e non vi sia presenza alcuna di lessemi derivati, quali блогерство o блоггерство, блогерский o блоггерский, блогерный o блоггерный, che possano confermare e smentire tale tendenza.

I dati emersi dall’analisi di *ruTenTen11* e *ruTenTen17* sono piuttosto controversi: infatti, in *ruTenTen11* il sostantivo maschile блогер compare 76.308 volte, mentre la variante con la consonante doppia, блоггер, compare 79.782 volte in tutto; in altre parole, contrariamente ai dati emersi da *NKRJa*, è maggiormente affermata la variante con la consonante doppia, sebbene la differenza sia di poco più di 3 mila occorrenze. Dall’analisi di *ruTenTen17*, invece, risulta più frequente la variante con consonante singola блогер, che occorre 55.129 volte in tutto, rispetto alle 30.955 volte della variante блоггер; questi dati, quindi, concordano con quelli presenti in *NKRJa*, ma sono in contrasto con quelli di *ruTenTen11*.

Come controprova è stato utilizzato un terzo web corpus, il *Timestamped JSI web corpus 2014-2021 Russian*: in questo caso emerge una schiacciante maggioranza per блогер (il 98% del totale, equivalente a 315.727 occorrenze, rispetto alle 6.332 della variante блоггер).

Sono stati riscontrati analoghi trend in relazione all’analisi dei lessemi derivati блогерство e блоггерство, блогерский e блоггерский, блогерный e блоггерный: in *ruTenTen17* e nel *Timestamped JSI web corpus 2014-2021 Russian* sono più frequenti le parole derivate con consonante singola, ossia блогерство e блогерский, mentre блогерный e блоггерный non compaiono affatto; al contrario, in *ruTenTen11* sono più frequenti блоггерство e блоггерский con consonante doppia, mentre per quanto concerne la coppia блогерный e блоггерный è più frequente la prima, che occorre 16 volte, rispetto alle 13 della variante con consonante doppia.

Riassumendo, attraverso la ricerca condotta è stato possibile riscontrare un progressivo affermarsi nella lingua russa della forma con consonante singola, ossia блогер, sia come sostantivo maschile, sia come elemento che compone parole derivate, nonostante in inglese la parola *blogger* mostri il raddoppiamento consonantico.

La parola inglese *off topic* è già stata presa in esame nella precedente sezione e si è visto che viene trascritta come un’unica parola oppure con un trattino di separazione

(cfr. analisi a pagina 169): dall'analisi era emersa una chiara preferenza per la grafia оффтоп, ossia per la trascrizione come parola singola e abbreviata, contrariamente alla grafia inglese che predilige la variante *off topic*, talvolta anche *off-topic*. Da questa ulteriore osservazione è emerso che anche in lingua russa, similmente all'inglese, viene utilizzata la doppia consonante fricativa sorda "f" nella trascrizione: tale tendenza è confermata dai dati dell'*Osnovnoj korpus* di *NKRJa* in cui, nonostante i dati siano quantitativamente esigui, vengono registrate 42 occorrenze di оффтоп, 15 di оффтопик, 9 di офф-топ e 2 di офф-топик, contro le 4 di офтоп e 2 di офтопик, mentre le varianti оф-топ e оф-топик non compaiono affatto. Anche in *ruTenTen11* e *ruTenTen17* è stata rilevata una netta maggioranza di оффтоп, che occorre 13.991 volte in *ruTenTen11* e 8.896 volte in *ruTenTen17*, contro le rispettive 1.967 e 1.241 volte della variante офтоп. Analogamente, la parola оффтопик compare 4.242 volte in *ruTenTen11* e 2.639 volte in *ruTenTen17*, contro le rispettive 714 e 477 volte della variante офтопик.

Per riassumere quanto descritto sinora, in russo, così come in inglese, il prestito è prevalentemente scritto tutto attaccato, con la consonante doppia, ma in forma abbreviata, ossia оффтоп.

Le successive analisi hanno mostrato una generale tendenza a mantenere nei prestiti la consonante doppia, lì dove è presente nell'originale inglese. È questo il caso di parole come троллинг e i suoi numerosi derivati, il sostantivo фолловер, il verbo derivato фолловить e il prestito фолловинг, il sostantivo челлендж e il derivato челленджер: tutti questi nomi sono prevalentemente scritti con la consonante doppia, così come avviene in inglese.

La parola *flash mob,* già analizzata precedentemente, è un interessante oggetto di studio. In inglese indica "A group of people who arrange by phone or online to meet suddenly in a public place to do something for a short time"[195] e raramente viene utilizzato il sostantivo *flash mobber*, tanto che se si effettua la ricerca «Flash mobber meaning» su Google, i risultati rimandano quasi tutti alla parola *flash mob*.

Anche in russo il prestito traslitterato флешмоббер non è molto frequente, ma è interessante la preferenza per la trascrizione senza la consonante doppia, ossia

[195] La definizione di *flash mob* secondo il *Collins Dictionary*. Per maggiori informazioni si veda al link: https://www.collinsdictionary.com/it/dizionario/inglese/flash-mob.

флешмобер[196]. Questo è confermato sia dai dati dell'*Osnovnoj korpus* di *NKRJa*, nonostante le cifre siano decisamente esigue (4 occorrenze di флешмобер e nessuna di флешмоббер), sia dai dati presenti in *ruTenTen11*, *ruTenTen17* e nel *Timestamped JSI web corpus 2014-2021 Russian* (флешмобер appare 146 volte in *ruTenTen11*, 48 in *ruTenTen17* e 17 in *Timestamped JSI web corpus 2014-2021 Russian*, contro le rispettive 68, 19 e 0 volte di флешмоббер).

Per quanto riguarda la parola *chill out* (o *chill-out*), già nella precedente sezione era emerso che il prestito in russo si discosta dalla tradizionale ortografia inglese, poiché viene trascritto чилаут, ossia tutto attaccato e con consonante singola. La medesima cosa avviene per il verbo derivato: nella lingua russa, infatti, si è affermata la variante con consonante singola sia del verbo чилить, sia del verbo riflessivo чилиться.

Riassumendo quanto osservato sino ad ora, nel 50%, ossia in блогер e derivati, флешмобер, чилаут e чилить, il prestito si allontana dal tradizionale *spelling* inglese, mentre nel restante 50% dei casi, ossia in оффтоп, троллинг, фолловер e челлендж, il prestito mantiene lo *spelling* con la consonante doppia delle parole originali. È interessante notare che di questi 3 su 4 presentano il raddoppiamento della stessa consonante, ossia della liquida laterale "l": sarebbe interessante condurre una ricerca al riguardo, per verificare se la "l" sia o meno una consonante che nel prestito tende a rimanere raddoppiata. In generale, in russo non è comune l'uso della consonante doppia, pertanto nel passaggio da una lingua all'altra vi è un processo di adeguamento allo standard grafico del russo, così come abbiamo osservato negli esempi sopra riportati.

4.4.5 Alternanza delle consonanti C (s) e З (z)

Nella lingua russa il grafema "c" rappresenta la fricativa alveolare sorda /s/, mentre il grafema "з" rappresenta la fricativa alveolare sonora /z/; in altre parole, queste due consonanti sono in opposizione 'sorda-sonora'. Nel corso dell'analisi è emersa una doppia modalità di trascrizione della parola inglese *dislike*: infatti, esiste la variante con la sorda дислайк e la variante con la sonora дизлайк.

[196] In questa analisi è stata presa in considerazione solo la variante con base *флешмоб-,* in quanto già dalle precedenti analisi 1.12 e 3.7 era emerso che in russo si è affermata la trascrizione con 'е' dolce e scritta tutta attaccata, come unica parola, non separata mediante trattino.

In inglese *dislike* si pronuncia /dɪˈslaɪk/[197], con la fricativa alveolare sorda, pertanto la trascrizione in russo dovrebbe essere дислайк, con la consonante sorda. Al tempo stesso, però, esistono in russo delle precise regole ortografiche che regolano l'uso di prefissi con "с" e "з"; in particolare, il manuale *Spravočnik po russkomu jazyku: orfografija i punktuacija* ('Manuale di lingua russa: ortografia e punteggiatura') riporta la seguente regola ortografica:

> Приставки *без-, воз- (вс-), из-, низ-, раз-, чрез- (через-)* пишутся с буквой *з* перед гласными и звонкими согласными (б, в, г, д, ж, з, л, м, н, р) […] и с буквой *с* перед глухими согласными (к, п, с, т, ф, х, ц, ч, ш, щ)[198] (Rozental', 2010: 51).

Attenendoci a questa regola, il prefisso inglese *dis-* in russo si dovrebbe traslitterare диз-, poiché è un prefisso seguito dalla consonante sonora "l". Nella sezione *Analisi dei dati* a pagina 173 è possibile consultare l'analisi in cui sono riportate tutte le occorrenze delle parole дислайк, дизлайк e i loro derivati; in questa sede ci limiteremo a riportare le osservazioni emerse nel corso dell'analisi.

Dai dati emerge una chiara preferenza per la traslitterazione дизлайк, con la fricativa alveolare sonora: tale tendenza è confermata dai dati di *ruTenTen17* (762 occorrenze) e di *Timestamped JSI web corpus 2014-2021 Russian* (4.954 occorrenze), che presentano qualche occorrenza anche dei verbi derivati дизлайкнуть (11 volte in *ruTenTen17* e 17 nel *Timestamped JSI*) e дизлайкать (20 volte in *ruTenTen17* e 26 nel *Timestamped JSI*). In altre parole, *dislike* ha subito un processo di russificazione e viene trascritta secondo le regole ortografiche russe e non secondo la pronuncia originale inglese; a mio parere, questo è un ulteriore esempio di come nella lingua russa i prestiti non sono trascritti così come si pronunciano o si scrivono in inglese, ma subiscono dei processi di adattamento e russificazione.

4.4.6 Alternanza delle consonanti Д (d) e Т (t)

Nel corso della ricerca è emersa una doppia modalità di trascrizione della parola inglese *upgrade*, già analizzata nel corso di questo capitolo, ossia la variante con la consonante sonora апгрейд e quella con la consonante sorda апгрейт.

---

[197] La pronuncia di *dislike,* secondo quanto riportato dal *Cambridge Dictionary.* è disponibile al link: https://dictionary.cambridge.org/it/dizionario/inglese/dislike.

[198] "I prefissi b*ez-, voz- (vs-), iz-, niz-, raz-, črez- (čerez-)* si scrivono con la lettera *z* prima delle vocali e delle consonanti sonore (*b, v, d, ž, z, l, m, n, p*) […] e con la lettera *s* prima delle consonanti sorde (*k, p, s, t, f, ch, c, č, š, šč*)".

Com'è possibile osservare dalla tabella a pagina 173, nella sezione *Analisi dei dati*, апгрейт occorre troppe poco per essere considerata una variante realmente affermata nella lingua russa (una sola occorrenza nell'*Osnovnoj korpus*¸ 968 occorrenze in *ruTenTen11* e 539 occorrenze in *ruTenTen17*), ma è necessario chiedersi come mai venga occasionalmente trascritta in questo modo.

In russo il grafema "д" rappresenta l'occlusiva alveolare sonora /d/, mentre il grafema "т" rappresenta l'occlusiva alveolare sorda /t/; formano quindi una coppia di consonanti 'sorda-sonora'.

Come riportato in *Sovremennyj russkij jazyk. Fonetika. Orfoepija*:

> В русском языке невозможно употребление звонкого шумного согласного в абсолютном конце слова, что приводит к одной особенности звукового оформления русского слова: во всех словах, заканчивающихся шумным согласным, этот согласный обязательно является глухим.[199] (Malyševa, Rogaleva, 2012: 46)

In altre parole, la sonora /d/ si pronuncia come sorda /t/ in fine di parola, pertanto апгрейд in russo viene pronunciato con l'occlusiva sorda /t/ in fine di parola. È questa la ragione che spiega come mai esista tale variante: i russofoni trascrivono erroneamente la parola inglese *upgrade*, basandosi sulla pronuncia russa della parola. È ancora più interessante notare che accanto al verbo апгрейдить, derivante dal sostantivo апгрейд, esiste anche la variante апгрейтить, derivante dal sostantivo апгрейт, nonostante quantitativamente le sue occorrenze siano poche (44 occorrenze di апгрейт, contro le 3.146 di апгрейд in *ruTenTen11*, mentre in *ruTenTen17* si registrano 25 occorrenze di апгрейт, contro le 2.080 di апгрейд).

Riassumendo, le regole riguardanti la fonetica della lingua russa e la corretta pronuncia delle parole hanno influenzato la trascrizione del prestito che, proprio perché errata, è un interessante oggetto di indagine della lingua.

4.4.7 Alternanza della consonante Ж (ž) e del nesso consonantico ДЖ (dž)

Nella lingua russa non esiste l'affricata post-alveolare sonora /dʒ/, tipica di parole inglesi come *cringe*, *challenge*, *magic*, *John*, *adjective*, *education*, *soldier*, ecc.

Solitamente, questo suono viene trascritto in russo mediante il nesso consonantico "дж" (dž), come nelle parole джаз ('jazz'), Джойс ('Joyce'), джип ('jeep'), бюджет

[199] "In lingua russa non è possibile usare una consonante ostruente in fine di parola e ciò determina una peculiarità fonetica della parola russa: in tutte le parole che terminano con una consonante sonora, questa consonante è necessariamente sorda". La traduzione è mia.

(‘budget’), ecc., tuttavia, nel corso della ricerca sono emerse due differenti modalità di trascrizione dell’affricata post-alveolare sonora /dʒ/.

Tale peculiarità è stata riscontrata nei seguenti prestiti:

1	Криндж, кринж; кринджовый, кринжовый.
2	Челлендж, челленж; челленджер, челленжер.

Tabella 9. Parole con alternanza della consonante Ж (ž) e del nesso consonantico ДЖ (dž).

Le singole analisi possono essere consultate nella sezione *Analisi dei dati* a pagina 174; in questa sede ci limiteremo a riportare le principali osservazioni emerse.

I sostantivi inglesi *challenge* e *challenger* sono traslitterati in russo secondo la tradizionale modalità, ossia mediante il nesso consonantico “дж”: i dati, infatti, mostrano una schiacciante maggioranza delle varianti челлендж e челленджер, in linea con le aspettative; al contrario, *cringe* si discosta dalla norma.

La parola inglese *cringe* si è diffusa molto recentemente: secondo i dati dell’*Accademia della Crusca* è diventata popolare a patire dall’1 Dicembre 2020[200], e questo spiegherebbe l’assoluta mancanza di dati in *NKRJa*, in *ruTenTen11* e *ruTenTen17*. Gli unici dati disponibili sono quelli tratti dal *Timestamped JSI web corpus 2014-2021 Russian*, che mostrano un’unica modalità di trascrizione, ossia кринж (202 occorrenze). L’analisi ha rilevato anche l’esistenza dell’aggettivo derivato кринжовый (48 occorrenze in tutto), mentre кринджовый, scritto con il nesso consonantico “дж” non compare.

A mio parere, questo sostantivo è sin troppo recente e poco affermato, perché possano essere tratte delle conclusioni certe sulla sua trascrizione in russo, tuttavia andrebbe monitorato, per capire se questa trascrizione si consoliderà nel corso del tempo o si adatterà alla tradizionale norma russa di realizzare l’affricata post-alveolare sonora /dʒ/ mediante il nesso consonantico “дж”.

Nei prossimi paragrafi saranno illustrate le considerazioni di natura morfologica emerse dall’analisi: si tratta dei cosiddetti processi di *slovoobrazovanie*¸ ossia di formazione e derivazione delle parole, come la suffissazione, i prestiti e i calchi, le parole composte, le abbreviazioni, gli acronimi, ecc.

[200] La parola *cringe* nel sito dell’*Accademia della Crusca*, disponibile al link: https://accademiadellacrusca.it/it/parole-nuove/cringe/18487.

I dati verranno presentati in base alla categoria morfologica di riferimento, ossia quella dei sostantivi, degli aggettivi e dei verbi.

4.5 I sostantivi: i processi di formazione delle parole, i prestiti e le irregolarità emerse

Tra i lessemi analizzati, i sostantivi sono la categoria morfologica più frequente, in relazione ai quali sono stati osservati numerosi fenomeni: dal semplice prestito traslitterato, alla russificazione dei sostantivi, dalle parole composte importate come prestiti, alle parole composte russificate e trasformate in calchi parziali. Inoltre, sono emerse anche delle "irregolarità", relative a quei sostantivi inglesi plurali traslitterati in russo e declinati come dei regolari sostantivi singolari terminanti in *-s*. Tutto questo verrà illustrato nei prossimi sotto-paragrafi.

4.5.1 Il prestito traslitterato di lessemi stranieri

Il prestito traslitterato dei sostantivi è il fenomeno più frequente, riscontrato in ben 41 casi, la cui lista è consultabile nella sezione *Analisi dei dati* a pagina 175.

Si tratta perlopiù di prestiti di lusso[201], ossia di sostantivi che hanno già un referente lessicale con egual significato nella lingua d'arrivo, ma vengono utilizzati per conferire una sfumatura esotica o prestigiosa al discorso, nei linguaggi tecnici e specialistici o, ancora, da una ristretta cerchia di persone che condivide uno specifico codice linguistico.

Inoltre, questi prestiti sono alla base di importanti processi di *inflectional affixation*[202] e di *derivational affixation*[203]: infatti, i sostantivi non solo vengono regolarmente declinati, mentre in passato si tendeva a lasciare la maggior parte dei

[201] I prestiti di lusso si contrappongono ai prestiti di necessità, ossia quei nuovi referenti lessicali che entrano a far parte di una lingua in seguito, per esempio, ad una scoperta. In altre parole, quando non esiste alcun equivalente nella lingua d'arrivo, si parla di prestito di necessità.

[202] Per *inflectional affixation*, o affissazione di tipo flessivo, si intendono quei processi di formazione delle parole mediante affissione, in particolare mediante suffissi flessivi, che non cambiano la categoria morfologica della parola. Un esempio a tal proposito sono i sostantivi e gli aggettivi regolarmente declinati, oppure i verbi coniugati; i morfemi flessivi non cambiano la categoria morfologica della parola, che rimarrà un sostantivo, un aggettivo o un verbo, a prescindere dal caso, persona, numero, tempo, ecc.

[203] Con *derivational affixation*, o affissazione di tipo derivazionale, si fa riferimento a quei processi di formazione delle parole mediante affissione, in particolare mediante suffissi derivativi, che, a differenza dei suffissi flessivi, cambiano la categoria morfologica della parola. Un esempio a tal proposito sono i sostantivi che danno vita, mediante specifici morfemi derivazionali, ad aggettivi o verbi; in altre parole, danno vita a nuove parole appartenenti ad una categoria morfologica differente rispetto all'originale.

prestiti invariati, ma danno vita anche a nuovi aggettivi e verbi derivati. Tutti questi processi flessivi e derivazionali sono indice di un buon grado di acclimatamento[204] ed integrazione[205] dei prestiti nella lingua d'arrivo e sono un ideale oggetto di studio per indagare la lingua russa della contemporaneità.

L'uso dei prestiti tende a regolarizzarsi con il passare del tempo: infatti, alcuni prestiti che inizialmente mostravano una o più modalità di trascrizione, nel corso del tempo hanno consolidato una trascrizione definitiva, come era già stato osservato nel paragrafo precedente relativamente alla trascrizione dei prestiti; è questo il caso di parole come абьюз, баттхерт, блогер, мейнстрим, трэш, флешмоб (*abuse, butthurt, blogger, mainstream, trash, flash mob*), ecc.

Interessante è il caso della parola селфи: secondo i dati di *ruTenTen11* questa parola nel 2011 occorreva 63 volte al solo nominativo e in tutto, ossia declinata, 679 volte; secondo i dati di *ruTenTen17*, селфи occorre 20.564 volte al nominativo e 20.861 volte in tutto. Sulla base di questi dati si può affermare che inizialmente il sostantivo veniva regolarmente declinato, ma con il passare del tempo si è consolidato nella lingua russa come un sostantivo invariato, non declinabile. Infatti, in russo i sostantivi terminanti in -и sono rari e solitamente sono dei prestiti invariati, come алиби, ассорти, бикини, виски, лобби ('alibi', 'assortimento', 'bikini', 'whisky', 'lobby'), ecc.

Un altro caso interessante è il sostantivo трэш che, secondo quanto riporta *Vikislovar'*, presenta due diverse declinazioni, con accento differente e solo con la forma singolare, in altre parole трэш farebbe parte della categoria dei *nomina singularia tantum*; tuttavia, attraverso l'analisi dei corpora è stato possibile stabilire che трэш viene regolarmente declinato anche al plurale.

4.5.2 Prestiti di *nomina agentis* e *nomina actionis* in *-er*

I *nomina agentis* sono i cosiddetti "nomi d'agente" indicanti l'entità che compie l'azione, mentre i *nomina actionis* sono dei sostantivi che indicano l'azione che si compie. Ogni lingua possiede i propri morfemi derivazionali per la creazione dei *nomina agentis* ed *actionis*: ad esempio, in inglese spesso viene utilizzato il modello

[204] Per acclimatamento si intende "l'uso che ne fa il parlate: tanto più egli familiarizzerà col neologismo, tanto più quest'ultimo risulterà acclimatato" (Gusmani, 1993: 25).
[205] Per integrazione di un prestito si intende "l'influsso esercitato dalla lingua ricevente nello sforzo di adeguare il termine di tradizione straniera alle sue strutture fonematiche, morfologiche, ecc." (Gusmani, 1993: 25)

derivazionale “Verbo + *-er*/*-or* → Sostantivo che indica un agente, un esperiente o uno strumento”.

Ecco alcuni esempi di sostantivi indicanti una persona o un oggetto:

Kill → *killer* (“uccidere → sicario”);

Translate → *translator* (“tradurre → traduttore”);

Amplify → *amplifier* (“amplificare → amplificatore”);

Calculate → *calculator* (“calcolare → calcolatrice”).

Nella lingua russa esistono numerosi morfemi derivazionali autoctoni per la formazione di sostantivi, tra cui -тель, -чик/-щик, -ич, -ок, -ник, ecc. Ad esempio:

Читать → чита-тель. Тот, кто читает (“leggere → lettore, colui che legge”);

Переводить → переводчик. Тот, кто переводит (“tradurre → traduttore”);

Учить → ученик. Тот, кто учить (“imparare → alunno, colui che impara”).

Nel corso dell’analisi sono stati rilevati vari sostantivi inglesi terminanti in *-er*, indicanti *nomina agentis* ed *actionis*, traslitterati in russo ed utilizzati nella lingua d’arrivo come prestiti, molto spesso di lusso. In altre parole, le radici di questi lessemi non vengono adattate alla lingua d’arrivo e trasformate in *nomina agentis* ed *actionis* mediante i suffissi derivazionali tipici russi, ma i sostantivi vengono traslitterati per intero e declinati come sostantivi maschili forti terminanti in *-er*.

I 22 sostantivi individuati sono consultabili a pagina 175 nella sezione *Analisi dei dati*; in 4 casi su 22, ossia nel caso delle parole зумер, спойлер, стайлер e стример, evidenziate in grigio nella sezione *Analisi dei dati*, i lessemi indicano sia un *nomen agentis* che *actionis*.

Il *nomen agentis* зумер (*zoomer*, giovane appartenente alla Generazione Z) non deriva da un verbo, si tratta bensì di un neologismo, indicante una classe generazionale, mentre per quanto concerne la parola хипстер (*hipster*, giovane anticonformista) è probabile che derivi da un verbo, ma non è stato possibile determinare quale; in tutti gli altri casi i sostantivi inglesi in *-er* derivano da verbi.

Nella prossima sezione saranno analizzati alcuni prestiti russificati, adattati cioè alla lingua russa.

4.5.3 Prestiti russificati in *–ств(о)*, indicanti sostantivi neutri astratti

In russo esistono numerosi sostantivi realizzati mediante il seguente modello derivazionale: “Radice di sostantivo/aggettivo + -ств(о) → Sostantivo neutro”.

Questi sostantivi neutri terminanti in -ств(о) hanno differenti significati, tra i principali ricordiamo: un’unione o un gruppo di persone, solitamente piuttosto numeroso (братство ʿfratellanza, confraternitaʾ, профессорство ʿprofessoratoʾ, дворянство ʿnobiltàʾ), un’istituzione o un’organizzazione (агентство ʿagenziaʾ, посольство ʿambasciataʾ, представительство ʿufficio di rappresentanzaʾ), una caratteristica astratta, tipica degli esseri animati, soprattutto umani (богатство ʿricchezzaʾ, изящество ʿeleganzaʾ, жеманство ʿleziosaggineʾ), una persona dotata delle caratteristiche indicate dalla radice della parola (божество ʿdivinitàʾ, высочество ʿmaestàʾ, ничтожество ʿnullitàʾ) e, infine, l’attività lavorativa di una persona (строительство ʿediliziaʾ, животноводство ʿallevamento di bestiameʾ, акушерство ʿostetriciaʾ).

Nel corso dell’analisi sono emersi 6 sostantivi terminanti in -ств(о), la cui radice non è un sostantivo o un aggettivo russo, bensì un prestito traslitterato di un sostantivo animato inglese. È possibile consultare nel dettaglio le occorrenze rilevate nella sezione *Analisi dei dati* a pagina 177; in particolare, quelli analizzati sono sostantivi neutri indicanti una categoria astratta, come блогерство (“un insieme di persone appartenenti alla categoria dei blogger”), геймерство (“un insieme di persone appartenenti alla categoria dei gamer”), лузерство (“un insieme di persone appartenenti alla categoria dei perdenti”), спойлерство (“un insieme di persone che spoilerano”), хейтерство (“un insieme di persone appartenenti alla categoria degli hater”) e хипстерство (“un insieme di persone appartenenti alla categoria degli hipster”), derivanti rispettivamente dai *nomina agentis blogger*, *gamer*, *loser*, *spoiler*, *hater* e *hipster*. Questi sostantivi animati inglesi sono, come si è detto, dei prestiti traslitterati russificati e ciò costituisce un interessante esempio di acclimatamento ed integrazione nella lingua russa, sebbene le loro occorrenze non siano molte. In generale, questo modello derivazionale non è molto diffuso, ma ha buone possibilità di affermarsi nel tempo.

4.5.4 Forme differenti dello stesso sostantivo

Nel corso delle analisi sono emerse varianti differenti dello stesso sostantivo: è questo il caso di крип e крипота e dei sostantivi лайк e лойс.

Крип е крипота derivano dal sostantivo inglese *creep* ed indicano qualcuno o qualcosa che suscita spavento, un sentimento di ansia, disagio o disgusto; l'unica differenza rilevata sta nella sfumatura di significato: крип viene perlopiù utilizzato per indicare una persona, mentre крипота per indicare un oggetto, ma la differenza non è sempre netta.

Il sostantivo крип è già presente da qualche tempo nella lingua russa con due differenti significati: infatti, indica un lento spostamento verso il basso di uno strato della terra[206], oppure una lenta deformazione dei metalli[207]; entrambi questi sostantivi derivano dal verbo inglese *to creep* ("muoversi lentamente, silenziosamente ed attentamente"). A questi due significati, oramai consolidati nella lingua russa, si è aggiunto recentemente questo ulteriore significato, indicante, come si è detto, qualcuno che provoca disagio, disgusto, ecc. Dal sostantivo крип ha successivamente avuto origine l'aggettivo russificato криповый e l'avverbio крипово. Da крипота, invece, è derivato l'aggettivo крипотный. Nella sezione *Analisi dei dati* a pagina 178 è possibile consultare le analisi dei lessemi sopra citati, eccetto le occorrenze del sostantivo крип, poiché non è stato possibile disambiguare i tre differenti significati del sostantivo e conteggiare le specifiche occorrenze.

I sostantivi лайк e лойс, invece, significano entrambi *like*; la differenza tra i due è che il primo, лайк, deriva dal social network *Facebook*, mentre лойс dal social network russo *VKontakte*. Il sostantivo лайк è nettamente più frequente rispetto a лойс, tuttavia non è possibile definire con puntualità le occorrenze di questi sostantivi, a causa degli strumenti di ricerca utilizzati: in primo luogo, se viene effettuata la ricerca del sostantivo лайк mediante la funzione *Concordance* di Sketch Engine, i risultati mostrano 7.906 occorrenze, ma del solo sostantivo al nominativo singolare; in altre parole, da questi risultati лайк sembrerebbe un sostantivo invariato, ma ciò non trova riscontro nella realtà, in quanto viene regolarmente declinato come sostantivo maschile forte. Inoltre, nei risultati della ricerca vengono inserite non solo le occorrenze del genitivo plurale della parola лайка (che significa 'cane eschimese, husky' e 'pelle di capretto'), ma anche le occorrenze del cognome traslitterato Лайк: ad esempio,

[206] Per una definizione accurata e completa si veda il sito *Academic* al link: https://dic.academic.ru/dic.nsf/enc_geolog/12141.
[207] Per una definizione accurata e completa si veda il sito Academic al link: https://dic.academic.ru/dic.nsf/dic_fwords/50673.

occorrono numerose volte i nomi Брукс Лайк, Янне Лайк, Иляна Лайк, Коринна Лайк, Настия Лайк, ecc. Per questi motivi i risultati non sono affidabili e non è possibile verificare le singole occorrenze dei casi nemmeno mediante *Word List*.

Non è possibile stabilire con certezza neanche le occorrenze del sostantivo лойс, poiché si sovrappongono a quelle del nome proprio di persona Лойс ("Lois"). Non è affidabile escludere tutte le occorrenze di Лойс scritte con l'iniziale maiuscola, nel tentativo di eludere le occorrenze del nome proprio di persona, perché la modalità di scrittura delle parole nel Web è spesso imprevedibile e i sostantivi indicanti la medesima cosa possono essere scritti interamente in maiuscolo, in minuscolo o con caratteri maiuscoli e minuscoli contemporaneamente. Lo stesso si può dire per il cognome Лайк.

Questo esempio ha mostrato i grandi limiti che ci possono essere nel realizzare analisi di tipo linguistico mediante web corpora di grandi dimensioni, ma poco affinati.

4.5.5 Sostantivi plurali traslitterati in russo e declinati come sostantivi singolari

Nel corso dell'analisi sono stati riscontrati casi in cui alcuni sostantivi plurali inglesi, terminanti quindi con desinenza *-s*, vengono traslitterati nella loro forma plurale e successivamente declinati come fossero dei sostantivi singolari terminanti in -с o -з (ossia "-s" e "-z").

I sostantivi che mostrano tale fenomeno sono i seguenti:

1.	Геймс, геймз;	5.	Спойлерс;
2.	Лайкс;	6.	Траблс, траблз;
3.	Лузерс;	7.	Хейтерс;
4.	Свайпс;	8.	Юзерс.

Tabella 10. Sostantivi plurali inglesi traslitterati e declinati come sostantivi singolari.

Nella sezione *Analisi dei dati* a pagina 178 è possibile consultare le analisi dei lessemi sopra citati; in questa sede ci limiteremo a riassumerle brevemente.

In tutti i casi le occorrenze vanno progressivamente diminuendo dal 2011 al 2021 e ciò dimostra una maggiore padronanza dei prestiti stranieri da parte dei russofoni.

Bisogna sottolineare che in alcuni casi l'utilizzo di questi sostantivi con -с/-з finali è motivato dalla traslitterazione di titoli o nomi di prodotti stranieri, terminanti per l'appunto in *-s*, pertanto non si tratta sempre di un errore commesso dal parlante russo, ma talvolta è giustificato da questioni di necessità. Il fatto più curioso ed interessante è

che questi sostantivi plurali vengono declinati, come se si trattasse di nominativi singolari, ma è evidente che spesso si tratta di una limitata conoscenza della lingua inglese o di un errore di distrazione del parlante.

4.5.6 Prestiti traslitterati o calchi parziali dello stesso nome composto

Nel corso della ricerca sono stati individuati dei nomi composti, aventi il medesimo significato, ma forme differenti: si tratta dei sostantivi селфи-стик e селфи-палка e dei nomi composti фейк-ньюс, фейкньюс e фейк-новости. Per quanto concerne селфи-стик (*selfie stick*) e фейк-ньюс o фейкньюс (*fake news*), si tratta di prestiti interamente traslitterati, mentre nel caso dei sostantivi селфи-палка ("bastone da *selfie*") e фейк-новости ("notizia *fake*") si può parlare di veri e propri calchi parziali.

In inglese questi sostantivi sono definiti *endocentric compounds*, ossia delle parole composte costituite da un centro semantico (in inglese *head*) portatore del significato semantico della parola composta, e un modificatore (dall'inglese *modifier*), che circoscrive il significato della parola. Nel caso di селфи-палка e фейк-новости, vengono traslitterati come prestiti stranieri soltanto i modificatori селфи- e фейк-, mentre nel caso dei centri semantici -палка ("bastone") e -новости ("notizie") si assiste al fenomeno del calco linguistico. È interessante osservare, peraltro, che in inglese la parola *news* fa parte dei sostantivi *pluralia tantum* e nel calco russo è stato mantenuto questo plurale.

Nella sezione *Analisi dei dati* a pagina 180 è possibile consultare le analisi dei lessemi sopra citati; in questa sede ci limiteremo a riassumerle brevemente.

Per quanto concerne la parola composta *selfie stick*, in russo si è affermato il calco parziale селфи-палка, rispetto al prestito traslitterato селфи-стик; questa tendenza è confermata dai dati di *ruTenTen17* e del *Timestamped JSI web corpus 2014-2021 Russian*. Per quanto riguarda la parola *fake news*, invece, si è affermato il prestito traslitterato фейк-ньюс, rispetto al calco parziale фейк-новости; anche in questo caso la tendenza è confermata dai dati di *ruTenTen17* e del *Timestamped JSI web corpus 2014-2021 Russian*.

Queste due parole composte sono un'ulteriore testimonianza degli scambi linguistici che possono avvenire tra due lingue, in contatto fra loro.

4.6 Gli aggettivi: processi derivazionali e forme differenti con lo stesso significato

Nel corso dell'analisi sono stati individuati numerosi aggettivi derivati; nei prossimi sotto-paragrafi analizzeremo i fenomeni osservati in relazione a questa categoria morfologica.

4.6.1 Gli aggettivi derivanti da prestiti stranieri traslitterati

Dall'analisi sono emersi 29 aggettivi derivanti da prestiti inglesi traslitterati, i cui dati sono disponibili nella sezione *Analisi dei dati* a pagina 180 e seguenti.

In 11 casi su 29 (38%) gli aggettivi si formano mediante il suffisso derivazionale -ов-/ -ев-, in 10 casi (34,5%), invece, si formano mediante il suffisso -н-, in 6 casi (20,7%) viene utilizzato il suffisso -ск- e in 2 casi (7%) il suffisso -овск-.

Inoltre, in 10 casi gli aggettivi russificati derivano da sostantivi inglesi traslitterati terminanti con il suffisso -*er*, ossia dai già citati *nomina agentis* e *actionis*, mentre in 2 casi gli aggettivi russificati derivano da prestiti di sostantivi inglesi terminanti con il suffisso -*ing*, ossia стриминговый (стриминг + -ов-+ ый), derivante dal sostantivo *streaming*[208], e стайлинговый (стайлинг + -ов- + ый), derivante dal sostantivo inglese *styling*[209]. Questi sostantivi inglesi in -*ing* generalmente indicano un'azione, un'attività o l'atto di fare qualcosa, pertanto gli aggettivi russificati derivati conservano questa azionalità nel significato.

Come è possibile osservare dalla tabella di analisi da pagina 180 a 183, sono stati analizzati alcuni aggettivi russi aventi suffissi derivazionali differenti, ma derivanti dal medesimo prestito traslitterato inglese. È questo il caso di aggettivi come апгрейдный e апгрейдовый, блогерный e блогерский, лузерный e лузерский, мейнстримный, мейнстримовский e мейнстримовый, спойлерный e спойлерский, стримовский e стримовый, трэшевый e трэшовый, derivanti rispettivamente dai sostantivi inglesi *upgrade*, *blogger*, *loser*, *mainstream*, *spoiler*, *stream* e *trash*.

Generalmente, ogni suffisso derivazionale conferisce una certa sfumatura di significato: ad esempio, i suffissi -ев-/-ов- indicano appartenenza, una proprietà dell'oggetto, un materiale, una terminologia tecnica; il suffisso -ск- può indicare

[208] Il sostantivo *streaming* viene definito dal *Cambridge Dictionary* come "the activity of listening to or watching sound or video directly from the internet"; per maggiori informazioni si veda al link: https://dictionary.cambridge.org/it/dizionario/inglese/streaming.

[209] Il sostantivo *styling* viene definito dal *Collins Dictionary* come "the act of creating or maintaining a hairstyle"; per altre notizie si veda al link: https://www.collinsdictionary.com/it/dizionario/inglese/styling.

relazione con una certa categoria o l'appartenenza alla categoria stessa; infine, il suffisso -н- indica una proprietà relativa ad un oggetto, un fenomeno, un'azione, un luogo, oppure l'essere soggetto ad una qualche azione o al risultato di un'azione, indicata dalla radice della parola.

Nonostante la specificità semantica di questi suffissi, gli aggettivi vengono utilizzati indistintamente, pertanto l'analisi dei dati può aiutare a comprendere quale forma dell'aggettivo si è affermata nella lingua russa. Bisogna ricordare che si tratta di neologismi coniati sulla base di prestiti inglesi e come tali si trovano in una fase di assestamento ed integrazione nella lingua. Mediante delle analisi diacroniche sarà possibile osservare il processo di consolidamento di questi neonati aggettivi nella lingua russa nel corso degli anni.

4.6.2 Aggettivi derivati da parole differenti, ma con la medesima radice

Nel corso dell'analisi sono stati osservati degli aggettivi, derivanti da lessemi differenti, aventi però la medesima radice. È questo il caso di криповый e крипотный, entrambi derivanti dal sostantivo inglese *creep*, con la sola differenza che il sostantivo крипота, da cui deriva l'aggettivo крипотный, è la variante russificata del prestito traslitterato крип (§ 4.5.4).

Стайлинговый e стайловый derivano da sostantivi differenti, ossia da *styling* e da *style*, ma la radice comune è *style*. Questi due aggettivi hanno differenti significati, in base al significato delle parole dalle quali derivano, che si conserva anche in russo: *styling* indica l'atto di creare qualcosa, in particolare un'acconciatura, oppure dei prodotti o degli strumenti utilizzati per dare una specifica forma o stile ai capelli, pertanto стайлинговый è molto specifico e solitamente viene utilizzato in riferimento all'acconciatura dei capelli; *style* è più generico, indica il modo in cui qualcosa è modellato o qualcuno è abbigliato, acconciato, ecc., pertanto стайловый viene solitamente usato per indicare la personalità o l'aspetto esteriore di qualcuno o qualcosa e potrebbe essere tradotto con 'stiloso', 'alla moda' o 'relativo allo stile'.

Gli aggettivi стримерный, стриминговый, стримовский e стримовый, derivano rispettivamente dai prestiti inglesi *streamer*, *streaming* e *stream*, ma hanno come radice comune la parola *stream*. Questi aggettivi hanno differenti sfumature di significato: come si è già detto (§ 4.5.2) il sostantivo *streamer* può essere al contempo un *nomen*

agentis, indicante una persona che fa una diretta streaming, e un *nomen actionis*, indicante un dispositivo di memoria su nastri magnetici, pertanto l'aggettivo da esso derivato indica una proprietà o una caratteristica con tale significato.

Il sostantivo *streaming* indica un'azione, il processo di trasmettere o scaricare in tempo reale dati Internet, perciò l'aggettivo ha un significato azionale, relativo a questo processo o attività. Infine, *stream* è la parola più generica, poiché può essere riferita sia allo scorrere di qualcosa, sia all'atto di guardare un film, un video o ascoltare la musica direttamente da Internet, senza scaricarla; in altre parole, questi due aggettivi, стримовский e стримовый, sono i più generici, poiché si possono applicare a una qualsiasi cosa che scorre o trascorre, come un fiume, un flusso, ecc.

4.7 I verbi: processi derivazionali e di russificazione

Nel corso dell'analisi sono emerse quattro principali tendenze, relative ai processi derivazionali dei verbi: la creazione di forme verbali aventi come radice sostantivi, acronimi o verbi inglesi traslitterati; la presenza di differenti varianti del verbo, derivanti dalla medesima radice, ma appartenenti a coniugazioni diverse ed anche il contrario, ossia la presenza di forme verbali con medesimo significato, ma derivanti da radici differenti; infine, la russificazione dei verbi derivanti da prestiti stranieri, mediante l'aggiunta del suffisso russo –ся per creare la forma riflessiva del verbo. Nei prossimi paragrafi vedremo nel dettaglio quanto illustrato.

4.7.1 Forme verbali derivanti dai prestiti inglesi traslitterati

Nel corso dell'analisi sono state individuate numerose forme verbali derivanti da sostantivi, da acronimi o da verbi inglesi, opportunamente traslitterati. Questo fenomeno è una manifestazione diretta dei processi d'integrazione dei prestiti e della loro successiva russificazione e consolidamento nella lingua russa.

I verbi individuati sono in tutto 40, consultabili nella sezione *Analisi dei dati* a pagina 184, e si formano perlopiù a partire da sostantivi inglesi traslitterati; solo in un caso è stato rilevato un verbo derivante da un acronimo, ossia ROFL (*Rolling On the Floor Laughing*), solitamente traslitterato in russo РОФЛ e che ha dato origine al verbo рофлить, che significa 'rotolarsi a terra dalle risate', 'ridere a crepapelle'.

Per quanto riguarda la formazione dei verbi, i principali suffissi derivazionali incontrati sono: -а-, -и-, -ирова-, -ну-, -ова-.

Il suffisso derivazionale più frequente è -и-, che compare in 25 volte su 40 (corrispondente al 62,5% del totale), seguito da -а-, utilizzato in 6 casi (15%). Il suffisso derivazionale -ну- è più raro, in quanto viene utilizzato in soli 4 casi (10%), così come i suffissi -ова- e -ирова- che vengono utilizzati rispettivamente 3 e 2 volte (7,5% e 5%).

In russo i suffissi derivazionali conferiscono un determinato significato ai verbi: in generale, il suffisso -и- indica un'azione legata al significato della radice del verbo (рыбачить, 'andare a pesca'), l'atto di agire mediante un oggetto o uno strumento, il verificarsi di eventi atmosferici (морозить, 'gelare'), il realizzarsi di una caratteristica indicata dalla radice del verbo (веселить, 'rallegrare', 'divertire') e molti altri; il suffisso -а- forma dei verbi denominali che indicano generalmente un'azione durativa o frequentativa (завтракать, 'fare colazione'); il suffisso -ну- forma verbi che spesso hanno il significato di azione momentanea, istantanea, singola, mentre il suffisso -ова- forma verbi imperfettivi che indicano l'atto di compiere qualcosa, dedicarsi ad una attività o trovarsi in un certo stato. In effetti, queste differenti sfumature di significato permangono anche in questi neonati verbi, ma non sempre: talvolta i verbi sembrano forme sinonimiche, altre volte hanno significato distintivo.

In vari casi sono state rilevate forme verbali derivanti dalla medesima parola: in primis, *upgrade* ha dato origine a 4 differenti forme verbali, 5 se si considera anche la variante апгрейтить; si tratta dei verbi апгрейдировать, il verbo riflessivo апгрейдиться, апгрейднуть e il più frequente di tutti, ossia апгрейдить.

Anche *game* ha dato origine a due differenti verbi in russo, гамать e геймить: apparentemente potrebbero sembrare verbi derivanti da prestiti stranieri differenti, ma in realtà la parola di origine è la medesima, ossia *game*, traslitterata in due differenti modi: гаме, come viene scritta in inglese, e гейм, secondo la pronuncia inglese. Dalle analisi la variante più affermata è гамать, sebbene solo degli studi diacronici della durata di decine di anni possono effettivamente mostrare quale delle due si affermerà definitivamente nella lingua.

Esistono due varianti russificate anche del verbo inglese *to dislike*: i verbi дизлайкать e дизлайкнуть; дизлайкать è leggermente più frequente rispetto a дизлайкнуть, sebbene le occorrenze siano talmente esigue non poter fare previsioni

certe. La principale differenza sta nel significato, oltre che nell'aspetto del verbo, poiché дизлайкнуть esprime un'azione singola ed istantanea, il 'cliccare con il cursore del mouse e mettere un *dislike*'.

I verbi свайпать, свайпить e свайпнуть derivano tutti dalla parola inglese *swipe*, che indica l'atto di scorrere il dito su uno schermo, ad esempio per sbloccare un dispositivo. Questo significato di movimento o azione singola ed istantanea spiegherebbe come mai свайпнуть sia la variante più frequente delle tre.

Спойлерить e спойлернуть derivano, invece, dalla parola *spoiler*, e significano 'spoilerare' o 'anticipare a voce una parte di un film, di una serie TV o di un libro, solitamente il finale'. La variante più frequente è спойлерить, tendenza confermata dai dati dell'*Osnovnoj korpus*, di *ruTenTen11* e *ruTenTen17*.

Dall'inglese *to troll* sono nati i verbi троллировать e троллить, che significano 'provocare qualcuno su Internet, in una chat, in un forum o in un blog; fomentare e disturbare'; троллировать è utilizzato di rado: 41 occorrenze in *ruTenTen11* e solo 14 in *ruTenTen17*, pertanto potrebbe essere destinato a scomparire, in favore della variante троллить.

Nel prossimo sotto-paragrafo analizzeremo delle forme verbali derivate da parole differenti, aventi però la medesima radice.

4.7.2 Verbi derivanti da parole differenti, ma con la medesima radice

Come è possibile osservare dalla tabella sopracitata, consultabile a pagina 183 e seguenti della sezione *Analisi dei dati*, nel corso della ricerca sono state rilevate alcune forme verbali derivanti da lessemi differenti, aventi però la medesima radice. È questo il caso di verbi come троллинговать e i già citati троллировать e троллить, tutti e tre derivanti dalla radice inglese *troll*, ma da parole differenti: in particolare, троллинговать deriva dal sostantivo inglese *trolling*, che indica un'azione[210], mentre троллировать e троллить derivano da *troll*, che indicherebbe la persona che compie questa azione, oppure il messaggio stesso che viene inviato[211]; in altre parole, il

[210] Secondo quanto riporta il *Cambridge Dictionary*, il sostantivo inglese *trolling* indica "the act of leaving an insulting message on the internet in order to annoy someone". Disponibile al link: https://dictionary.cambridge.org/it/dizionario/inglese/trolling.

[211] Secondo quanto riporta il *Cambridge Dictionary* il sostantivo *troll* indica "1) Someone who leaves an intentionally annoying or offensive message on the internet, in order to upset someone or to get attention

significato intrinseco è differente, sulla base della parola dalla quale derivano. Per quanto riguarda la loro frequenza, il verbo троллить è senza dubbio il più frequente, seguito da троллинговать e infine da троллировать.

I verbi хейтить e хейтерить, invece, derivano dalla radice *hate*, ma da parole differenti, rispettivamente *hate* e *hater*, pertanto il significato dei due verbi sarà differente: хейтить, il più frequente tra i due, ha un significato generico e può essere tradotto con ʻodiare', mentre хейтерить significa ʻcomportarsi da hater', ʻfare l'hater'.

4.7.3 Forme verbali derivanti da prestiti, rese riflessive mediante suffisso -ся

In russo i verbi riflessivi vengono creati mediante un processo definito *postfiksacija* (letteralmente ʻpostfissazione', sebbene in italiano si traduca ʻsuffissazione'), che si distingue dalla normale *suffiksacia* (ʻsuffissazione'): infatti la particella riflessiva -ся viene posta in fine di verbo, dopo i suffissi derivazionali che trasformano la radice della parola in una forma verbale.

Ad esempio, il verbo сорриться (ʻscusarsi') è costituito da una radice, due suffissi e un postfisso "copp + -и- + -ть- + -ся": la radice è copp-, derivante dal prestito inglese *sorry*, seguita dai due suffissi derivazionali, ossia il suffisso verbale -и- e il suffisso che indica l'infinito della forma verbale -ть- a seguito del quale viene aggiunto il "postfisso" riflessivo -ся.

Nel corso della ricerca sono stati rilevati 6 verbi riflessivi su 40 (corrispondenti al 15% del totale), tutti derivanti da prestiti stranieri traslitterati. Spesso questi verbi presentano anche la variante non riflessiva, come nel caso di агрить ʻfar arrabbiare, provocare', e il meno frequente агриться ʻarrabbiarsi, irritarsi', il verbo апгрейдить (o le sue numerose varianti апгрейдировать, апгрейднуть e апгрейтить) che significa ʻfare un upgrade' e il verbo riflessivo апгрейдиться ʻfarsi un upgrade', ʻmigliorarsi'. Il verbo селфить ʻfare un selfie' presenta la variante riflessiva селфиться ʻfarsi un selfie', così come il verbo чилить ʻriposare, dormire', derivante dall'inglese *chill*, che presenta la forma riflessiva, ma poco utilizzata, чилиться ʻrilassarsi, riposarsi'; citiamo poi il verbo юзать ʻusare, utilizzare' e la forma meno frequente юзаться, che solitamente si traduce con ʻsi usa/si usano'.

---

or cause trouble; 2) A message that someone leaves on the internet that is intended to annoy people". Disponibile al link: https://dictionary.cambridge.org/it/dizionario/inglese/troll.

Interessante è il verbo сорриться, ossia ‘scusarsi’, che come abbiamo già visto si forma dall’aggettivo inglese traslitterato *sorry*, al quale viene aggiunto il postfisso -ся per renderlo riflessivo: in inglese, infatti, non esiste il verbo *to sorry*, ma si usano formule come *to be sorry*, *to feel sorry* o *to say sorry*; pertanto questo verbo ha subito un profondo processo di russificazione, adattandosi al modello russo извинятьсяIMP – извинитьсяPF, anch’esso riflessivo.

4.8 Gli acronimi, le abbreviazioni, le interiezioni russe e le espressioni russificate

Durante la ricerca sono emersi interessanti acronimi di origine inglese traslitterati o russificati, nonché varianti russe di acronimi inglesi. Anche le abbreviazioni sono un aspetto molto interessante dello slang giovanile: la quasi totalità delle abbreviazioni riguarda parole russe già esistenti e solo in un caso si tratta dell’abbreviazione di una parola inglese traslitterata; infine, sono interessanti alcune espressioni russe ed espressioni di origine inglese russificate.

Nei prossimi tre paragrafi illustreremo i fenomeni appena citati.

4.8.1 Gli acronimi: traslitterazione e russificazione del modello alloglotto

Gli acronimi analizzati nel corso della ricerca sono in tutto 5: tre sono di origine inglese e traslitterati in caratteri cirillici, uno è un acronimo russo, creato però da un modello anglofono, uno è un acronimo inglese russificato. Le singole analisi sono consultbili a pagina 187 nella sezione *Analisi dei dati*.

Il primo prestito analizzato è ИМХО, corrispondente alla traslitterazione dell’acronimo inglese *In My Humble Opinion*, che in russo è assai frequente: i dati hanno rilevato 726 occorrenze nell’*Osnovnoj korpus*, 204.954 in *ruTenTen11* e 130.881 in *ruTenTen17*. Solitamente viene scritto con tutte le lettere maiuscole, fra due virgole, come un inciso, o fra parentesi. Questo acronimo è così popolare che è nata una variante russa, un vero e proprio calco del modello inglese, ossia ПМСМ, acronimo di по моему скромному мнению (‘secondo il mio modesto parere’); questo calco è di gran lunga meno frequente rispetto all’originale inglese: 3 sole occorrenze nell’*Osnovnoj korpus*, 1.371 in *ruTenTen11* e 756 in *ruTenTen17*. Similmente a ИМХО, anche ПМСМ viene scritto con le lettere maiuscole, ad inizio di frase seguito dalla virgola, o all’interno di una frase, tra due virgole o tra parentesi come un inciso.

Il secondo acronimo inglese analizzato è ЛОЛ, derivante dall'inglese *LOL*, che sta per *Lot Of Laughs*, oppure *Laughing Out Loud*; per quanto concerne la frequenza d'uso, si registrano 21 occorrenze nell'*Osnovnoj korpus*, 3 nell'*Ustnyj korpus*, 17.198 in *ruTenTen11* e 15.207 in *ruTenTen17*.

L'acronimo РОФЛ, invece, deriva dall'inglese *ROFL*, che significa *Rolling On the Floor Laughing*, ossia 'rotolarsi a terra dalle risate'. Il prestito non è molto frequente: 2 sole entrate nell'*Osnovnoj korpus*, 73 in *ruTenTen11* e 60 in *ruTenTen17*, eppure ha dato origine al verbo рофлить, già precedentemente trattato (§ 4.7.1).

L'ultimo acronimo analizzato si potrebbe definire un calco parziale: si tratta di СЗОТ, dall'inglese *sorry for off-topic*, ma derivato dalla variante russificata соррu за оффтопик ('scusa/scusate per l'off-topic'); in altre parole, una parte è stata traslitterata come un normale prestito, mentre la preposizione *za* è il calco russo della preposizione inglese *for*. Pertanto, si può parlare sia di prestito, che di parziale calco strutturale[212], in quanto in russo per esprimere le proprie scuse si usa l'espressione "извини/извините + preposizione за + caso accusativo del sostantivo" ('scusa/scusate per'). Per quanto concerne la frequenza d'uso, non sono molte le occorrenze registrate: 446 volte in *ruTenTen11* e 123 in *ruTenTen17*.

4.8.2 Le abbreviazioni di parole russe o di prestiti stranieri

Nello slang giovanile e nel mondo del Web le abbreviazioni sono molto frequenti; spesso il loro utilizzo è motivato da limiti di spazio (la possibilità di digitare solo un certo numero di caratteri) o di tempo (la necessità di digitare velocemente un messaggio). Nel corso della ricerca sono state individuate 7 abbreviazioni di parole russe ed 1 abbreviazione di una parola inglese. Le analisi dei dati possono essere consultate a pagina 188.

Il sostantivo варик è l'abbreviazione della parola russa вариант ('variante'), mentre nel gergo dei *gamer* è utilizzato in riferimento al gioco *World of Warcraft*; inoltre, варик era già presente nella lingua russa, come abbreviazione di alcuni nomi come *Varfolomej* ed altri[213]. Per quanto riguarda le occorrenze, nell'*Osnovnoj korpus* si

[212] Il *calco formale o strutturale* riproduce la struttura morfosintattica del modello originario. Il calco strutturale è solitamente anche semantico, mentre un calco semantico non è necessariamente strutturale.
[213] È possibile consultare tale accezione del significato di варик in *Academic* al link: https://dic.academic.ru/dic.nsf/lastnames/1908.

registra una sola entrata, mentre nei web corpora è nettamente più frequente: 2.435 occorrenze in *ruTenTen11* e 2.402 in *ruTenTen17*; spesso questo acronimo è preceduto da un numerale ordinale o seguito da un numerale cardinale (es. *pervyj varik* ʻprima variante', *varik 3* ʻvariante 3').

La seconda abbreviazione individuata è вписка che, similmente a варик, ha molteplici significati, stratificatisi nel corso del tempo: il significato più antico è ʻannotazione, biglietto'[214], derivante dal verbo вписывать^{IMP} – вписать^{PF} (ʻinscrivere, iscrivere'), ed infatti nell'*Osnovnoj korpus* la prima entrata con questo significato risale al periodo 1846-1874. Il secondo significato, risalente al periodo sovietico, è quello di ʻalloggio notturno temporaneo e gratuito'; più recentemente si è affermato il significato di ʻfesta notturna, generalmente in casa', quale sinonimo di вечеринка (ʻfesticciola, serata'), ma anche di ʻentrata gratuita in un club o in discoteca'[215]. Per quanto riguarda la frequenza d'uso si registrano 24 occorrenze nell'*Osnovnoj korpus*, 6 nell'*Ustnyj korpus*, 2.868 in *ruTenTen11* e 2.941 in *ruTenTen17*.

Il sostantivo днюха è l'abbreviazione di день рождения (ʻgiorno del compleanno') e, in effetti, si trova spesso associato a delle date ed esiste anche l'augurio "С днюхой!", ossia ʻBuon compleanno!'. Nell'*Osnovnoj korpus* compare 4 volte, nell'*Ustnyj korpus* ben 32, mentre nei web corpora è di gran lunga più frequente: in *ruTenTen11* si registrano 4.749 occorrenze e 3.873 in *ruTenTen17*.

Жи́за oppure жизá nel linguaggio giovanile corrisponde all'abbreviazione della parola жизнь (ʻvita'). Жиза occorre 3 volte nell'*Osnovnoj korpus*, mentre non compare nell'*Ustnyj korpus*. In *ruTenTen11* жиза occorre 8 volte, mentre in *ruTenTen17* occorre 97 volte.

La parola кста è l'unica abbreviazione che non riguarda un sostantivo, bensì un avverbio: si tratta infatti della forma abbreviata di кстати (ʻa proposito, tempestivamente, nel momento giusto'), utilizzata nel parlato o nello slang giovanile. È molto frequente, tanto che in *ruTenTen11* occorre ben 10.012 volte e 4.793 in *ruTenTen17*; in *NKRJa* le entrate registrate sono poche: 13 nell'*Osnovnoj korpus*, e 3 nell'*Ustnyj korpus*.

Il sostantivo maschile музон è l'abbreviazione di музыка (ʻmusica'), più raramente corrisponde al nome traslitterato del fiume *Mouzon* e dell'omonimo comune

[214] Per maggiori informazioni si veda al link: https://dic.academic.ru/dic.nsf/ushakov/765570.
[215] Per vedere i significati di вписка si veda al link: https://argo.academic.ru/845.

francese[216]. È mediamente frequente: nell'*Osnovnoj korpus* compare 31 volte, 11 nell'*Ustnyj korpus*, in *ruTenTen11* si registrano 4.283 occorrenze e 2.772 in *ruTenTen17*.

L'ultima abbreviazione individuata, relativa ad una parola russa, è forse il caso più interessante: si tratta di 7Я, abbreviazione della parola семья ('famiglia'). È un particolare tipo di abbreviazione, nato oltreoceano, in cui alcune sillabe della parola vengono sostituite da numeri omofoni e/o omografi. In questo caso 7Я è costituito da: "семь ('sette') + я → семья".

Questo tipo di abbreviazione è estremamente frequente in inglese, per citare alcuni esempi:

gr8 → *gr* + *eight* → *great* (per omofonia);

2day → *two* + *day* → *today* (per omofonia);

2nite → *two* + *nite* → *tonight* (per omofonia);

no1 → *no* + *one* → *nessuno* (per omofonia ed omografia).

Nonostante sia un fenomeno molto interessante, in lingua russa non è molto diffuso: non appare né in *NKRJa* né in *ruTenTen11*, mentre in *ruTenTen17* occorre 1.718 volte.

Nel corso della ricerca è emersa una sola abbreviazione derivante da un prestito traslitterato inglese, ossia Кэп: nello slang giovanile sta per *Captain Obvious*, un epiteto che si rivolge a chi fa affermazioni ovvie o banali; esiste anche la variante russificata, il calco semantico e strutturale Капитан Очевидность ('Capitan Ovvio'), abbreviato КО.

In realtà, la parola Кэп era già presente nella lingua russa, ma con differenti significati: indica numerosi acronimi russi[217], in secondo luogo è un tipo di nucleotide nell'ambito della genetica e della microbiologia[218], infine può indicare un massimo fisso nei tassi di interesse nel mondo del business e della finanza[219].

Data questa molteplicità di significati, in presenza di numerose occorrenze è difficile stabilire quando queste si riferiscano al significato indagato; secondo i dati rilevati, Кэп appare 80 volte nell'*Osnovnoj korpus*, ma solo 4 volte con il significato di

[216] Secondo i dati riportati da *Academic.ru* al link: https://dic.academic.ru/dic.nsf/ruwiki/1834562.
[217] Per consultare tutti gli acronimi si veda al link: https://sokrasheniya.academic.ru/807.
[218] Per maggiori informazioni relative a questo significato si veda al link: https://technical_translator_dictionary.academic.ru/105579.
[219] Per maggiori informazioni relative a questo significato si veda al link: https://dic.academic.ru/dic.nsf/business/7101.

Капитан Очевидность, tra gli anni 2009 e 2015. Nell'*Ustnyj korpus* occorre 1 volta, ma non con il significato indagato, infine occorre 13.446 in *ruTenTen11* e 7.760 in *ruTenTen17*, ma non è possibile determinare con precisione il numero di entrate relative al significato indagato.

Nel corso dell'analisi è stato rilevato questo limite significativo del corpus utilizzato, ovvero l'impossibilità di effettuare operazioni di disambiguazione del significato, al fine di verificare le occorrenze di una parola con un preciso significato.

4.8.3 Le interiezioni russe e le espressioni russificate

Nel corso della ricerca è stata individuata una particolare interiezione russa, molto diffusa nel linguaggio giovanile e nel mondo di Internet: si tratta di Ауф, sinonimo di Вау ('Wow'), ma poco frequente, tanto che non è presente in nessuno dei dizionari cartacei ed online consultati per la ricerca. Tuttavia, ricercando la parola mediante il motore di ricerca *Yandex*, sono numerose le pagine web che ne illustrano il significato ed alcuni esempi d'uso.

Anche l'analisi dei corpora dimostra che questo neologismo non si è ancora affermato nella lingua russa, nonostante sia presente da almeno 15 anni: nell'*Osnovnoj korpus* è presente una sola occorrenza, risalente all'anno 2007, mentre in *ruTenTen11* e in *ruTenTen17* occorre rispettivamente 744 e 405 volte.

Infine, l'ultima analisi di questa ricerca riguarda un'espressione inglese russificata, ossia по фану ('per divertimento'), corrispondente all'inglese *for fun*. Da un lato si tratta di un prestito linguistico, in questo caso della parola inglese *fun*, traslitterato фан, dall'altro si può parlare di calco strutturale della preposizione finale *for*, corrispondente al russo по, seguito dal caso dativo del prestito traslitterato. Sebbene non sia frequente (267 occorrenze in *ruTenTen11* e 249 in *ruTenTen17*), a mio avviso è un interessante fenomeno di adattamento di un'espressione anglofona calcata in una lingua flessiva, quale quella russa, ed adattata ad essa. I dati emersi dall'analisi possono essere consultati a pagina 189.

4.9 Conclusioni del capitolo

In questo capitolo sono stati esposti gli obiettivi della ricerca, gli strumenti e le modalità di analisi scelte. In seguito si è passati ad illustrare le principali osservazione emerse nel

corso delle analisi: la discussione è stata suddivisa in una prima sezione, riguardante le varie modalità di trascrizione delle parole, e in una seconda sezione, riguardante gli aspetti morfologici dei lessemi analizzati. Questa seconda parte della ricerca è stata a sua volta suddivisa in differenti sezioni: la prima tratta dei sostantivi, dal semplice prestito traslitterato, alla russificazione dei sostantivi, dalle parole composte importate come prestiti, alle parole composte russificate e trasformate in calchi parziali; la seconda parte riguarda gli aggettivi, in particolare gli aggettivi derivanti da prestiti stranieri e gli aggettivi derivanti da parole differenti, ma con la medesima radice.

La terza sezione si occupa di alcuni aspetti relativi ai verbi analizzati, come la creazione di forme verbali aventi come radice sostantivi, acronimi o verbi inglesi traslitterati; la presenza di differenti varianti del verbo, derivanti dalla medesima radice, ma appartenenti a coniugazioni diverse; la presenza di forme verbali con medesimo significato, ma derivanti da radici differenti; infine, la russificazione dei verbi derivanti da prestiti stranieri, mediante l'aggiunta del suffisso russo -ся per creare la forma riflessiva del verbo.

L'ultima sezione del capitolo si occupa degli acronimi, delle abbreviazioni, delle interiezioni russe e delle espressioni russificate emerse nel corso della ricerca.

Nelle prossime pagine verranno illustrare le conclusioni alle quali questo lavoro è giunto, ma anche i limiti della ricerca e degli strumenti di ricerca.

Conclusioni

Nella presente ricerca si è cercato di proporre una disamina teorica della linguistica dei corpora e, al tempo stesso, un esempio di analisi del *netspeak* russo contemporaneo mediante l'uso di corpora.

I primi tre capitoli indagano rispettivamente la linguistica dei corpora, intesa come branca della linguistica tradizionale (Capitolo 1), i corpora linguistici (Capitolo 2) e, infine, l'approccio più avanguardistico, il *Web as Corpus* e la creazione dei web corpora (Capitolo 3).

Il primo capitolo ha aperto nuove prospettive d'indagine circa le origini della linguistica dei corpora in Italia: mentre la tradizione italiana fa risalire le origini della disciplina all'epoca di Dante, secondo il mondo anglofono e russofono il capostipite dell'intera disciplina è il *Concordantiae Morales* di Sant'Antonio da Padova, vissuto circa un secolo prima di Dante. È emersa quindi una discrepanza nella periodizzazione di questa particolare branca della linguistica, quale punto di partenza per interessanti ricerche future.

Il secondo capitolo ha approfondito il concetto di corpus linguistico: innanzitutto, sono state analizzate le caratteristiche dei corpora, come il formato elettronico e l'autenticità dei dati linguistici, la rappresentatività, il bilanciamento e le dimensioni del corpus, nonché la ripetibilità e riproducibilità dei risultati d'analisi. Successivamente, si è passati alla rassegna delle varie tipologie di corpora, riportando numerosi esempi di corpora russi.

Il terzo capitolo ha ripercorso la storia della *Web Linguistics* e ha analizzato due dei più recenti approcci alla disciplina, il *Web as Corpus* e il *Web for Corpus*. Inoltre, sono state esaminate le principali critiche mosse nei confronti dell'approccio *Web as Corpus*, nonché i pareri di alcuni studiosi nei riguardi di questo approccio. Successivamente, nel capitolo sono state illustrate le numerose procedure per la creazione dei Web corpora, secondo l'approccio *Web for Corpus*, ed infine sono state analizzate le principali problematiche relative alla loro compilazione, ovvero il copyright e la lingua del Web, il *netspeak*.

Sulla base di quanto emerso nel corso del capitolo si può affermare che sebbene il Web sia un'importante fonte di materiale testuale, è ancora piuttosto sconosciuto,

incontrollato e incontrollabile per poter essere considerato uno strumento affidabile d'indagine. Al contrario, i Web corpora, basati sul materiale tratto da Internet, scaricato e successivamente filtrato, ripulito ed organizzato, risultano essere degli strumenti di indagine che rispettano appieno tutte le caratteristiche dei corpora linguistici, garantendo al tempo stesso una varietà e una ricchezza di dati testuali senza precedenti.

Il quarto ed ultimo capitolo offre un esempio di analisi del *netspeak* contemporaneo russo, effettuata mediante strumenti di ricerca differenti, quali dizionari cartacei ed online, corpora tradizionali e web corpora. In questo modo, è stato possibile indagare lo slang giovanile russo, i numerosi prestiti inglesi utilizzati nella lingua russa, il grado di acclimatamento ed integrazione di questi prestiti nella lingua russa e la tipologia di parole derivate e composte da questi prestiti inglesi.

Inoltre, è stata indagata la modalità di trascrizione dei prestiti; dalle osservazioni emerse è stata proposta la denominazione di "traslitterazione fonetica" in riferimento alla lingua russa, poiché la traslitterazione dai caratteri latini a quelli cirillici avviene sulla base della realizzazione fonetica di queste parole in russo e non sulla base dell'originale grafia in lingua inglese.

Per quanto concerne i risultati delle analisi, è emerso che il dizionario cartaceo non è uno strumento adeguato per la ricerca di neologismi e prestiti stranieri, infatti non è stato trovato alcun lessema oggetto di indagine. Al contrario, i dizionari online sono degli strumenti idonei: in soli 5 casi (nel caso di Ауф! 'Wow!', зумер 'zoomer', лойс 'like', стэнить 'comportarsi da fan accanito', 7Я 'famiglia') né *Vikislovar'*, né *Academic* presentavano alcun risultato di ricerca, mentre in 75 casi (corrispondenti al 93,75% del totale) contenevano la parola ricercata e talvolta alcuni lessemi derivati e/o composti.

Inoltre, è stato possibile tracciare un confronto statistico tra *Vikislovar'* e *Academic*: 55 volte su 75 (73,3%) entrambi contenevano la parola ricercata; in 16 casi su 75 (21,3%) solo *Vikislovar'* conteneva la parola ricercata, mentre 4 volte su 75 (5,3%) solo *Academic* conteneva la risposta. È possibile affermare, quindi, che lo strumento più aggiornato ed adeguato per indagare il *netspeak* russo è *Vikislovar'*, che nel 94,6% dei casi ha una risposta per la *query* di ricerca.

Per quanto concerne i due sotto-corpora di *NKRJa*, l'*Osnovnoj korpus* è più adatto rispetto all'*Ustnyj korpus*: infatti, in 56 casi su 80 (70%) l'*Osnovnoj korpus* presentava

delle occorrenze della parola ricercata, a differenza dell'*Ustnyj korpus,* che in soli 26 casi su 80 (32,5%) conteneva delle occorrenze. Nonostante la loro indubbia utilità, questi due sotto-corpora di *NKRJa* non possono essere paragonati ai web corpora utilizzati per le analisi, sia dal punto di vista della quantità, che della varietà dei dati.

RuTenTen11 e *ruTenTen17* sono senza dubbio gli strumenti più efficaci per indagare i neologismi e i prestiti inglesi della lingua russa: in soli 3 casi su 80 (nel caso di байтить ʻcopiare, imitare', стэнить ʻcomportarsi da fan accanito' e чапалах ʻceffone, sberla'), corrispondenti al 3,75% dei casi, non è stata trovata alcuna occorrenza della parola ricercata, mentre per il restante 96,25% dei casi sono state trovate delle occorrenze della parola ricercata e diversi lessemi derivati o composti.

Non è stato tracciato un confronto tra questi due corpora, in quanto essi sono stati realizzati in anni differenti, rispettivamente nel 2011 e nel 2017, pertanto *ruTenTen11* non contiene le occorrenze dei lessemi entrati nella lingua russa successivamente.

Per quanto concerne l'analisi lessicale, sono stati rilevati numerosi prestiti di origine straniera, che generalmente mostrano un buon grado di acclimatamento ed integrazione, tanto da originare nuove parole russificate derivate o composte.

La traslitterazione fonetica di queste parole spesso si discosta sia dalla trascrizione fonetica, sia dalla traslitterazione dell'originale parola inglese: ad esempio, le parole composte vengono scritte unite e non separate da uno spazio o da un trattino, così come avviene in inglese (il 70% dei casi studiati mostra questa caratteristica); similmente, il 50% dei sostantivi analizzati viene scritto con consonante singola, al contrario dello spelling originale che presenta consonante doppia. Inoltre, è emerso che la traslitterazione fonetica dei prestiti non dipende dall'originale trascrizione fonetica delle parole inglesi, ma viene adattata alle regole ortografiche e alla pronuncia della lingua russa.

I sostantivi sono il prestito più frequente, tanto che ne sono stati individuati ben 41 in questa analisi; solitamente si tratta di prestiti di lusso che danno origine a parole derivate e/o composte russificate. Un fenomeno interessante è la presenza di *nomina agentis* e *nomina actionis* inglese terminanti in *-er*, traslitterati per intero e declinati in russo come sostantivi maschili con terminazione forte in -ep. Nel corso dell'analisi ne sono stati individuati 22 in tutto.

Meno frequente, ma altrettanto interessante, è la derivazione - da questi *nomina agentis* terminanti in *-er* - di sostantivi di genere neutro con suffisso -ств(о), al fine di creare nuovi sostantivi dal significato astratto, indicanti solitamente un gruppo di persone.

La seconda categoria lessicale più produttiva è quella dei verbi: nel corso dell'analisi sono state individuate 40 forme verbali differenti, derivanti prevalentemente da sostantivi inglesi e in un solo caso da un acronimo inglese. Dalla ricerca è emerso che esistono forme verbali differenti, derivanti dal medesimo lemma inglese, ma in base al suffisso derivazionale utilizzato (-а-, -и-, -ирова-, -ну-, -ова-) questi verbi acquisiscono particolari caratteristiche di significato e aspettuali differenti, così come accade nei verbi russi.

Un fenomeno molto interessante riguarda quei verbi originati da prestiti inglesi traslitterati, ai quali viene aggiunto il suffisso -ся per renderli riflessivi; il caso più interessante è il verbo сорриться 'scusarsi', che è stato adattato al modello russo della coppia di verbi riflessivi извинятьсяIPF – извинитьсяPF, a differenza di quanto si trova nella lingua inglese con formule come *to be sorry*, *to feel sorry* o *to say sorry*, con verbo non riflessivo.

La terza categoria più frequente è quella degli aggettivi, derivanti da prestiti traslitterati e creati mediante differenti suffissi derivazionali tipici degli aggettivi russi (-ов-/-ев-, -н-, -ск-, -овск-). Nel corso dell'indagine sono stati individuati 29 aggettivi, ma, a differenza della categoria dei verbi, l'utilizzo di un suffisso derivazionale piuttosto che di un altro non incide sul significato dell'aggettivo, tanto che alcune forme aggettivali derivate stanno progressivamente scomparendo dalla lingua perché poco utilizzate.

Infine, la ricerca ha evidenziato che spesso gli acronimi russi sono il risultato di un calco semantico e strutturale del modello inglese, più raramente un prestito traslitterato, a differenza delle abbreviazioni, che derivano quasi esclusivamente da parole russe.

I principali problemi riscontrati nella preparazione del presente lavoro riguardano gli stessi strumenti utilizzati per la ricerca: in particolare, i web corpora utilizzati talvolta risultano poco affidabili e precisi. Ad esempio, come è stato evidenziato nel corso del quarto capitolo, spesso i risultati della ricerca mostrano le occorrenze relative

al solo caso ricercato, escludendo, dunque, le occorrenze con forme declinate in altri casi. Per esempio, nel caso venga ricercato il nominativo singolare di una parola, i risultati di ricerca potrebbero trarre l'utente in errore e fargli credere che il sostantivo sia un prestito straniero invariato. Un altro problema legato all'utilizzo dei due corpora è la mancata segnalazione dell'errore: gli esempi d'uso spesso mostrano errori grammaticali, sintattici o di errato *spelling* di una parola, che vanno ad inficiare la qualità dei dati e dei risultati di analisi.

Infine, sono stati rilevati casi di esempi d'uso ripetuti, sintomo che non sempre i testi sono esaustivamente de-duplicati.

In conclusione, la presente ricerca ha soddisfatto, a mio avviso, gli obiettivi preposti e in alcuni casi li ha anche superati. Al tempo stesso rimangono aperte alcune questioni che necessitano di ulteriori indagini: 1) l'analisi dei verbi derivati da prestiti, per verificare se la loro reggenza venga adattata o meno alla sintassi russa mediante calco strutturale, o se invece rimanga ancorata al modello inglese; 2) l'analisi dei sostantivi derivati da un prestito inglese, per verificare se siano retti dagli stessi verbi della lingua d'origine, oppure se vengano adattati alla lingua d'arrivo.

Nel primo caso, un esempio è rappresentato dai verbi гамать e геймит, entrambi derivati dalla parola *game*: gli esempi d'uso dimostrano che questi verbi hanno reggenza "гамать/геймить + в + caso accusativo", esattamente come il verbo russo играть, a differenza del modello anglofono "*to game/play something*".

Per quanto concerne il secondo campo d'analisi, un esempio è rappresentato dal sostantivo *selfie*: in inglese si utilizza l'espressione "*take a selfie*" ('scattare un selfie'), mentre in russo è di gran lunga più frequente l'espressione "делать^{IMP} – сделать^{PF} селфи" ('fare un selfie') rispetto a "снимать^{IMP} – снять^{PF} селфи" ('scattare un selfie'), segno che il prestito si è perfettamente integrato ed adattato alla lingua d'arrivo.

Questi sono solo alcuni esempi che danno l'idea di quanto sia vasto e complesso il campo d'indagine. La presente ricerca, infatti, non è un punto d'arrivo, bensì un punto di partenza con migliaia di destinazioni.

Appendice

Le tipologie di corpora linguistici

Tipologia	Breve descrizione relativa alla tipologia
Corpus annotato	Il corpus annotato contiene uno o più livelli di annotazione, effettuata in modo automatico, semi-automatico o manualmente. L'annotazione può essere di tipo morfosintattico, sintattico, semantico; per i corpora di parlato esiste anche quella fonetica, fonologica, prosodica o pragmatica.
Corpus grezzo	Il corpus grezzo non contiene alcuna annotazione. È lo strumento principale dell'approccio *corpus-based* che utilizza i dati linguistici puri.
Corpus generico o di riferimento	È un corpus di grandi dimensioni, rappresentativo della lingua in tutti i suoi aspetti: tipologie testuali, varietà linguistiche, periodi storici, caratteristiche socioculturali e sociolinguistiche della popolazione, fonti scritte ed orali. Questo tipo di corpora sono uno strumento utile per la creazione di grammatiche, dizionari o tesauri.
Corpus specialistico	È un corpus contenente dati testuali riguardanti uno specifico ambito di competenza, gruppo sociolinguistico, periodo storico o una determinata tipologia testuale. Solitamente sono di dimensioni ridotte.
Corpus diacronico	Il corpus diacronico indaga il mutamento linguistico di una lingua, di una particolare varietà linguistica, del lessico o di un fenomeno linguistico in un ampio arco temporale.
Corpus sincronico	Il corpus sincronico contiene dati testuali appartenenti ad una determinata, e solitamente ristretta, finestra temporale, al fine di condurre indagini linguistiche riguardanti una specifica fase della lingua.
Corpus storico	Un corpus storico se viene creato per condurre indagini su una lingua o una varietà linguistica di tipo storico, focalizzate su una determinata epoca nel passato.
Corpus dinamico	Il corpus dinamico è costantemente aggiornato mediante l'aggiunta, ad intervalli regolari, di dati testuali sempre nuovi ed in virtù di questo è solitamente di grandi dimensioni ed adatto per studi diacronici.
Corpus statico	Il corpus statico non è soggetto ad aggiornamento o integrazione di dati testuali, per questo motivo viene anche chiamato corpus "a campione chiuso".
Corpus di scritto	È un tipo di corpus realizzato medianti fonti scritte di vario tipo.

Corpus di parlato	È un corpus realizzato mediante fonti orali di vario tipo, successivamente trascritte.
Corpus misto	È un corpus che contiene sia fonti scritte, sia fonti orali successivamente trascritte.
Corpus audio	È un corpus contenente campioni di linguaggio parlato in forma di segnale acustico.
Corpus multimediale	È un corpus contenente registrazioni audio e video dell'atto comunicativo e permette di condurre studi riguardanti il linguaggio dei segni, la comunicazione verbale e non verbale e lo studio delle emozioni.
Corpus monolingue	Il corpus monolingue contiene dati testuali in una sola lingua.
Corpus multilingue	Il corpus multilingue contiene dati testuali in due o più lingue. Di questa tipologie si distinguono i corpora paralleli e comparabili.
Corpus parallelo	Contiene dati testuali originali e la loro traduzione in due o più lingue. Solitamente i corpora paralleli sono allineati, affinché vi sia una costante corrispondenza tra il testo originale e la sua traduzione.
Corpus comparabile	Due corpora si definiscono comparabili quando ciò che li accomuna non è la traduzione, ma modalità di campionamento affini (testi di dimensioni, tipologie o argomenti simili, ecc.).
Learner corpus	Un tipo di corpus che presenta tutte le tradizionali caratteristiche dei corpora, ma i cui dati testuali, orali e scritti, sono prodotti da apprendenti di una lingua.

I corpora della lingua russa

Araneum Russicum Minus, Maius e Maximum: sono dei web corpora appartenenti alla famiglia di corpora Aranea. L'Araneum Russicum Maius contiene 1,2 milioni di *token*, mentre l'Araneum Russicum Minus ne contiene 120 mila.
Link: http://ucts.uniba.sk/aranea_about/_russicum.html.

Araneum Russicum Russicum Minus e Maius: si tratta di web corpora appartenenti alla famiglia Aranea. Attualmente su Sketch Engine è possibile consultare l'Araneum Russicum Russicum Maius, contenente oltre 859 mila *token*.
Link: http://ucts.uniba.sk/aranea_about/index.html.

Častotnyj slovar' russkogo jazyka: pubblicato nel 1977, il dizionario si basava su un corpus di testi appartenenti a quattro differenti generi testuali, per un totale di 1 milione di parole.
Link: http://project.phil.spbu.ru/lib/data/slovari/zasorina/zasorina.html.

Chel'sinskij annotirovannyj korpus: è un corpus di lingua russa nato in Finlandia a Helsinki, di circa 100 mila parole provenienti da testi tratti dal settimanale socio-politico russo *Itogi*.
Link: http://h248.it.helsinki.fi/hanco/index.html.

CHILDES Russian Corpus: è un web corpus contenente trascrizioni di linguaggio infantile di bambini monolingui, registrati mentre conversano con i propri genitori, fratelli o sorelle. Il corpus contiene anche trascrizioni di bambini bilingui, bambini in età scolastica, adulti apprendenti di una L2, bambini con disturbi del linguaggio e afasie, ecc.
Link: https://childes.talkbank.org/access/Slavic/Russian/Protassova.html.

Corpus of Russian Spontaneous Speech (CoRuSS): si basa su 30 ore di conversazione di 60 madrelingua russi di età compresa tra i 16 e i 77 anni. È in parte trascritto ed annotato prosodicamente.

Link: https://aclanthology.org/L16-1309.pdf.

Dinamičeskij korpus tekstov po sovremennoj publicistike (90-e gody): creato dal Dipartimento di Lessicografia Sperimentale dell'Istituto di Lingua Russa dell'Accademia Russa delle Scienze, al fine di analizzare i cambiamenti nel linguaggio dei media e nel discorso politico durante il periodo della perestrojka e quello successivo alla caduta dell'Unione Sovietica.

Emotional Child Russian Speech Corpus (EmoChildRu): contiene 30 ore di registrazioni audio di 100 bambini di età compresa tra i 3 e 7 anni. Il corpus è stato progettato per studiare come si manifesta un certo stato emotivo nella voce e nel linguaggio.
Link:
https://www.researchgate.net/publication/281583846_EmoChildRu_Emotional_Child_Russian_Speech_Corpus.

Gutenberg Russian 2020: si tratta di un corpus creato mediante dei testi di libri russi presenti nel sito Progetto Gutenberg. Attualmente contiene 13,6 mila *token* ed è consultabile mediante Sketch Engine.
Link: https://www.sketchengine.eu/corpora-and-languages/russian-text-corpora/.

General'nyj Internet-Korpus Russkogo Jazyka: un web corpus generale della lingua russa di Internet, creato nel 2015 e contenente 20 miliardi di parole tratte da testi presenti su *Runet*.
Link: http://www.webcorpora.ru.

Istoričeskij korpus: un sotto-corpus di *NKRJa* che contiene testi scritti tra il XI e il XVIII secolo. Questo sotto-corpus è suddiviso in quattro distinte sezioni: *Drevnerusskij korpus*, *Berestjanye gramoty*, *Starorusskij korpus*, *Cerkovslavjanskij korpus*.
Link: https://ruscorpora.ru/new/search-old_rus.html.

Komp'juternyj korpus tekstov russkich gazet konca XX veka: comprende oltre 11 milioni di parole provenienti da articoli di giornale e di riviste, pubblicati tra il 1994 e il 1997.
Link: http://www.philol.msu.ru/~lex/corpus/corp_descr.html.

Korpus nesoveršennych perevodov: chiamato anche *Russian Learner Translator Corpus* (*RusLTC*). È un corpus parallelo inglese-russo, contenente traduzioni allineate di studenti provenienti da 14 università russe. Nell'ottobre 2018 *RusLTC* conteneva oltre 2,3 milioni di *token*.
Link: http://rus-ltc.org/search.

Korpus poetičeskich tekstov: è un sotto-corpus di *NKRJa*, inaugurato nel 2006 e contenente testi poetici scritti dal XVIII secolo ai giorni nostri. Presenta annotazione metatestuale, morfologica, semantica e una particolare annotazione poetica che indica il tipo di rima, di strofa, di ritmo, di metro e così via.
Link: https://ruscorpora.ru/new/search-poetic.html.

Korpus russkich učebnych tekstov: chiamato anche *Corpus of Russian Student Texts* (*CoRST*). È una raccolta di testi di 3,1 milioni di *token* scritti in russo da studenti iscritti all'università o a un master. Il corpus presenta informazioni metalinguistiche, annotazione morfosintattica e, soprattutto, l'error-tagging, che indica il tipo di errore linguistico e il motivo dell'errore.
Link: http://web-corpora.net/learner_corpus.

Korpus russkogo rasskaza pervoj treti XX v.: diviso in tre periodi cronologicamente sequenziali: l'inizio del XX secolo (1900-1913), il periodo della prima guerra mondiale, le rivoluzioni di febbraio e ottobre e la successiva guerra civile (1914-1922), l'epoca post-rivoluzionaria e il primo periodo sovietico (1923-1930).
Link: https://russian-short-stories.ru/.

Korpus russkogo žestovogo jazyka: contiene 230 video-testi di 43 parlanti della lingua russa dei segni. Contiene due diverse varianti della lingua dei segni russa, quella

"siberiana" e quella "moscovita", che consentono di studiare non solo la struttura interna e il funzionamento della lingua, ma anche la sua variazione territoriale.
Link: http://rsl.nstu.ru/site/index/language/ru.

Korpus sintaksičeskich kombinacij (KoSiKo): chiamato anche *Corpus of Syntactic Co-occurrences* (*CoSyCo*). Il corpus permette di ottenere elenchi di combinazioni di parole ed esempi d'uso in contesto reale, tratti da Internet; inoltre, fornisce informazioni sulle co-occorrenze delle parole e sulle relazioni sintattiche tra le parole.
Link: https://cosyco.ru/.

Korpus ustnoj reči: un sotto-corpus di *NKRJa* che contiene 13,4 milioni di parole , per un arco temporale che va dagli anni '30 del secolo scorso ad oggi.
Link: https://ruscorpora.ru/new/search-spoken.html.

Mašinnyj Fond Russkogo Jazyka: creato nel 1985, conteneva testi teatrali, di prosa e poesia russa dei secoli XIX e XX, un corpus di giornali russi degli anni '90 del secolo scorso, alcuni dizionari russi ed infine testi di storia e folklore russo.
Link: http://cfrl.ruslang.ru/.

Mul'timedijnyj korpus dialektnych tekstov «Žiznennyj krug»: si tratta di un corpus multimediale, contenente registazioni audio e videoregistrazioni, nonchè fonti testuali scritte. Per maggiori informazioni si vedano i tre articoli di J. N. Dračeva e N. N. Zubova in bibliografia.

Mul'timedijnyj russkij korpus (MURKO): è un sotto-corpus di *NKRJa*, inaugurato nel dicembre 2010. È possibile effettuare la ricerca di parole, ma anche di gesti (come cenni del capo, pacche sulla spalla, ecc.). Attualmente contiene oltre 5 mila *token*.
Link: https://ruscorpora.ru/new/search-murco.html.

Mul'timedijnyj Saratovskij dialektologičeskij tekstovyj korpus (СарДК): chiamato anche *Saratov Dialect Corpus (SarDC)*. Si tratta di un corpus dialettale multimediale

che presenta i testi trascritti con la versione originale audio-registrata o video-registrata in parallelo.
Per maggiori informazioni si consulti Krjučkova e Gol'din (2011).

Nacional'nyj Korpus Russkogo Jazyka: è il corpus nazionale della lingua russa; attualmente contiene più di 2,6 milioni di testi, per un totale di oltre 1 miliardo di parole. Oltre ad *Osnovnoj korpus*, contiene 15 sotto-corpora differenti.
Link: https://ruscorpora.ru/new/index.html.

Odin rečevoj den': creato mediante registrazione h24, raccoglie circa 1250 ore di registrazioni di 128 parlanti che interagiscono con più di 1000 persone appartenenti a diversi gruppi sociali. Il corpus contiene 1 milione di parole ed un sotto-corpus annotato di 125 mila parole.
Link: http://www.ord-corpus.spbu.ru/SocialStudies/ORD.html.

OPUS2 Russian: la famiglia di corpora OPUS contiene corpora paralleli in 40 lingue. Attualmente il corpus OPUS2 in lingua russa, contenente quasi 308 mila *token*, è consultabile tramite Sketch Engine.
Link: https://opus.nlpl.eu/.

Otkrytyj korpus: noto anche come *OpenCorpora*, si tratta di un corpus di lingua russa annotato morfologicamente, sintatticamente e semanticamente. Si chiama "Open" in quanto si tratta i un progetto open source a cui chiunque può prendere parte.
Link: http://opencorpora.org/.

Regensburgskij diachroničeskij korpus russkogo jazyka: La nuova versione include oltre 100 mila parole e contiene le opere di Kirill Turovskij e testi come il *Domostroj*, *Choždenie Bogorodicy po mukam*, *Povest' vremennych let* e la *Novgorodskaja pervaja letopis'*.
Link: https://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/ netzwerke/ regensburger-korpora/index.html.

RuSkELL corpus: *RuSKELL* è l'abbreviazione di *Russian Corpus for SKELL interface*. Il corpus è composto da testi che provengono dai principali e più frequenti domini Web russi, come kontrol'naja.ru, news.yandex.ru, alterauto.ru, pressarchive.ru and com.sibpress.ru. Attualmente su Sketch Engine è consultabile la versione *RuSKELL 1.6*, contenente quasi 1 milione di *token*.
Link: https://www.sketchengine.eu/russian-skell-corpus/.

Russian Error-Annotated Learner English Corpus: Un corpus con oltre 3,3 milioni di parole provenienti da testi in inglese, scritti da studenti madrelingua russi della Vysšaja Škola Ekonomiki. Include l'error-tagging.
Link: https://realec.org/.

Russian emotional corpus: la prima sezione include 295 videoregistrazioni di esami universitari orali per un totale di quasi 30 ore; la seconda sezione comprende videoregistrazioni di conversazioni riguardanti il pagamento delle utenze a uno sportello pubblico.

Russian-Finnish Parallel Corpus of Literary Texts (ParRus 2016): contiene testi letterari russi classici e del XX secolo e le loro traduzioni in finlandese allineate a livello di paragrafo. Attualmente ha raggiunto i 5,9 milioni di *token*.
Link: https://metashare.csc.fi/repository/browse/parrus-2016-russian-finnish-parallel-corpus-of-literary-texts/870bdc20fccc11e18b49005056be118e14a557fc52e5430ebfa6df946eba6e59/

RUSsian LANguage Affective speech (RUSLANA): raccoglie le registrazioni di parlato di 61 studenti universitari, 49 femmine e 12 maschi, di età compresa tra i 16 e i 28 anni. Agli studenti è stato chiesto di pronunciare dieci frasi in modo neutrale (senza alcuna emozione) e successivamente con sorpresa, felicità, rabbia, tristezza e paura.
Per maggiori informazioni si veda Makarova V. e Petrušin V. A. (2002).

RUssian LEarner Corpus of Academic Writing (RULEC): si tratta di un corpus contenente testi scritti, prodotti da studenti americani apprendenti il russo come lingua

straniera o come lingua ereditaria. Il materiale, raccolto in quattro anni, attualmente contiene circa 3800 testi scritti.
Link: http://www.web-corpora.net/RLC/rulec.

Russian Romani Corpus: contiene circa 720.000 *token* di testi pubblicati in URSS negli anni '20 e '30. Il corpus comprende tutti i testi originali (sia narrativa che stampa), nonché alcuni testi tradotti (narrativa, saggistica e stampa).
Link: http://web-corpora.net/RomaniCorpus/search/?interface_language=ru.

Russian spoken language corpus for speech synthesis (RUSLAN): contiene 22200 campioni audio con annotazioni di testo, per un totale di più di 31 ore di parlato di alta qualità.
Link: https://ruslan-corpus.github.io/.

Russkij korpus predloženij: nato nel 2014, si tratta di un corpus che studia i movimenti oculari degli adulti durante la lettura in lingua russa. Il corpus verrà utilizzato per confrontare i meccanismi di lettura in adulti madrelingua russi e senza patologie rispetto ad altri campioni della popolazione (bambini di lingua russa e bambini bilingui, adulti bilingui, bambini e adulti con dislessia, anziani e adulti di lingua russa con afasia).
Link: https://www.hse.ru/neuroling/research/RSC/.

Russkij učebnyj korpus: chiamato anche *Russian Learner Corpus* (*RLC*). Raccoglie dati orali e scritti di due particolari categorie di parlanti russi: coloro che studiano il russo come lingua straniera e i cosiddetti eritažnye govorjaščie, coloro che hanno appreso la lingua russa tramite esposizione informale.
Link: http://www.web-corpora.net/RLC/search/.

Russian Web corpus (ruTenTen): si tratta di un Web corpora creato sulla base di testi tratti da internet ed appartenente alla famiglia di corpora *TenTen*. Attualmente esistono il Russian Web 2006, 2011 e 2017.
Link: https://www.sketchengine.eu/russian-web-corpus/.

Saint Petersburg EFL Learner Corpus (SPbEFL LC): è un learner corpus relativamente piccolo, contenente testi orali e scritti di varia natura (saggi, lettere personali, monologhi e dialoghi). Il campione ricomprende 90 studenti delle scuole superiori di San Pietroburgo (Russia) e 12 loro coetanei, immigrati ma con un livello di russo intermedio o avanzato.

Sankt-Peterburgskij korpus agiografičeskich tekstov (SKAT): è un corpus elettronico di testi di letteratura agiografica russa del periodo tra il XV e il XVII secolo, creato presso il Dipartimento di Linguistica Matematica della Facoltà di Filologia dell'Università Statale di San Pietroburgo.
Link: http://project.phil.spbu.ru/scat/page.php?page=project.

Corpus of Russian Dialogue Speech (SibLing): contiene 90 dialoghi, di durata compresa tra i 25 e 60 minuti, di 100 parlanti. Il corpus presenta annotazione ortografica e fonetica.
Per maggiori informazioni si veda Kačkovskaja T. *et alii* (2020).

Testovyj korpus parallel'noj cintaksičeskoj razmetkoj: contiene 1 milione di parole tratte da testi di diverso genere, compresa la letteratura scientifica e narrativa, nonché testi di cronaca.
Link: http://otipl.philol.msu.ru/~soiza/testsynt/.

TOROT Corpus: è un treebank con annotazione morfologica e sintattica. Si tratta di un ampliamento del corpus PROIEL ed è stato avviato come parte del progetto di ricerca *Birds and Beasts: Shaping Events in Old Russian* , finanziato dal Norwegian Research Council. Il TOROT Corpus contiene testi in paleoslavo, o slavo ecclesiastico antico, in slavo meridionale e in russo moderno.
Link: https://nestor.uit.no/users/sign_in.

Učebnyj Mul'timodal'nyj KOrpus (UMKO): contiene 28 brevi videoregistrazioni di dialoghi spontanei di madrelingua russa e apprendenti di lingua russa di nazionalità

cinese o tedesca. Il corpus è stato creato nel 2011 ad opera dell'Istituto Linguistico Eurasiatico dell'Università Linguistica Statale di Mosca.

UH's Russian E-thesis corpus: il corpus contiene tesi di laurea magistrale e di dottorato pubblicate tra il 1999 e il 2016, per un totale di 1,1 milioni di parole.
Link: https://korp.csc.fi/korp/?mode=other_languages#?lang=en&cqp=%5B%5D&corpus=ethesis_ru&stats_reduce=word.

Uppsal'skij Korpus Russkogo Jazyka: composto da circa 600 testi russi specialistici e letterari, per un totale di 1 milione di parole. I testi specialistici appartengono al periodo compreso tra il 1985 e il 1989, mentre i testi letterari sono degli anni 1960-1988.
Link: https://www.lingexp.uni-tuebingen.de/sfb441/b1/en/korpora.html.

Lista di dizionari e siti consultati per l'individuazione dei lessemi da analizzare

La data di consultazione ultima di tutti i presenti link è il 7 Dicembre 2021

Dizionari:

Slovar' molodëžnogo slenga: https://slang.su/.

Slovar' molodëžnogo slenga nel sito *DicsOnline*:
https://www.dicsonline.ru/zhargonnye-slovari.

Slovar' molodëžnogo slenga nel sito *ZnačenieSlova*:
https://znachenieslova.ru/slovar/youthslang/.

Slovar' sovremennogo slenga nel sito *AntiSlang.RU*: https://antislang.ru.

Sitografia:

Un articolo della *Komsomol'skaja Pravda* del 22 Novembre 2017:
https://www.kp.ru/daily/26761.3/3790678/.

Un articolo della *Volžskaja Pravda* del 13 Maggio 2020:
https://yandex.ru/turbo/gazeta-vp.ru/s/budesh-livat-offni-svet-razbiraem-sleng-sovremennyh-podrostkov/.

Un articolo di *ANews* del 22 Giugno 2020:
https://anews.com/novosti/131249527-molodezhnyj-sleng-2021-chto-takoe-krash-krinzh-vpiska-i-padra-slovary.html.

Un articolo di *Life* del 27 Giugno 2016:
https://life.ru/p/422651.

Un articolo di *iSmart* dell'8 Novembre 2021:
https://ismart.org/capabilities/library/slovar-podrostkovogo-slenga-dlya-tekh-kto-ne-otlichaet-krash-ot-krinzha/.

Un articolo di *Masterlang* del 25 Gennaio 2020:
https://masterlang.ru/top-desyat-samikh-modnyh-slengovyh-slov-2019/.

Un articolo di *Memepedia* dell'8 Agosto 2019:
https://memepedia.ru/kto-takie-bumery-zumery-i-dumery/.

Un articolo di *MTS Media* del 10 Settembre 2021:
https://media.mts.ru/technologies/197036-sleng-tinejdzherov/.

Un articolo di *Nižnij Novgorod Online* del 12 Luglio 2019:
https://www.nn.ru/text/culture/2019/07/12/66159211/.

Un articolo di *Škola Novogo Pokolenija NGS*:
https://ngs-school.kz/ru/generation/vocabulary.

Un articolo di *Udmurtija – Informacionnoe agentstvo* del 29 Gennaio 2021:
https://udmurt.media/articles/obshchestvo/103798/.

Un articolo di *Yandex Zen* del 12 Agosto 2019:
https://zen.yandex.ru/media/id/5d34ec7df2df2500ae87a3a2/20-slov-iz-molodejnogo-leksikona-kotorye-navriatli-poimet-vcherashniaia-molodej-5d5089561d656a00adddc4e4.

Indice dei lessemi analizzati

1. Абьюз, абъюз, абюз; абьюзер, абъюзер, абюзер; абьюзить, абъюзить, абюзить; абьюзный, абъюзный, абюзный.
2. Агрить, агриться.
3. Апгрейд, апгрэйд; апгрейд, ап-грейд; апгрейт, апгрэйт; апгрейдер, апгрэйдер; апгрейдинг, апгрэйдинг; апгрейдный, апгрэйдный; апгрейдовый, апгрэйдовый; апгрейденный, апгрэйденный.
4. Апгрейдить, апгрэйдить; апгрейдиться, апгрэйдиться; апгрейднуть, апгрэйднуть; апгрейтить, апгрэйтить; апгрейдировать, апгрейдировать.
5. Ауф.
6. Байтить.
7. Баттхерт, баттхёрт, баттхёртить, баттхертить.
8. Блогер, блоггер; блогерский, блоггерский; блогерный, блоггерный; блогерство, блоггерство.
9. Бодиарт, боди-арт.
10. Бодипозитив, боди-позитив; бодипозитивный, боди-позитивный.
11. Бодишейминг, боди-шейминг; бодишеймер, боди-шеймер.
12. Буллинг, кибербуллинг, кибер-буллинг, кибербулинг, кибер-булинг; буллер, кибербуллер, кибер-буллер.
13. Вайб, вайбер
14. Варик
15. Вписка
16. Гамать
17. Гейм, гэйм; гейминг, гэйминг; геймингoвой, гэймингoвой; геймер, гэймер; геймерский, гэймерский; геймерство; геймить, гэймить; геймиться, гэймиться; геймплей, гэймплей; геймплэй, гэймплэй; геймплейный, гэймплейный; геймификация, гэймификация.
18. Дакфейс, дакфейсинг.
19. Дислайк, дизлайк, дизлайкнуть, дизлайкать.
20. Днюха.
21. Жиза.

22. Зашквар, зашквариться, зашкварить, зашкваренный, зашкваривания.
23. Зумер.
24. ИМХО.
25. Инфлюенсер, инфлюэнсер.
26. Краш.
27. Крейз, крэйз; крейзи, крэйзи; крейзер, крэйзер; крейзинг, крэйзинг; крейзанутый, крэйзанутый; крейзовый, крэйзовый.
28. Криндж, кринж; кринджовый, кринжовый.
29. Крипово, крипота, криповый, крипотный.
30. Кста.
31. Кэп.
32. Лайк, лайковый, лайкер.
33. Лайфхак, лайфхакер, лайфхакинг.
34. Лойс.
35. ЛОЛ.
36. Лузер, лузерство, лузерский, лузерный, лузерс.
37. Мейнстрим, мэйнстрим; мейнстримовый, мэйнстримовый; мейнстримный, мэйнстримный; мейнстримовский, мэйнстримовский; мейнстриминг, мэйнстриминг.
38. Муд.
39. Музон.
40. Нетикет, нэтикет.
41. Офтоп, оф-топ; оффтоп, офф-топ; офтопик, оф-топик; оффтопик, офф-топик
42. ПМСМ.
43. По фану.
44. Пранк , пранкер, пранковать.
45. РОФЛ, рофлить, рофлер.
46. Свайп, свайпнуть, свайпать, свайпить, свайпинг, свайпс.
47. Селфи, сэлфи; селфить, сэлфить; селфиться, сэлфиться; селфи-камера, сэлфи-камера; селфи-палка, сэлфи-палка; селфи-стик, сэлфи-стик.
48. Сетикет.
49. СЗОТ.

50. Сорри, сорриться.
51. Спойлер, спойлерство, спойлерить, спойлернуть, спойлерный.
52. Стайл, стайлс, стайлер, стайлинг, стайлинговый, стайлиш, стайловый.
53. Стрим, стример, стримерный, стриминг, стриминговый, стримить, стримовый, стримовский.
54. Стэнить.
55. Трабл, траблс, траблер, траблмейкер, траблшутинг, траблшутить.
56. Трип, трипп, трипер, триппер, триперный, трипперный, триповый, триповать, трип-хоп, трип-хоповый трип-компьютер.
57. Тролинг, троллинг; тролинговый, троллинговый; тролить, троллить; тролинговать, троллинговать; тролировать, троллировать.
58. Треш, трэш; трешер, трэшер; трешевый, трэшевый; трешовый, трэшовый.
59. Фейк, фэйк; фейковый, фэйковый; фейк-новости, фэйк-новости; фейк-ньюс, фэйк-ньюс; фейкньюс, фэйкньюс.
60. Фейспалм, фэйспалм; фейспалмить, фэйспалмить.
61. Флексить.
62. Флуд, флудер, флудерский, флудинг, флудить.
63. Флешмоб, флэшмоб; флеш-моб, флэш-моб; флешмобер, флэшмобер; флешмоббер, флэшмоббер; флеш-мобер, флэш-мобер; флеш-мобер, флэш-моббер.
64. Фоловер, фолловер; фоловинг, фолловинг; фоловить, фолловить.
65. Френдзона, френд-зона, френд зона.
66. Хайп, хайповый, хайпануть, хайпер.
67. Хейтер, хэйтер; хейтерить, хэйтерить; хейтить, хейтерс; хейтерство, хэйтерство; хейтинг, хэйтинг; хейтерский, хэйтерский.
68. Хипстер, хипстерство, хипстерский.
69. Хэштег, хэштэг, хештег.
70. Чайлдфри, чайлд-фри, чайлд фри.
71. Чапалах.
72. Чекать.
73. Челендж, челлендж; челенж, челленж; челенджер, челленджер, челенжер, челленжер.

74. Чилл-аут, чиллаут, чил-аут, чилаут, чиллер, чиллерный.
75. Чилить, чиллить; чилиться, чиллиться.
76. Шейм, шэим; шейминг, шэйминг.
77. Юзер, юзерский, юзать, юзаться.
78. Юзернейм.
79. Юзерпик.
80. 7Я.

Analisi dei dati

4.4 Le modalità di trascrizione dei prestiti stranieri

4.4.1 Utilizzo di Э (*e* tonica) e di E (*e* palatalizzata)

1. Апгрейд, апгрэйд; апгрейт, апгрэйт; апгрейдер, апгрэйдер; апгрейдинг, апгрэйдинг; апгрейдный, апгрэйдный; апгрейдовый, апгрэйдовый; апгрейденный, апгрэйденный.

2. Апгрейдить, апгрэйдить; апгрейдиться, апгрэйдиться; апгрейднуть, апгрэйднуть; апгрейтить, апгрэйтить; апгрейдировать, апгрейдировать.

3. Гейм, гэйм; гейминг, гэйминг; геймингoвый, гэймингoвый; геймер, гэймер; геймерский, гэймерский; геймить, гэймить; геймиться, гэймиться; геймплей, гэймплей; геймплэй, гэймплэй; геймплейный, гэймплейный; геймификация, гэймификация.

4. Инфлюенсер, инфлюэнсер.

5. Крейз, крэйз; крейзи, крэйзи; крейзер, крэйзер; крейзинг, крэйзинг; крейзанутый, крэйзанутый; крейзовый, крэйзовый.

6. Мейнстрим, мэйнстрим; мейнстримовый, мэйнстримовый; мейнстримный, мэйнстримный; мейнстримовский, мэйнстримовский; мейнстриминг, мэйнстриминг.

7. Нетикет, нэтикет.

8. Селфи, сэлфи; селфить, сэлфить; селфиться, сэлфиться; селфи-камера, сэлфи-камера; селфи-палка, сэлфи-палка; селфи-стик, сэлфи-стик.

9. Треш, трэш; трешер, трэшер; трешевый, трэшевый; трешовый, трэшовый.

10. Фейк, фэйк; фейковый, фэйковый; фейк-новости, фэйк-новости; фейк-ньюс, фэйк-ньюс; фейкньюс, фэйкньюс.

11. Фейспалм, фэйспалм; фейспалмить, фэйспалмить.

12. Флешмоб, флэшмоб; флеш-моб, флэш-моб; флешмобер, флэшмобер; флешмоббер, флэшмоббер; флеш-мобер, флэш-мобер; флеш-мобер, флэш-моббер.

13. Хейтер, хэйтер; хейтерить, хэйтерить; хейтить, хейтерс; хейтерство, хэйтерство; хейтинг, хэйтинг; хейтерский, хэйтерский.

14. Хэштег, хэштэг, хештег.

15. Шейм, шэим; шейминг, шэйминг.

1. Апгрейд, апгрэйд; апгрейт, апгрэйт; апгрейдер, апгрэйдер; апгрейдинг, апгрэйдинг; апгрейдный, апгрэйдный; апгрейдовый, апгрэйдовый; апгрейденный, апгрэйденный.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Апгрейд, апгрейдовый; *Academic*: Апгрейд, апгрейдер.
NKRJa Osnovnoj korpus	**Апгрейд**: 49 volte. **Апгрейт**: 1 volta. **Апгрейдер**: No.	**Апгрэйд**: No. **Апгрэйт**: No. **Апгрэйдер**: No.

NKRJa Ustnyj korpus	**Апгрейдинг**: No. **Апгрейдный**: No. **Апгрейдовый**: No. **Апгрейденный**: No. **Апгрейд**: No. **Апгрейт**: No. **Апгрейдер**: No. **Апгрейдинг**: No. **Апгрейдный** No. **Апгрейдовый**: No. **Апгрейденный**: No.	**Апгрэйдинг**: No. **Апгрэйдный**: No. **Апгрэйдовый**: No. **Апгрэйденный**: No. **Апгрэйд**: No. **Апгрэйт**: No. **Апгрэйдер**: No. **Апгрэйдинг**: No. **Апгрэйдный**: No. **Апгрэйдовый**: No. **Апгрэйденный**: No.
ruTenTen11	**Апгрейд**: 39.915 volte. **Апгрейт**: 968 volte. **Апгрейдер**: 82 volte. **Апгрейдинг**: 37 volte. **Апгрейдный**: 74 volte. **Апгрейдовый**: 20 volte. **Апгрейденный**: 16 volte.	**Апгрэйд**: 427 volte. **Апгрэйт**: 22 volte. **Апгрэйдер**: No. **Апгрэйдинг**: No. **Апгрэйдный**: No. **Апгрэйдовый**: No. **Апгрэйденный**: No.
ruTenTen17	**Апгрейд**: 24.780 volte. **Апгрейт**: 539 volte. **Апгрейдер**: 39 volte. **Апгрейдинг**: 19 volte. **Апгрейдный**: 30 volte. **Апгрейдовый**: No. **Апгрейденный**: 16 volte.	**Апгрэйд**: 202 volte. **Апгрэйт**: No. **Апгрэйдер**: No. **Апгрэйдинг**: No. **Апгрэйдный**: No. **Апгрэйдовый**: No. **Апгрэйденный**: No.

2. Апгрейдить, апгрэйдить; апгрейдиться, апгрэйдиться; апгрейднуть, апгрэйднуть; апгрейтить, апгрэйтить; апгрейдировать, апгрейдировать.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Апгрейдить, апгрейтить апгрейдиться, апгрейднуть; *Academic*: No.
NKRJa Osnovnoj korpus	**Апгрейдить**: 2 volte. **Апгрейдиться**: No. **Апгрейднуть**: No. **Апгрейтить**: No. **Апгрейдировать**: No.	**Апгрэйдить**: No. **Апгрэйдиться**: No. **Апгрэйднуть**: No. **Апгрэйтить**: No. **Апгрэйдировать**: No.
NKRJa Ustnyj korpus	**Апгрейдить**: No. **Апгрейдиться**: No. **Апгрейднуть**: No. **Апгрейтить**: No. **Апгрейдировать**: No.	**Апгрэйдить**: No. **Апгрэйдиться**: No. **Апгрэйднуть**: No. **Апгрэйтить**: No. **Апгрэйдировать**: No.
ruTenTen11	**Апгрейдить**: 3.146 volte. **Апгрейдиться**: 681 volte. **Апгрейднуть**: 59 volte. **Апгрейтить**: 44 volte. **Апгрейдировать**: 36 volte.	**Апгрэйдить**: 40 volte. **Апгрэйдиться**: No. **Апгрэйднуть**: No. **Апгрэйтить**: No. **Апгрэйдировать**: No.
ruTenTen17	**Апгрейдить**: 2.080 volte. **Апгрейдиться**: 354 volte. **Апгрейднуть**: 47 volte.	**Апгрэйдить**: 39 volte. **Апгрэйдиться**: No. **Апгрэйднуть**: No.

	Апгрейтить: 25 volte. **Апгрейдировать**: 16 volte.	**Апгрэйтить**: No. **Апгрэйдировать**: No.

3. Гейм, гэйм; гейминг, гэйминг; гейминговый, гэйминговый; геймер, гэймер; геймерский, гэймерский; геймить, гэймить; геймиться, гэймиться; геймплей, гэймплей; геймплэй, гэймплэй; геймплейный, гэймплейный; геймификация, гэймификация.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Гейм, гейминг, геймер, гейминговый, геймерский, геймплей, геймплейный, геймификация; *Academic*: No.
NKRJa Osnovnoj korpus	**Гейм**: 177 volte. **Гейминг**: No. **Гейминговый**: No. **Геймер**: 117 volte. **Геймерский**: 4 volte. **Геймить**: No. **Геймиться**: No. **Геймплей** 5 volte. **Геймплэй**: No. **Геймплейный**: No. **Геймификация**: 3 volte.	**Гэйм**: 4 volte. **Гэйминг**: No. **Гэйминговый**: No. **Гэймер**: No. **Гэймерский**: No. **Гэймить**: No. **Гэймиться**: No. **Гэймплей**: No. **Гэймплэй**: No. **Гэймплейный**: No. **Гэймификация**: No.
NKRJa Ustnyj korpus	**Гейм**: 9 volte. **Гейминг**: No. **Гейминговый**: No. **Геймер**: No. **Геймерский**: No. **Геймить**: No. **Геймиться**: No. **Геймплей**: No. **Геймплэй**: No. **Геймплейный**: No. **Геймификация**: No.	**Гэйм**: No. **Гэйминг**: No. **Гэйминговый**: No. **Гэймер**: No. **Гэймерский**: No. **Гэймить**: No. **Гэймиться**: No. **Гэймплей**: No. **Гэймплэй**: No. **Гэймплейный**: No. **Гэймификация**: No.
ruTenTen11	**Гейм**: 21.485 volte. **Гейминг**: 1.383 volte. **Гейминговый**: 29 volte. **Геймер**: 65.289 volte. **Геймерский**: 5.750 volte. **Геймить**: 16 volte. **Геймиться**: 33 volte. **Геймплей**: 48.852 volte. **Геймплэй**: 948 volte. **Геймплейный**: 3.429 volte. **Геймификация**: 94 volte.	**Гэйм**: 86 volte. **Гэйминг**: 68 volte. **Гэйминговый**: No. **Гэймер**: 59 volte. **Гэймерский**: No. **Гэймить**: No. **Гэймиться**: No. **Гэймплей**: 34 volte. **Гэймплэй**: 55 volte. **Гэймплейный**: No. **Гэймификация**: No.
ruTenTen17	**Гейм**: 11.016 volte. **Гейминг**: 1.023 volte. **Гейминговый**: 7 volte. **Геймер**:49.774 volte. **Геймерский**: 3.799 volte. **Геймить**: 34 volte.	**Гэйм**: 92 volte. **Гэйминг**: No. **Гэйминговый**: No. **Гэймер**: 32 volte. **Гэймерский**: No. **Гэймить**: No.

	Геймиться: 5 volte. **Геймплей**: 39.621 volte **Геймплэй**: 402 volte. **Геймплейный**: 2.882 volte **Геймификация**: 1.390 volte	**Гэймиться**: No. **Гэймплей**: 8 volte. **Гэймплэй**: 14 volte. **Гэймплейный**: No. **Гэймификация**: No.

4. Инфлюенсер, инфлюэнсер.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Инфлюэнсер; *Academic*: No.
NKRJa Osnovnoj korpus	**Инфлюенсер**: No. **Инфлуенсер**: No.	**Инфлюэнсер**: No. **Инфлуэнсер**: No.
NKRJa Ustnyj korpus	**Инфлюенсер**: No. **Инфлуенсер**: No.	**Инфлюэнсер**: No. **Инфлуэнсер**: No.
ruTenTen11	**Инфлюенсер**: No.	**Инфлюэнсер**: No.
ruTenTen17	**Инфлюенсер**: 31 volte. **Инфлуенсер**: No.	**Инфлюэнсер**: No. **Инфлуэнсер**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Инфлюенсер**: 2.882 volte. **Инфлуенсер**: 37 volte.	**Инфлюэнсер**: 843 volte. **Инфлуэнсер**: 5 volte.

5. Крейз, крэйз; крейзи, крэйзи; крейзер, крэйзер; крейзинг, крэйзинг; крейзанутый, крэйзанутый; крейзовый, крэйзовый.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Крейзи, крейзанутый, крейзовый; *Academic*: No.
NKRJa Osnovnoj korpus	**Крейз**: No. **Крейзи**: 27 volte. **Крейзер**: 17 volte. **Крейзинг**: 1 volta. **Крейзанутый**: 1 volta. **Крейзовый**: No.	**Крэйз**: No. **Крэйзи**: 10 volte. **Крэйзер**: No. **Крэйзинг**: No. **Крэйзанутый**: No. **Крэйзовый**: No.
NKRJa Ustnyj korpus	**Крейз**: No. **Крейзи**: No. **Крейзер**: No. **Крейзинг**: No. **Крейзанутый**: No. **Крейзовый**: No.	**Крэйз**: No. **Крэйзи**: No. **Крэйзер**: No. **Крэйзинг**: No. **Крэйзанутый**: No. **Крэйзовый**: No.
ruTenTen11	**Крейз**: 347 volte. **Крейзи**: 8.032 volte. **Крейзер**: 536 volte. **Крейзинг**: 67 volte. **Крейзанутый**: 61 volte. **Крейзовый**: 5 volte.	**Крэйз**: 19 volte. **Крэйзи**: 660 volte. **Крэйзер**: No. **Крэйзинг**: No. **Крэйзанутый**: No. **Крэйзовый**: No.
ruTenTen17	**Крейз**: 96 volte. **Крейзи**: 2.410 volte.	**Крэйз**: 5 volte. **Крэйзи**: 335 volte.

	Крейзер: 251 volte. **Крейзинг**: 33 volte. **Крейзанутый**: 21 volte. **Крейзовый**: 5 volte.	**Крэйзер**: No. **Крэйзинг**: No. **Крэйзанутый**: No. **Крэйзовый**: No.

6. Мейнстрим, мэйнстрим; мейнстримовый, мэйнстримовый; мейнстримный, мэйнстримный; мейнстримовский, мэйнстримовский; мейнстриминг, мэйнстриминг.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Мейнстрим, мейнстримовый, мейнстримный, мейнстримовский; *Academic*: Мейнстрим, мейнстримовский.
NKRJa Osnovnoj korpus	**Мейнстрим**: 137 volte. **Мейнстримовый**: 2 volte. **Мейнстримный**: 11 volte. **Мейнстримовский**: 10 volte. **Мейнстриминг**: No.	**Мэйнстрим**: 42 volte. **Мэйнстримовый**: 2 volte. **Мэйнстримный**: 2 volte. **Мэйнстримовский**: 6 volte. **Мэйнстриминг**: No.
NKRJa Ustnyj korpus	**Мейнстрим**: 14 volte. **Мейнстримовый**: No. **Мейнстримный**: No. **Мейнстримовский**: No. **Мейнстриминг**: No.	**Мэйнстрим**: 2 volte. **Мэйнстримовый**: 1 volta. **Мэйнстримный**: No. **Мэйнстримовский**: No. **Мэйнстриминг**: No.
ruTenTen11	**Мейнстрим**: 12.545 volte. **Мейнстримовый**: 1.065 volte. **Мейнстримный**: 790 volte. **Мейнстримовский**: 461 volte. **Мейнстриминг**: 20 volte.	**Мэйнстрим**: 4.986 volte. **Мэйнстримовый**: 476 volte. **Мэйнстримный**: 270 volte. **Мэйнстримовский**: 159 volte. **Мэйнстриминг**: 19 volte.
ruTenTen17	**Мейнстрим**: 7.746 volte. **Мейнстримный**: 774 volte. **Мейнстримовый**: 623 volte. **Мейнстримовский**: 252 volte. **Мейнстриминг**: 24 volte.	**Мэйнстрим**: 2.271 volte. **Мэйнстримный**: 119 volte. **Мэйнстримовый**: 241 volte. **Мэйнстримовский**: 80 volte. **Мэйнстриминг**: 13 volte.

7. Нетикет, нэтикет.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Нетикет; *Academic*: Нетикет, нэтикет.
NKRJa Osnovnoj korpus	**Нетикет**: 2 volte.	**Нэтикет**: No.
NKRJa Ustnyj korpus	**Нетикет**: No.	**Нэтикет**: No.
ruTenTen11	**Нетикет**: 285 volte.	**Нэтикет**: 23 volte.
ruTenTen17	**Нетикет**: 183 volte.	**Нэтикет**: 11 volte.

8. Селфи, сэлфи; селфить, сэлфить; селфиться, сэлфиться; селфи-камера, сэлфи-камера; селфи-палка, сэлфи-палка; селфи-стик, сэлфи-стик.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Селфи, селфить, селфиться, селфи-камера; *Academic*: No.
NKRJa Osnovnoj	**Селфи**: 15 volte. **Селфить**: No.	**Сэлфи**: 1 volta. **Сэлфить**: No.

korpus	**Селфиться**: No. **Селфи-камера**: No. **Селфи-палка**: 3 volte. **Селфи-стик**: No.	**Сэлфиться**: No. **Сэлфи-камера**: No. **Сэлфи-палка**: 3 volte. **Сэлфи-стик**: No.
NKRJa Ustnyj korpus	**Селфи**: 1 volta. **Селфить**: No. **Селфиться**: No. **Селфи-камера**: No. **Селфи-палка**: No. **Селфи-стик**: No.	**Сэлфи**: No. **Сэлфить**: No. **Сэлфиться**: No. **Сэлфи-камера**: No. **Сэлфи-палка**: No. **Сэлфи-стик**: No.
ruTenTen11	**Селфи**: 63 entrate al nominativo. In totale appare 679 volte. **Селфить**: 5 volte. **Селфиться**: No. **Селфи-камера**: No. **Селфи-палка**: No. **Селфи-стик**: No.	**Сэлфи**: 8 volte. **Сэлфить**: No. **Сэлфиться**: No. **Сэлфи-камера**: No. **Сэлфи-палка**: No. **Сэлфи-стик**: No.
ruTenTen17	**Селфи**: 20.564 entrate al nominativo. In totale appare 20.861 volte. **Селфить**: 6 volte. **Селфиться**: 61 volte. **Селфи-камера**: 282 volte. **Селфи-палка**: 448 volte. **Селфи-стик**: 32 volte.	**Сэлфи**: 819 volte. **Сэлфить**: No. **Сэлфиться**: No. **Сэлфи-камера**: 11 volte. **Сэлфи-палка**: 17 volte. **Сэлфи-стик**: No.

9. Треш, трэш; трешер, трэшер; трешевый, трэшевый; трешовый, трэшовый.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Треш, трэш, трешер, трешевый, трэшевый, трешовый; *Academic*: Треш, трэш.
NKRJa Osnovnoj korpus	**Треш**: 10 volte. **Трешер**: No. **Трешевый**: 2 volte. **Трешовый**: 1 volta.	**Трэш**: 72 volte. **Трэшер**: No. **Трэшевый**: 8 volte. **Трэшовый**: 1 volta.
NKRJa Ustnyj korpus	**Треш**: 2 volte. **Трешер**: No. **Трешевый**: No. **Трешовый**: 1 volta.	**Трэш**: 2 volte. **Трэшер**: No. **Трэшевый**: No. **Трэшовый**: No.
ruTenTen11	**Треш**: 7.744 volte. **Трешер**: 471 volte. **Трешевый**: 642 volte. **Трешовый**: 450 volte.	**Трэш**: 10.425 volte. **Трэшер**: 379 volte. **Трэшевый**: 1.118 volte. **Трэшовый**: 380 volte.
ruTenTen17	**Треш**: 5.097 volte. **Трешер**: 276 volte. **Трешевый**: 365volte. **Трешовый**: 496volte.	**Трэш**: 6.584 volte. **Трэшер**: 378 volte. **Трэшевый**: 720 volte. **Трэшовый**: 245 volte.

10. Фейк, фэйк; фейковый, фэйковый; фейк-новости, фэйк-новости; фейк-ньюс, фэйк-ньюс; фейкньюс, фэйкньюс.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Фейк, фейковый, фейк-ньюс;

		Academic: Фейк, фэйк, фейковый, фэйковый.
NKRJa Osnovnoj korpus	**Фейк**: 45 volte. **Фейкер**: No. **Фейковый**: 9 volte. **Фейк-новости**: No. **Фейк-ньюс**: No. **Фейкньюс**: No	**Фэйк**: 2 volte. **Фэйкер**: No. **Фэйковый**: No. **Фэйк-новости**: No. **Фэйк-ньюс**: No. **Фэйкньюс**: No.
NKRJa Ustnyj korpus	**Фейк**: 3 volte. **Фейкер**: No. **Фейковый**: No. **Фейк-новости**: No. **Фейк-ньюс**: No. **Фейкньюс**: No	**Фэйк**: 2 volte. **Фэйкер**: No. **Фэйковый**: No. **Фэйк-новости**: No. **Фэйк-ньюс**: No. **Фэйкньюс**: No.
ruTenTen11	**Фейк**: 8.639 volte. **Фейкер**: 36 volte. **Фейковый**: 2.083 volte. **Фейк-новости**: No. **Фейк-ньюс**: No. **Фейкньюс**: No.	**Фэйк**: 1.764 volte. **Фэйкер**: 5 volte. **Фэйковый**: 299 volte. **Фэйк-новости**: No. **Фэйк-ньюс**: No. **Фэйкньюс**: No.
ruTenTen17	**Фейк**: 16.090 volte. **Фейковый**: 6.794 volte. **Фейкер**: 18 volte. **Фейк-новости**: 16 volte. **Фейк-ньюс**: 24 volte. **Фейкньюс**: 13 volte.	**Фэйк**: 1.330 volte. **Фэйкер**: No. **Фэйковый**: 423 volte. **Фэйк-новости**: No. **Фэйк-ньюс**: 7 volte. **Фэйкньюс**: No.

11. Фейспалм, фэйспалм; фейспалмить, фэйспалмить.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Фейспалм, фэйспалм. *Academic*: Фейспалм, фэйспалм.
NKRJa Osnovnoj korpus	**Фейспалм**: No. **Фейспалмить**: No.	**Фэйспалм**: No. **Фэйспалмить**: No.
NKRJa Ustnyj korpus	**Фейспалм**: No. **Фейспалмить**: No.	**Фэйспалм**: No. **Фэйспалмить**: No.
ruTenTen11	**Фейспалм**: 489 volte. **Фейспалмить**: 16 volte.	**Фэйспалм**: 205 volte. **Фэйспалмить**: 6 volte.
ruTenTen17	**Фейспалм**: 997 volte. **Фейспалмить**: 92 volte.	**Фэйспалм**: 287 volte. **Фэйспалмить**: 6 volte.

12. Флешмоб, флэшмоб; флеш-моб, флэш-моб; флешмобер, флэшмобер; флешмоббер, флэшмоббер; флеш-мобер, флэш-мобер; флеш-моббер, флэш-моббер.		
Dizionario	Cartaceo: No.	*Vikislovar'*: флешмоб, флэшмоб, флешмобер; *Academic*: флешмоб, флэшмоб, флеш-моб, флэш-моб, флэшмоббер.
NKRJa Osnovnoj	**Флешмоб**: 79 volte. **Флеш-моб**: 4 volte.	**Флэшмоб**: 10 volte. **Флэш-моб**: No.

korpus	**Флешмобер**: 4 volte. **Флешмоббер**: No. **Флеш-мобер**: No. **Флеш-моббер**: No.	**Флэшмобер**: No. **Флэшмоббер**: No. **Флэш-мобер**: No. **Флэш-моббер**: No.
NKRJa Ustnyj korpus	**Флешмоб**: No. **Флеш-моб**: No. **Флешмобер**: No. **Флешмоббер**: No. **Флеш-мобер**: No. **Флеш-моббер**: No.	**Флэшмоб**: 2 volte. **Флэш-моб**: No. **Флэшмобер**: No. **Флэшмоббер**: No. **Флэш-мобер**: No. **Флэш-моббер**: No.
ruTenTen11	**Флешмоб**: 9.866 volte. **Флеш-моб**: 4.106 volte. **Флешмобер**: 146 volte. **Флешмоббер**: 68 volte. **Флеш-мобер**: 14 volte. **Флеш-моббер**: 25 volte.	**Флэшмоб**: 4.936 volte. **Флэш-моб**: 3.638 volte. **Флэшмобер**: 95 volte. **Флэшмоббер**: 113 volte. **Флэш-мобер**: 36 volte. **Флэш-моббер**: 38 volte.
ruTenTen17	**Флешмоб**: 14.312 volte. **Флеш-моб**: 2.561 volte. **Флешмобер**: 48 volte. **Флешмоббер**: 19 volte. **Флеш-мобер**: No. **Флеш-моббер**: No.	**Флэшмоб**: 3.925 volte. **Флэш-моб**: 1.959 volte. **Флэшмобер**: 30 volte. **Флэшмоббер**: 27 volte. **Флэш-мобер**: 11 volte. **Флэш-моббер**: 8 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Флешмоб**: 57.948 volte. **Флеш-моб**: 2.201 volte. **Флешмобер**: 17 volte. **Флешмоббер**: No. **Флеш-мобер**: No. **Флеш-моббер**: No.	**Флэшмоб**: 4.874 volte. **Флэш-моб**: 1.149 volte. **Флэшмобер**: No. **Флэшмоббер**: No. **Флэш-мобер**: No. **Флэш-моббер**: No.

13. Хейтер, хэйтер; хейтерить, хэйтерить; хейтить, хейтерс; хейтерство, хэйтерство; хейтинг, хэйтинг; хейтерский, хэйтерский.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Хейтер, хэйтер, хейтить, хейтинг хейтерский; *Academic*: Хейтер.
NKRJa Osnovnoj korpus	**Хейтер**: 4 volte. **Хейтерить**: 30 volte. **Хейтить**: 1 volta. **Хейтерство**: 1 volta. **Хейтинг**: No. **Хейтерский**: 1 volta.	**Хэйтер**: No. **Хэйтерить**: No. **Хэтить**: No. **Хэйтерство**: No. **Хэйтинг**: No. **Хэйтерский**: No.
NKRJa Ustnyj korpus	**Хейтер**: No. **Хейтерить**: No. **Хейтить**: 2 volte. **Хейтерство**: No. **Хейтинг**: No. **Хейтерский**: No.	**Хэйтер**: No. **Хэйтерить**: No. **Хэтить**: No. **Хэйтерство**: No. **Хэйтинг**: No. **Хэйтерский**: No.
ruTenTen11	**Хейтер**: 1243 volte. **Хейтерить**: 30 volte. **Хейтить**: 33 volte.	**Хэйтер**: 244 volte. **Хэйтерить**: 6 volte. **Хэтить**: No.

	Хейтерство: 52 volte. **Хейтинг**: 437 volte. **Хейтерский**: No.	**Хэйтерство**: No. **Хэйтинг**: No. **Хэйтерский**: No.
ruTenTen17	**Хейтер**: 2.372 volte. **Хейтерить**: 54 volte. **Хейтить**: 360 volte. **Хейтерство**: 160 volte. **Хейтинг**: 133 volte. **Хейтерский**: 35 volte.	**Хэйтер**: 168 volte. **Хэйтерить**: No. **Хэтить**: 51 volte. **Хэйтерство**: No. **Хэйтинг**: No. **Хэйтерский**: No.

14. Хештег, хэштэг, хэштег.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Хештег, хэштег; *Academic*: No.
NKRJa Osnovnoj korpus	**Хештег**: 2 volte.	**Хэштэг**: 10 volte. **Хэштег**: 5 volte.
NKRJa Ustnyj korpus	**Хештег**: No.	**Хэштэг**: No. **Хэштег**: No.
ruTenTen11	**Хештег**: 809 volte.	**Хэштэг**: 380 volte. **Хэштег**: 1.491 volte.
ruTenTen17	**Хештег**: 3.278 volte.	**Хэштэг**: 641 volte. **Хэштег**: 5.426 volte.

15. Шейм, шэим; шейминг, шэйминг.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Шейминг; *Academic*: No.
NKRJa Osnovnoj korpus	**Шейм**: 1 volta. **Шейминг**: No.	**Шэйм**: 6 volte. **Шэйминг**:
NKRJa Ustnyj korpus	**Шейм**: No. **Шейминг**: No.	**Шэйм**: No. **Шэйминг**: No.
ruTenTen11	**Шейм**: 233 volte. **Шейминг**: 14 volte.	**Шэйм**: 458 volte. **Шэйминг**: No.
ruTenTen17	**Шейм**: 129 volte. **Шейминг**: 19 volte.	**Шэйм**: 116 volte. **Шэйминг**: No.

4.4.2 Utilizzo di *mjagkij snak* (ь), *tvërdyj snak* (ъ) o omissione del segno

1. Абьюз, абъюз, абюз; абьюзер, абъюзер, абюзер; абьюзить, абъюзить, абюзить; абьюзный, абъюзный, абюзный.			
Dizionario	Cartaceo: No.	*Vikislovar'*: Абьюз, абьюзер.	*Academic*: Абьюз.
NKRJa Osnovnoj korpus	**Абьюз**: No. **Абьюзер**: No. **Абьюзить**: No. **Абьюзный**: 1 volta.	**Абъюз**: No. **Абъюзер**: No. **Абъюзить**: No. **Абъюзный**: No.	**Абюз**: No. **Абюзер**: No. **Абюзить**: No. **Абюзный**: No.
NKRJa Ustnyj korpus	**Абьюз**: No. **Абьюзер**: No.	**Абъюз**: No. **Абъюзер**: No.	**Абюз**: No. **Абюзер**: No.

	Абьюзить: No. **Абьюзный**: No.	**Абъюзить**: No. **Абъюзный**: No.	**Абюзить**: No. **Абюзный**: No.
ruTenTen11	**Абьюз**: 254 volte. **Абьюзер**: No. **Абьюзить**: No. **Абьюзный**: No.	**Абъюз**: 127 volte. **Абъюзер**: No. **Абъюзить**: No. **Абъюзный**: No.	**Абюз**: 6 volte. **Абюзер**: No. **Абюзить**: No. **Абюзный**: No.
ruTenTen17	**Абьюз**: 375 volte. **Абьюзер**: 248 volte. **Абьюзить**: 19 volte. **Абьюзный**: 10 volte.	**Абъюз**: 52 volte. **Абъюзер**: 7 volte. **Абъюзить**: **Абъюзный**: No.	**Абюз**: No. **Абюзер**: 5 volte. **Абюзить**: No. **Абюзный**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Абьюз**: 1.530 volte. **Абьюзер**: 2.054 volte. **Абьюзить**: 52 volte. **Абьюзный**: 12 volte.	**Абъюз**: 34 volte. **Абъюзер**: 24 volte. **Абъюзить**: No. **Абъюзный**: No.	**Абюз**: No. **Абюзер**: No. **Абюзить**: No. **Абюзный**: No.

4.4.3 Nomi composti uniti o separati da un trattino

1. Апгрейд, ап-грейд.
2. Бодиарт, боди-арт.
3. Бодипозитив, боди-позитив; бодипозитивный, боди-позитивный.
4. Бодишейминг, боди-шейминг; бодишеймер, боди-шеймер.
5. Кибербуллинг, кибер-буллинг; кибербуллер, кибер-буллер.
6. Офтоп, оф-топ; оффтоп, офф-топ; офтопик, оф-топик; оффтопик, офф-топик.
7. Флешмоб, флэшмоб; флеш-моб, флэш-моб; флешмобер, флэшмобер; флешмоббер, флэшмоббер; флеш-мобер, флэш-мобер; флеш-мобер, флэш-моббер.
   Si veda l'analisi 12 della sezione 4.4.1 (pagina 164).
8. Френдзона, френд-зона, френд зона.
9. Чайлдфри, чайлд-фри, чайлд фри.
10. Чилаут , чил-аут, чиллаут, чилл-аут

1. Апгрейд, ап-грейд		
Dizionario	Cartaceo: No.	*Vikislovar'*: Апгрейд; *Academic*: Апгрейд.
NKRJa Osnovnoj korpus	**Апгрейд**: 49 volte.	**Ап-грейд**: No.
NKRJa Ustnyj korpus	**Апгрейд**: No.	**Ап-грейд**: No.
ruTenTen11	**Апгрейд**: 39.915 volte.	**Ап-грейд**: 65 volte.
ruTenTen17	**Апгрейд**: 24.780 volte.	**Ап-грейд**: 19 volte.

2. Бодиарт, боди-арт		
Dizionario	Cartaceo: No.	*Vikislovar'*: Боди-арт; *Academic*: Бодиарт, боди-арт.
NKRJa Osnovnoj korpus	**Бодиарт**: 1 volta.	**Боди-арт**: No.
NKRJa Ustnyj korpus	**Бодиарт**: No.	**Боди-арт**: No.
ruTenTen11	**Бодиарт**: 2.277 volte.	**Боди-арт**: 3.128 volte.
ruTenTen17	**Бодиарт**: 1.700 volte.	**Боди-арт**: 7.129 volte.

3. Бодипозитив, боди-позитив; бодипозитивный, боди-позитивный.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Бодипозитив, бодипозитивный; *Academic*: No.
NKRJa Osnovnoj korpus	**Бодипозитив**: 2 volte. **Бодипозитивный**: No.	**Боди-позитив**: No. **Боди-позитивный**: No.
NKRJa Ustnyj korpus	**Бодипозитив**: No. **Бодипозитивный**: No.	**Боди-позитив**: No. **Боди-позитивный**: No.
ruTenTen11	**Бодипозитив**: No. **Бодипозитивный**: No.	**Боди-позитив**: No. **Боди-позитивный**: No.
ruTenTen17	**Бодипозитив**: 317 volte. **Бодипозитивный**: 56 volte.	**Боди-позитив**: 20 volte. **Боди-позитивный**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Бодипозитив**: 3.066 volte. **Бодипозитивный**: 778 volte.	**Боди-позитив**: 148 volte. **Боди-позитивный**: 10 volte.

4. Бодишейминг, боди-шейминг; бодишеймер, боди-шеймер.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Бодишейминг, бодишеймер; *Academic*: No.
NKRJa Osnovnoj korpus	**Бодишейминг**: No. **Бодишеймер**: No.	**Боди-шейминг**: No. **Боди-шеймер**: No.
NKRJa Ustnyj korpus	**Бодишейминг**: No. **Бодишеймер**: No.	**Боди-шейминг**: No. **Боди-шеймер**: No.
ruTenTen11	**Бодишейминг**: No. **Бодишеймер**: No.	**Боди-шейминг**: No. **Боди-шеймер**: No.
ruTenTen17	**Бодишейминг**: 16 volte. **Бодишеймер**: 5 volte.	**Боди-шейминг**: No. **Боди-шеймер**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Бодишейминг**: 393 volte. **Бодишеймер**: 64 volte.	**Боди-шейминг**: 5 volte. **Боди-шеймер**: No.

5. Кибербуллинг, кибер-буллинг; кибербуллер, кибер-буллер.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Кибербуллинг, кибербуллер. *Academic*: No.
NKRJa Osnovnoj korpus	**Кибербуллинг**: 2 volte. **Кибербулинг**: No. **Кибербуллер**: 1 volta.	**Кибер-буллинг**: No. **Кибер-булинг**: No. **Кибер-буллер**: No.
NKRJa Ustnyj korpus	**Кибербуллинг**: No. **Кибербулинг**: No. **Кибербуллер**: No.	**Кибер-буллинг**: No. **Кибер-булинг**: No. **Кибер-буллер**: No.

ruTenTen11	**Кибербуллинг**: 132 volte. **Кибербулинг**: 8 volte. **Кибербуллер**: No.	**Кибер-буллинг**: 20 volte. **Кибер-булинг**: No. **Кибер-буллер**: No.
ruTenTen17	**Кибербуллинг**: 559 volte. **Кибербулинг**: 7 volte. **Кибербуллер**: 12 volte.	**Кибер-буллинг**: 15 volte. **Кибер-булинг**: No. **Кибер-буллер**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Кибербуллинг**: 3.068 volte. **Кибербулинг**: 93 volte. **Кибербуллер**: 26 volte.	**Кибер-буллинг**: 34 volte. **Кибер-булинг**: No. **Кибер-буллер**: No.

6. Офтоп, оф-топ; оффтоп, офф-топ; офтопик, оф-топик; оффтопик, офф-топик.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Оффтопик, офтоп; *Academic*: Оффтопик, офтопик.
NKRJa Osnovnoj korpus	**Офтоп**: 4 volte. **Оффтоп**: 42 volte. **Офтопик**: 2 volte. **Оффтопик**: 15 volte.	**Оф-топ**: No. **Офф-топ**: 9 volte. **Оф-топик**: No. **Офф-топик**: 2 volte.
NKRJa Ustnyj korpus	**Офтоп**: No. **Оффтоп**: No. **Офтопик**: No. **Оффтопик**: No.	**Оф-топ**: No. **Офф-топ**: No. **Оф-топик**: No. **Офф-топик**: No.
ruTenTen11	**Офтоп**: 1.967 volte. **Оффтоп**: 13.991 volte. **Офтопик**: 714 volte. **Оффтопик**: 4.242 volte.	**Оф-топ**: 70 volte. **Офф-топ**: 1.317 volte. **Оф-топик**: e 15 volte. **Офф-топик**: 378 volte.
ruTenTen17	**Офтоп**: 1.241 volte. **Оффтоп**: 8.896 volte. **Офтопик**: 477 volte. **Оффтопик**: 2.639 volte.	**Оф-топ**: 39 volte. **Офф-топ**: 744 volte. **Оф-топик**: 15 volte. **Офф-топик**: 260 volte.

8. Френдзона, френд-зона, френд зона			
Dizionario	Cartaceo: No.	*Vikislovar'*: Френдзона.	*Academic*: Френдзона.
NKRJa Osnovnoj korpus	**Френдзона**: No.	**Френд-зона**: No.	**Френд зона**: 1 volta.
NKRJa Ustnyj korpus	**Френдзона**: 1 volta.	**Френд-зона**: No.	**Френд зона**:
ruTenTen11	**Френдзона**: 209 volte.	**Френд-зона**: 13 volte	**Френд зона**: 3 volte.
ruTenTen17	**Френдзона**: 511 volte.	**Френд-зона**: 44 volte.	**Френд зона**: 17 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Френдзона**: 1.217 volte.	**Френд-зона**: 20 volte.	**Френд зона**: 4 volte.

9. Чайлдфри, чайлд-фри, чайлд фри.			
Dizionario	Cartaceo: No.	*Vikislovar'*: Чайлдфри.	*Academic*: Чайлдфри, чайлд-фри.
NKRJa Osnovnoj korpus	**Чайлдфри**: No.	**Чайлд-фри**: 2 volte.	**Чайлд фри**: No.

NKRJa *Ustnyj korpus*	**Чайлдфри**: 2 volte.	**Чайлд-фри**: No.	**Чайлд фри**: No.
ruTenTen11	**Чайлдфри**: 2.044 volte.	**Чайлд-фри**: 234 volte.	**Чайлд фри**: 61 volte.
ruTenTen17	**Чайлдфри**: 1.837 volte.	**Чайлд-фри**: 182 volte.	**Чайлд фри**: 50 volte.

10. Чилаут , чил-аут, чиллаут, чилл-аут.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Чилаут, чиллаут; *Academic*: Чилаут, чиллаут.
NKRJa *Osnovnoj korpus*	**Чилаут**: No. **Чиллаут**: 2 volte.	**Чил-аут**: 2 volte. **Чилл-аут**: 4 volte.
NKRJa *Ustnyj korpus*	**Чилаут**: 1 volta. **Чиллаут**: No.	**Чил-аут**: No. **Чилл-аут**: No.
ruTenTen11	**Чилаут**: 1.432 volte. **Чиллаут**: 625 volte.	**Чил-аут**: 581 volte. **Чилл-аут**: 827 volte.
ruTenTen17	**Чилаут**: 811 volte. **Чиллаут**: 358 volte.	**Чил-аут**: 211 volte. **Чилл-аут**: 280 volte.

4.4.4 Utilizzo della consonante doppia e singola

1. Блогер, блоггер; блогерство, блоггерство; блогерский, блоггерский; блогерный, блоггерный.

2. Офтоп, оффтоп; оф-топ, офф-топ; офтопик, оффтопик; оф-топик, офф-топик.
Si veda l'analisi 6 della sezione 4.4.3 (pagina 169).

3. Тролинг, троллинг; тролинговый, троллинговый; тролить, троллить; тролинговать, троллинговать; тролировать, троллировать.

4. Флешмобер, флешмоббер; флеш-мобер, флеш-моббер; флэшмобер, флэшмоббер; флэш-мобер; флэш-моббер.
Si veda l'analisi 12 della sezione 4.4.1 (pagina 164).

5. Фоловер, фолловер; фоловинг, фолловинг; фоловить, фолловить.

6. Челендж, челлендж; челенж, челленж; челенджер, челленджер, челенжер, челленжер.

7. Чилаут, чиллаут; чил-аут, чилл-аут.
Si veda l'analisi 10 della sezione 4.4.3 (pagina 170).

8. Чилить, чиллить; чилиться, чиллиться.

1. Блогер, блоггер; блогерство, блоггерство; блогерский, блоггерский; блогерный, блоггерный.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Блогер, блоггер, блогерство, блогерский; *Academic*: Блогер, блоггер.
NKRJa Osnovnoj korpus	**Блогер**: 216 volte. **Блогерство**: No. **Блогерский**: No. **Блогерный**: No.	**Блоггер**: 36 volte. **Блоггерство**: No. **Блоггерский**: No. **Блоггерный**: No.
NKRJa Ustnyj korpus	**Блогер**: 16 volte. **Блогерство**: No. **Блогерский**: No. **Блогерный**: No.	**Блоггер:** 3 volte. **Блоггерство**: No. **Блоггерский**: No. **Блоггерный**: No.
ruTenTen11	**Блогер**: 76.308 volte. **Блогерство**: 425 volte. **Блогерский**: 1.404 volte. **Блогерный**: 16 volte.	**Блоггер**: 79.782 volte. **Блоггерство**: 691 volte. **Блоггерский**: 1.648 volte. **Блоггерный**: 13 volte.
ruTenTen17	**Блогер**: 55.129 volte. **Блогерство**: 295 volte. **Блогерский**: 772 volte. **Блогерный**: No.	**Блоггер**: 30.955 volte. **Блоггерство**: 267 volte. **Блоггерский**: 540 volte. **Блоггерный**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Блогер**: 315.727 volte. **Блогерство**: 1.356 volte. **Блогерский**: 1.363 volte. **Блогерный**: No.	**Блоггер**: 6.332 volte. **Блоггерство**: 54 volte. **Блоггерский**: 40 volte. **Блоггерный**: No.

3. Тролинг, троллинг; тролинговый, троллинговый; тролить, троллить; тролинговать, троллинговать; тролировать, троллировать.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Троллинг, троллинговый; *Academic*: Троллинг.
NKRJa Osnovnoj korpus	**Тролинг**: No. **Тролинговый**: No. **Тролить**: 1 volta. **Тролинговать**: No.	**Троллинг**: 18 volte. **Троллинговый**: No. **Троллить**: 6 volte. **Троллинговать**: No.
NKRJa Ustnyj korpus	**Тролинг**: No. **Тролинговый**: No. **Тролить**: No. **Тролинговать**: No.	**Троллинг**: No. **Троллинговый**: No. **Троллить**: No. **Троллинговать**: No.
ruTenTen11	**Тролинг**: 739 volte. **Тролинговый**: 69 volte. **Тролить**: 1.477 volte. **Тролинговать**: 5 volte.	**Троллинг**: 15.279 volte. **Троллинговый**: 1.317 volte. **Троллить**: 4.467 volte. **Троллинговать**: 74 volte.
ruTenTen17	**Тролинг**: 660 volte. **Тролинговый**: 53 volte. **Тролить**: 1.115 volte. **Тролинговать**: No.	**Троллинг**: 17.973 volte. **Троллинговый**: 1.939 volte. **Троллить**: 4.386 volte. **Троллинговать**: 52 volte.

5. Фоловер, фолловер; фоловинг, фолловинг; фоловить, фолловить.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Фолловер, фолловинг, фолловить. *Academic*: Фоловер, фолловер.
NKRJa Osnovnoj korpus	**Фоловер**: No. **Фоловить**: No. **Фоловинг**: No.	**Фолловер**: 5 volte. **Фолловить**: No. **Фолловинг**: No.
NKRJa Ustnyj korpus	**Фоловер**: No. **Фоловить**: No. **Фоловинг**: No.	**Фолловер**: No. **Фолловить**: No. **Фолловинг**: No.
ruTenTen11	**Фоловер**: 1.259 volte. **Фоловить**: 481 volte. **Фоловинг**: 191 volte.	**Фолловер**: 5.355 volte. **Фолловить**: 1.219 volte. **Фолловинг**: 874 volte.
ruTenTen17	**Фоловер**: 532 volte. **Фоловить**: 126 volte. **Фоловинг**: 53 volte.	**Фолловер**: 4.038 volte. **Фолловить**: 464 volte. **Фолловинг**: 375 volte.

6. Челендж, челлендж; челенж, челленж; челенджер, челленджер, челенжер, челленжер.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Челендж, челлендж, челленджер; *Academic*: Челендж, челенджер, челленджер.
NKRJa Osnovnoj korpus	**Челендж**: 4 volte. **Челенж**: No. **Челенджер**: 6 volte. **Челенжер**: No.	**Челлендж**: 3 volte. **Челленж**: No. **Челленджер**: 102 volte. **Челленжер**: 1 volta.
NKRJa Ustnyj korpus	**Челендж**: No. **Челенж**: No. **Челенджер**: No. **Челенжер**: No.	**Челлендж**: 1 volta. **Челленж**: No. **Челленджер**: 5 volte. **Челленжер**: No.
ruTenTen11	**Челендж**: 571 volte. **Челенж**: 58 volte. **Челенджер**: 618 volte. **Челенжер**: 68 volte.	**Челлендж**: 1.886 volte. **Челленж**: 39 volte. **Челленджер**: 5.106 volte. **Челленжер**: 69 volte.
ruTenTen17	**Челендж**: 457 volte. **Челенж**: 39 volte. **Челенджер**: 317 volte. **Челенжер**: 40 volte.	**Челлендж**: 1.944 volte. **Челленж**: 54 volte. **Челленджер**: 3.289 volte. **Челленжер**: 25 volte.

8. Чилить, чиллить; чилиться, чиллиться.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Чилить, чиллить, чилиться; *Academic*: No.
NKRJa Osnovnoj korpus	**Чилить**: No. **Чилиться**: No.	**Чиллить**: No. **Чиллиться**: No.
NKRJa Ustnyj korpus	**Чилить**: No. **Чилиться**: No.	**Чиллить**: No. **Чиллиться**: No.
ruTenTen11	**Чилить**: 1.220 volte. **Чилиться**: 7 volte.	**Чиллить**: 11 volte. **Чиллиться**: No.

ruTenTen17	**Чилить**: 683 volte. **Чилиться**: 107 volte.	**Чиллить**: 14 volte. **Чиллиться**: No.

4.4.5 Alternanza delle consonanti C (s) e З (z)

1. Дислайк, дизлайк; дислайкнуть, дизлайкнуть; дислайкать, дислайкать.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Дислайк, дизлайк; *Academic*: No.
NKRJa *Osnovnoj korpus*	**Дислайк**: No. **Дислайкнуть**: No. **Дислайкать**: No.	**Дизлайк**: 1 volta. **Дизлайкнуть**: No. **Дизлайкать**: No.
NKRJa *Ustnyj korpus*	**Дислайк**: No. **Дислайкнуть**: No. **Дислайкать**: No.	**Дизлайк**: No. **Дизлайкнуть**: No. **Дизлайкать**: No.
ruTenTen11	**Дислайк**: 40 volte. **Дислайкнуть**: No. **Дислайкать**: No.	**Дизлайк**: 18 volte. **Дизлайкнуть**: No. **Дизлайкать**: No.
ruTenTen17	**Дислайк**: 196 volte. **Дислайкнуть**: No. **Дислайкать**: No.	**Дизлайк**: 762 volte. **Дизлайкнуть**: 11 volte. **Дизлайкать**: 20 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Дислайк**: 367 volte. **Дислайкнуть**: No. **Дислайкать**: No.	**Дизлайк**: 4.953 volte. **Дизлайкнуть**: 17 volte. **Дизлайкать**: 26 volte.

4.4.6 Alternanza delle consonanti Д (d) e Т (t)

1. Апгрейд, апгрейдить; апгрейт, апгрейтить.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Апгрейд, апгрейдить; *Academic*: Апгрейд, апгрейдить.
NKRJa *Osnovnoj korpus*	**Апгрейд**: 51 volte. **Апгрейдить**: 4 volte.	**Апгрейт**: 1 volta. **Апгрейтить**: No.
NKRJa *Ustnyj korpus*	**Апгрейд**: No. **Апгрейдить**: No.	**Апгрейт**: No. **Апгрейтить**: No.
ruTenTen11	**Апгрейд**: 39.915 volte. **Апгрейдить**: 3.146 volte.	**Апгрейт**: 968 volte. **Апгрейтить**: 44 volte.
ruTenTen17	**Апгрейд**: 24.780 volte. **Апгрейдить**: 2.080 volte.	**Апгрейт**: 539 volte. **Апгрейтить**: 25 volte.

4.4.7 Alternanza della consonante Ж (ž) e del nesso consonantico ДЖ (dž)

1.	Криндж, кринж; кринджовый, кринжовый.
2.	Челлендж, челленж; челленджер, челленжер.

1. Криндж, кринж; кринджовый, кринжовый.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Кринж, кринжовый; *Academic*: No.
NKRJa *Osnovnoj korpus*	**Криндж**: No. **Кринджовый**: No.	**Кринж**: No. **Кринжовый**: No.
NKRJa *Ustnyj korpus*	**Криндж**: No. **Кринджовый**: No.	**Кринж**: No. **Кринжовый**: No.
ruTenTen11	**Криндж**: No. **Кринджовый**: No.	**Кринж**: No. **Кринжовый**: No.
ruTenTen17	**Криндж**: No. **Кринджовый**: No.	**Кринж**: No. **Кринжовый**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Криндж**: No. **Кринджовый**: No.	**Кринж**: 202 volte. **Кринжовый**:48 volte.

2. Челлендж, челленж; челленджер, челленжер.		
Dizionario	Cartaceo: No.	*Vikislovar'*: Челендж, челлендж, челленджер; *Academic*: Челендж, челенджер, челленджер.
NKRJa *Osnovnoj korpus*	**Челлендж**: 3 volte. **Челленджер**: 102 volte.	**Челленж**: No. **Челленжер**: 1 volta.
NKRJa *Ustnyj korpus*	**Челлендж**: 1 volta. **Челленджер**: 5 volte.	**Челленж**: No. **Челленжер**: No.
ruTenTen11	**Челлендж**: 1.886 volte. **Челленджер**: 5.106 volte.	**Челленж**: 39 volte. **Челленжер**: 69 volte.
ruTenTen17	**Челлендж**: 1.944 volte. **Челленджер**: 3.289 volte.	**Челленж**: 54 volte. **Челленжер**: 25 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Челлендж**: 18.370 volte. **Челленджер**: 6.153 volte.	**Челленж**: 84 volte. **Челленжер**: 13 volte.

4.5 I sostantivi: i processi di formazione delle parole, i prestiti e le irregolarità emerse

4.5.1 Il prestito traslitterato

1.	Абьюз	22.	Селфи-камера
2.	Апгрейд	23.	Селфи-стик
3.	Баттхерт	24.	Сетикет
4.	Боди-арт	25.	Стайл
5.	Бодипозитив	26.	Стрим
6.	Вайб	27.	Трабл
7.	Гейм	28.	Трип
8.	Геймплей	29.	Треш
9.	Дакфейс	30.	Фейк
10.	Дизлайк	31.	Фейк-ньюс
11.	Краш	32.	Фейспалм
12.	Кринж	33.	Флуд
13.	Лайк	34.	Флешмоб
14.	Лайфхак	35.	Хайп
15.	Мейнстрим	36.	Хэштег
16.	Муд	37.	Челлендж
17.	Нетикет	38.	Чилаут
18.	Оффтопик (оффтоп)	39.	Шэим
19.	Пранк	40.	Юзернейм
20.	Свайп	41.	Юзерпик
21.	Селфи		

4.5.2 Prestiti di *nomina agentis* e *nomina actionis* in *-er*

1.	**Абьюзер**: *To abuse* + *-er*. Qualcuno che abusa.	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 248 volte.
2.	**Апгрейдер**: *To upgrade* + *-er*. Qualcuno che fa un upgrade (poco usato in Inglese).	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: 82 volte. *ruTenTen17*: 39 volte.
3.	**Блогер**: *Blog/ to blog* + *-er*. Qualcuno che scrive un blog, un blogger.	*Osnovnoj korpus*: 190 volte. *Ustnyj korpus*: 16 volte. *ruTenTen11*: 76.308 volte.

		ruTenTen17: 55.129 volte.
5.	**Кибербуллер**: *Cyber + to bully + -er.* Un cyber-bullo.	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 12 volte.
6.	**Геймер**: *To game + -er.* Un giocatore.	*Osnovnoj korpus*: 117 volte. *Ustnyj korpus*: No. *ruTenTen11*: 65.289 volte. *ruTenTen17*: 49.774 volte.
7.	**Зумер**: Persona appartenente alla Generazione Z. **Зумер**: *To zoom + -er.* Un oggetto che permette di fare lo zoom.	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 387 volte. *ruTenTen17*: 252 volte.
8.	**Инфлюенсер**: *To influence + -er.* Qualcuno che influenza.	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 31 volte.
9.	**Лайкер**: *To like + -er.* Qualcuno che mette “Mi piace”.	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 205 volte. *ruTenTen17*: 180 volte.
10.	**Лайфхакер**: *Lifehack/ to lifehack + -er.* Qualcuno che usa dei trucchetti per semplificarsi la vita.	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 355 volte. *ruTenTen17*: 620 volte.
11.	**Лузер**: *To loose + -er.* Un perdente.	*Osnovnoj korpus*: 17 volte. *Ustnyj korpus*: 5 volte. *ruTenTen11*: 11.138 volte. *ruTenTen17*: 7.020 volte.
12.	**Пранкер**: *To prank + -er.* Qualcuno che fa uno scherzo o una burla telefonica.	*Osnovnoj korpus*: 8 volte. *Ustnyj korpus*: No. *ruTenTen11*: 1.354 volte. *ruTenTen17*: 1.785 volte.
13.	**Спойлер**: *To spoil + -er.* Una persona che rovina qualcosa; Anticipazione di una parte di film, serie TV, libro, ecc., soprattutto del finale; In aeronautica indica il *diruttore* o *disruttore* degli aerei; In ambito automobilistico s’intende una o più parti aerodinamiche dell’auto.	*Osnovnoj korpus*: 103 volte. *Ustnyj korpus*: 10 volte. *ruTenTen11*: 53.141 volte. *ruTenTen17*: 31.698 volte.
14.	**Стайлер**: *To style + -er.* Parrucchiere; Un utensile per acconciare i capelli.	*Osnovnoj korpus*: 8 volte. *Ustnyj korpus*: No. *ruTenTen11*: 2.693 volte. *ruTenTen17*: 1.801 volte.
15.	**Стример**: *To stream + -er.* Una persona che fa una diretta streming. Dispositivo di memoria su nastri magnetici.	*Osnovnoj korpus*: 15 volte. *Ustnyj korpus*: 1 volta. *ruTenTen11*: 6.714 volte. *ruTenTen17*: 6.117 volte.
16.	**Траблмейкер**:*Trouble + to make + -er.* Una persona che crea problemi.	*Osnovnoj korpus*: No. *Ustnyj korpus*: No.

		ruTenTen11: 99 volte. *ruTenTen17*: 46 volte.
17.	**Фейкер**: *To fake + -er.* Una persona che finge.	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 36 volte. *ruTenTen17*: 18 volte.
18.	**Флешмобер**: *Flashmob + -er/ flash + to mob + -er.* Un partecipante ad un flash mob.	*Osnovnoj korpus*: 4 volte. *Ustnyj korpus*: No. *ruTenTen11*: 146 volte. *ruTenTen17*: 48 volte.
19.	**Фолловер**: *To follow + -er.* Un seguace.	*Osnovnoj korpus*: 5 volte. *Ustnyj korpus*: No. *ruTenTen11*: 5.355 volte. *ruTenTen17*: 4.038 volte.
20.	**Хейтер**: *To hate + -er.* Una persona che su internet usa un linguaggio aggressivo, volgare, espressioni di odio o insulta.	*Osnovnoj korpus*: 4 volte. *Ustnyj korpus*: No. *ruTenTen11*: 1.240 volte. *ruTenTen17*: 2.372 volte.
21.	**Хипстер**: Hipster. Un giovane anticonformista.	*Osnovnoj korpus*: 65 volte. *Ustnyj korpus*: 1 volta. *ruTenTen11*: 3.824 volte. *ruTenTen17*: 4.366 volte.
22.	**Юзер**: *To use + -er.* Un utente, una persona che utilizza qualcosa.	*Osnovnoj korpus*: 88 volte. *Ustnyj korpus*: 3 volte. *ruTenTen11*: 96.274 volte. *ruTenTen17*: 40.529 volte.

4.5.3 Prestiti russificati in *–ств(о)*, indicanti sostantivi neutri astratti

1.	**Блогерство**:	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 425 volte. *ruTenTen17*: 295 volte. *Timestamped JSI web corpus 2014-2021*: 1.356 volte.
2.	**Геймерство**	*Osnovnoj korpus*: 8 volte. *Ustnyj korpus*: No. *ruTenTen11*: 225 volte. *ruTenTen17*: 114 volte. *Timestamped JSI web corpus 2014-2021*: 55 volte.
3.	**Лузерство**	*Osnovnoj korpus*: 4 volte. *Ustnyj korpus*: 2 volte. *ruTenTen11*: 451 volte. *ruTenTen17*: 169 volte. *Timestamped JSI web corpus 2014-2021*: 167 volte.
4.	**Спойлерство**	*Osnovnoj korpus*: No. *Ustnyj korpus*: No.

		ruTenTen11: 60 volte. *ruTenTen17*: 50volte. *Timestamped JSI web corpus 2014-2021*: 155 volte.
5.	**Хейтерство**	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: 160 volte. *ruTenTen17*: 160 volte. *Timestamped JSI web corpus 2014-2021*: 330 volte.
6.	**Хипстерство**	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 50 volte. *ruTenTen17*: 59 volte. *Timestamped JSI web corpus 2014-2021*: 47 volte.

4.5.4 Forme differenti dello stesso sostantivo

1. Крипово, криповый, крипота, крипотный		
NKRJa *Osnovnoj korpus*	**Крипово**: 4 volte. **Криповый**: No.	**Крипота**: 1 volta. **Крипотный**: No.
NKRJa *Ustnyj korpus*	**Крипово**: No. **Криповый**: No.	**Крипота**: No. **Крипотный**: No.
ruTenTen11	**Крипово**: No. **Криповый**: 21 volte.	**Крипота**: 34 volte. **Крипотный**: No.
ruTenTen17	**Крипово**: 22 volte. **Криповый**: 71 volte.	**Крипота**: 116 volte. **Крипотный**: 17 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Крипово**: 79 volte. **Криповый**: 259 volte.	**Крипота**: 64 volte. **Крипотный**: No.

4.5.5 Sostantivi plurali traslitterati in russo e declinati come sostantivi singolari

1.	Геймс, геймз;	5.	Спойлерс;
2.	Лайкс;	6.	Траблс, траблз;
3.	Лузерс;	7.	Хейтерс;
4.	Свайпс;	8.	Юзерс.

1. Геймс, геймз		
NKRJa Osnovnoj korpus	**Геймс**: 1 volta.	**Геймз**: No.
NKRJa Ustnyj korpus	**Геймс**: 1 volta.	**Геймз**: No.
ruTenTen11	**Геймс**: 517 volte.	**Геймз**: 128 volte.
ruTenTen17	**Геймс**: 719 volte.	**Геймз**: 71 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Геймс**: 85 volte.	**Геймз**: 55 volte.

2. Лайкс	
NKRJa Osnovnoj korpus	**Лайкс**: No.
NKRJa Ustnyj korpus	**Лайкс**: No.
ruTenTen11	**Лайкс**: 126 volte.
ruTenTen17	**Лайкс**: 16 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Лайкс**: 66 volte.

3. Лузерс	
NKRJa Osnovnoj korpus	**Лузерс**: No.
NKRJa Ustnyj korpus	**Лузерс**: No.
ruTenTen11	**Лузерс**: 11 volte.
ruTenTen17	**Лузерс**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Лузерс**: No.

4. Свайпс	
NKRJa Osnovnoj korpus	**Свайпс**: No.
NKRJa Ustnyj korpus	**Свайпс**: No.
ruTenTen11	**Свайпс**: 12 volte.
ruTenTen17	**Свайпс**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Свайпс**: No.

5. Спойлерс	
NKRJa Osnovnoj korpus	**Спойлерс**: No.
NKRJa Ustnyj korpus	**Спойлерс**: No.
ruTenTen11	**Спойлерс**: 5 volte.
ruTenTen17	**Спойлерс**: 12 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Спойлерс**: No.

6. Траблс, траблз		
NKRJa Osnovnoj korpus	**Траблс**: No.	**Траблз**: No.
NKRJa Ustnyj korpus	**Траблс**: No.	**Траблз**: No.
ruTenTen11	**Траблс**: 16 volte.	**Траблз**: 19 volte.
ruTenTen17	**Траблс**: 9 volte.	**Траблз**: No.
Timestamped JSI web corpus 2014-2021 Russian	**Траблс**: No.	**Траблз**: No.

7. Хейтерс	
NKRJa Osnovnoj korpus	**Хейтерс**: No.

NKRJa Ustnyj korpus	**Хейтерс**: No.
ruTenTen11	**Хейтерс**: 5 volte.
ruTenTen17	**Хейтерс**: 5 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Хейтерс**: No.

8. Юзерс	
NKRJa Osnovnoj korpus	**Юзерс**: No.
NKRJa Ustnyj korpus	**Юзерс**: No
ruTenTen11	**Юзерс**: 34 volte.
ruTenTen17	**Юзерс**: 17 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Юзерс**: No.

4.5.6 Prestiti traslitterati o calchi parziali dello stesso nome composto

1. Селфи-палка, селфи-стик		
NKRJa Osnovnoj korpus	**Селфи-палка**: 3 volte.	**Селфи-стик**: No.
NKRJa Ustnyj korpus	**Селфи-палка**: No.	**Селфи-стик**: No.
ruTenTen11	**Селфи-палка**: No.	**Селфи-стик**: No.
ruTenTen17	**Селфи-палка**: 435 volte.	**Селфи-стик**: 32 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Селфи-палка**: 1.316 volte.	**Селфи-стик**: 23 volte.

2. Фейк-новости, фейк-ньюс, фейкньюс		
NKRJa Osnovnoj korpus	**Фейк-новости**: No.	**Фейк-ньюс**: No. **Фейкньюс**: No.
NKRJa Ustnyj korpus	**Фейк-новости**: No.	**Фейк-ньюс**: No. **Фейкньюс**: No.
ruTenTen11	**Фейк-новости**: No.	**Фейк-ньюс**: No. **Фейкньюс**: No.
ruTenTen17	**Фейк-новости**: 16 volte.	**Фейк-ньюс**: 24 volte. **Фейкньюс**: 13 volte.
Timestamped JSI web corpus 2014-2021 Russian	**Фейк-новости**: 388 volte.	**Фейк-ньюс**: 2.020 volte. **Фейкньюс**: 379 volte.

4.6 Gli aggettivi: processi derivazionali e forme differenti con lo stesso significato

1.	**Абьюзный**: Абьюз + -н- + ый	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 10 volte. *Timestamped JSI web corpus 2014-2021*: 12 volte.

2.	**Апгрейдный**: Апгрейд + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 74 volte. *ruTenTen17*: 30 volte. *Timestamped JSI web corpus 2014-2021*: 6 volte.
3.	**Апгрейдовый**: Апгрейд + -ов- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 20 volte. *ruTenTen17*: No. *Timestamped JSI web corpus 2014-2021*: No.
4.	**Блогерный**: Блогер + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 16 volte. *ruTenTen17*: No. *Timestamped JSI web corpus 2014-2021*: No.
5.	**Блогерский**: Блогер + -ск- + ий	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 1.404 volte. *ruTenTen17*: 772 volte. *Timestamped JSI web corpus 2014-2021*: 1.363 volte.
6.	**Бодипозитивный**: Бодипозитив + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No.. *ruTenTen17*: 56 volte. *Timestamped JSI web corpus 2014-2021*: 778 volte.
7.	**Геймерский**: Геймер + -ск- + ий	*Osnovnoj korpus*: 4 volte. *Ustnyj korpus*: No. *ruTenTen11*: 5.750 volte. *ruTenTen17*: 3.799 volte. *Timestamped JSI web corpus 2014-2021*: 5.622 volte.
8.	**Геймплейный**: Геймплей + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 3.429 volte. *ruTenTen17*: 2.882 volte. *Timestamped JSI web corpus 2014-2021*: 6.428 volte.
9.	**Криповый**: Крип + -ов- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 21 volte. *ruTenTen17*: 71 volte. *Timestamped JSI web corpus 2014-2021*: 259 volte.
10.	**Крипотный**: Крипот(а) + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 17 volte. *Timestamped JSI web corpus 2014-2021*: No.
11.	**Лайковый**: Лайк + -ов- + ый L'aggettivo è relative non solo ai *like*, ma si riferisce anche agli altri due significati della parola, ossia "cane eschimese" e "pelle di capretto".	*Osnovnoj korpus*: 354 volte. *Ustnyj korpus*: No. *ruTenTen11*: 1.781 volte. *ruTenTen17*: 1.030 volte. *Timestamped JSI web corpus 2014-2021*: 114 volte.

12.	**Лузерный**: Лузер + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 59 volte. *ruTenTen17*: No. *Timestamped JSI web corpus 2014-2021*: No.
13.	**Лузерский**: Лузер + -ск- + ий	*Osnovnoj korpus*: 3 volte. *Ustnyj korpus*: No. *ruTenTen11*: 362 volte. *ruTenTen17*: 140 volte. *Timestamped JSI web corpus 2014-2021*: 78 volte.
14.	**Мейнстримный**: Мейнстрим + -н- + ый	*Osnovnoj korpus*: 13 volte. *Ustnyj korpus*: No. *ruTenTen11*: 790 volte. *ruTenTen17*: 774 volte. *Timestamped JSI web corpus 2014-2021*: 2.057 volte.
15.	**Мейнстримовский**: Мейнстрим + -овск- + ий	*Osnovnoj korpus*: 10 volte. *Ustnyj korpus*: No. *ruTenTen11*: 461 volte. *ruTenTen17*: 252 volte. *Timestamped JSI web corpus 2014-2021*: 372 volte.
16.	**Мейнстримовый**: Мейнстрим + -ов- + ый	*Osnovnoj korpus*: 2 volte. *Ustnyj korpus*: No. *ruTenTen11*: 1.065 volte. *ruTenTen17*: 623 volte. *Timestamped JSI web corpus 2014-2021*: 709 volte.
17.	**Спойлерный**: Спойлер + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 325 volte. *ruTenTen17*: 404 volte. *Timestamped JSI web corpus 2014-2021*: 94 volte.
18.	**Спойлерский**: Спойлер + -ск- + ий	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 91 volte. *ruTenTen17*: 45 volte. *Timestamped JSI web corpus 2014-2021*: I 190 volte.
19.	**Стайлинговый**: Стайлинг + -ов- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 977 volte. *ruTenTen17*: 870 volte. *Timestamped JSI web corpus 2014-2021*: 189 volte.
20.	**Стайловый**: Стайл + -ов- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 23 volte. *ruTenTen17*: 7 volte. *Timestamped JSI web corpus 2014-2021*: No.
21.	**Стримерный**: Стример + -н- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: 1 volta. *ruTenTen11*: 703 volte. *ruTenTen17*: 322 volte. *Timestamped JSI web corpus 2014-2021*: 5 volte.

22.	**Стриминговый**: Стриминг + -ов-+ ый	*Osnovnoj korpus*: 2 volte. *Ustnyj korpus*: No. *ruTenTen11*: 114 volte. *ruTenTen17*: 1.692 volte. *Timestamped JSI web corpus 2014-2021*: 22.355 volte.
23.	**Стримовский**: Стрим + -овск- + ий	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 93 volte. *ruTenTen17*: 49 volte. *Timestamped JSI web corpus 2014-2021*: I 6 volte.
24.	**Стримовый**: Стрим + -ов- + ый	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 5 volte. *ruTenTen17*: 51 volte. *Timestamped JSI web corpus 2014-2021*: 7 volte.
25.	**Трэшевый**: Трэш + -ев- + ый	*Osnovnoj korpus*: 8 volte. *Ustnyj korpus*: No. *ruTenTen11*: 1.118 volte. *ruTenTen17*: 380 volte. *Timestamped JSI web corpus 2014-2021*: 133 volte.
26.	**Трэшовый**: Трэш + -ов- + ый	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: 720 volte. *ruTenTen17*: 245 volte. *Timestamped JSI web corpus 2014-2021*: 193 volte.
27.	**Фейковый**: Фейк + -ов- + ый	*Osnovnoj korpus*: 9 volte. *Ustnyj korpus*: No. *ruTenTen11*: 2.110 volte. *ruTenTen17*: 6.794 volte. *Timestamped JSI web corpus 2014-2021*: 78.904 volte.
28.	**Хейтерский**: Хейтер + -ск- + ий	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: 35 volte. *ruTenTen17*: 35 volte. *Timestamped JSI web corpus 2014-2021*: 411 volte.
29.	**Хипстерский**: Хипстер + -ск- + ий	*Osnovnoj korpus*: 7 volte. *Ustnyj korpus*: No. *ruTenTen11*: 595 volte. *ruTenTen17*: 1.126 volte. *Timestamped JSI web corpus 2014-2021*: 1.415 volte.

4.7 I verbi: processi derivazionali e di russificazione

1.	**Абьюзить**: Abusare. Абьюз + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 19 volte.
2.	**Агрить**: Fare arrabbiare, provocare. Агр (da *anger*) + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 352 volte. *ruTenTen17*: 321 volte.
3.	**Агриться**: Arrabbiarsi, irritarsi. Агр + -и- + -ть- + -ся	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 224 volte. *ruTenTen17*: 109 volte.
4.	**Апгрейдировать**: Fare un upgrade. Апгрейд + -ирова- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 36 volte. *ruTenTen17*: 16 volte.
5.	**Апгрейдить**: Fare un upgrade. Апгрейд + -и- + -ть	*Osnovnoj korpus*: 2 volte. *Ustnyj korpus*: No. *ruTenTen11*: 3.146 volte. *ruTenTen17*: 2.080 volte.
6.	**Апгрейдиться**: Farsi un upgrade, migliorarsi. Апгрейд + -и- + -ть- + -ся	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 681 volte. *ruTenTen17*: 354 volte.
7.	**Апгрейднуть**: Fare un upgrade. Апгрейд + -ну- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 59 volte. *ruTenTen17*: 47 volte.
8.	**Апгрейтить**: Fare un upgrade- Апгрейт + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*:44 volte. *ruTenTen17*: 25 volte.
9.	**Байтить**: Copiare lo stile di qualcuno (dall'inglese *to bite one's style*). Байт + -и- + -ть	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: 22 volte. *ruTenTen17*: 7 volte.
10.	**Гамать**: Giocare, specialmente con i videogiochi. Гам (da *game*) + -а- + -ть	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: 1 volta. *ruTenTen11*: 1.864 volte. *ruTenTen17*: 802 volte.
11.	**Геймить**: Giocare, specialmente con i videogiochi. Гейм + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 16 volte. *ruTenTen17*: 34 volte.
12.	**Дизлайкать**: mettere un "non mi piace", non gradire. Дизлайк + -а- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No.

		ruTenTen17: 20 volte.
13.	**Дизлайкнуть**: mettere un “non mi piace”, non gradire. Дизлайк + -ну- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 11 volte.
14.	**Пранковать**: Fare uno scherzo telefonico, scherzare. Пранк + -ова- + -ть	*Osnovnoj korpus*: 2 volte. *Ustnyj korpus*: No. *ruTenTen11*: 5 volte. *ruTenTen17*: 5 volte.
15.	**Рофлить**: Rotolarsi a terra dalle risate, ridere a crepapelle. Рофл (da *ROFL - Rolling On the Floor Laughing*) + -и- + -ть	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 27 volte.
16.	**Свайпать**: scorrere il dito su uno schermo, fare swipe. Свайп + -а- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 34 volte.
17.	**Свайпить**: scorrere il dito su uno schermo, fare swipe. Свайп + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 15 volte. *ruTenTen17*: 25 volte.
18.	**Свайпнуть**: : scorrere il dito su uno schermo, fare swipe. Свайп + -ну- + -ть	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: 5 volte. *ruTenTen17*: 87 volte.
19.	**Селфить**: Fare un selfie. Селф (da *selfie*) + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 5 volte. *ruTenTen17*: 17 volte.
20.	**Селфиться**: Farsi un selfie. Селф (da *selfie*) + -и- + -ть- + -ся	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 61 volte.
21.	**Сорриться**: Scusarsi. Сорр (da *sorry*) + -и- + -ть- + -ся	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 35 volte. *ruTenTen17*: 24 volte.
22.	**Спойлерить**: Spoilerare; anticipare una parte di un film o di un libro, solitamente il finale. Спойлер + -и- + -ть	*Osnovnoj korpus*: 2 volte. *Ustnyj korpus*: No. *ruTenTen11*: 859 volte. *ruTenTen17*: 1.356 volte.
23.	**Спойлернуть**: Spoilerare; anticipare una parte di un film o di un libro, solitamente il finale. Спойлер + -ну- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 10 volte. *ruTenTen17*: 51 volte.
24.	**Стримить**: Guardare un video o un film direttamente da Internet, in streaming. Стрим + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 435 volte. *ruTenTen17*: 705 volte.
25.	**Триповать**: Fare un viaggio mentale, essere in trip.	*Osnovnoj korpus*: 1 volta.

	Трип + -ова- + -ть	*Ustnyj korpus*: No. *ruTenTen11*: 38 volte. *ruTenTen17*: 7 volte.
26.	**Троллинговать**: Trollare; provocare, disturbare e fomentare nel Web. Троллинг + -ова- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 74 volte. *ruTenTen17*: 52 volte.
27.	**Троллировать**: Trollare; provocare, disturbare e fomentare nel Web. Тролл + -ирова- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 41 volte. *ruTenTen17*: 14 volte.
28.	**Троллить**: Trollare; provocare, disturbare e fomentare nel Web. Тролл + -и- + -ть	*Osnovnoj korpus*: 6 volte. *Ustnyj korpus*: No. *ruTenTen11*: 4.467 volte. *ruTenTen17*: 4.386 volte.
29.	**Фейспалмить**: coprirsi il viso con la mano. Фейспалм + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 16 volte. *ruTenTen17*: 92 volte.
30.	**Флексить**: Mettersi in mostra, vantarsi; ballare. Флекс (da *to flex*) + -и- + -ть	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 14 volte.
31.	**Флудить**: Inviar a grande velocità una serie di messaggi, solitamente non inerenti all'argomento trattato. Deriva di *flood*, che significa "inondazione". Флуд + -и- + -ть	*Osnovnoj korpus*: 25 volte. *Ustnyj korpus*: No. *ruTenTen11*: 8.577 volte. *ruTenTen17*: 5.191 volte.
32.	**Фолловить**: Seguire qualcuno, specialmente sui social. Фоллов + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 1.219 volte. *ruTenTen17*: 464 volte.
33.	**Френдзонить**: Friendzonare. Френдзон(а) + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: No. *ruTenTen17*: 35 volte.
34.	**Хейтить**: Odiare. Хейт + -и- + -ть	*Osnovnoj korpus*: 1 volta. *Ustnyj korpus*: No. *ruTenTen11*: 33 volte. *ruTenTen17*: 360 volte.
35.	**Хейтерить**: Fare l'hater nei social network. Хейтер + -и- + -ть	*Osnovnoj korpus*: 30 volte. *Ustnyj korpus*: No. *ruTenTen11*: 30 volte. *ruTenTen17*: 54 volte.
36.	**Чекать**: verificare, controllare, fare un check. Чек + -а- + -ть	*Osnovnoj korpus*: 10 volte. *Ustnyj korpus*: 1 volta. *ruTenTen11*: 1.086 volte. *ruTenTen17*: 296 volte.

37.	**Чилить**: Riposare, dormire Чил + -и- + -ть	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 1.220 volte. *ruTenTen17*: 683 volte.
38.	**Чилиться**: Rilassarsi, riposarsi Чил + -и- + -ть- + -ся	*Osnovnoj korpus*: No. *Ustnyj korpus*: No. *ruTenTen11*: 7 volte. *ruTenTen17*: 107 volte.
39.	**Юзать**: Utilizzare, usare. Юз + -а- + -ть	*Osnovnoj korpus*: 28 volte. *Ustnyj korpus*: No. *ruTenTen11*: 32.451 volte. *ruTenTen17*: 14.542 volte.
40.	**Юзаться**: Usare, si usa/si usano. Юз + -а- + -ть- + -ся	*Osnovnoj korpus*: 2 volte. *Ustnyj korpus*: No. *ruTenTen11*: 1.020 volte. *ruTenTen17*: 597 volte.

4.8 Gli acronimi, le abbreviazioni, le interiezioni russe e le espressioni russificate

4.8.1 Gli acronimi: traslitterazione e russificazione del modello alloglotto

ИМХО (IMHO, In My Humble Opinion)	
NKRJa Osnovnoj korpus	**ИМХО**: 726 volte.
NKRJa Ustnyj korpus	**ИМХО**: No.
ruTenTen11	**ИМХО**: 204.954 volte.
ruTenTen17	**ИМХО**: 130.881 volte.

ПМСМ (*PMSM, Po Moemu Skromnomu Mneniju*)	
NKRJa Osnovnoj korpus	**ПМСМ**: 3 volte.
NKRJa Ustnyj korpus	**ПМСМ**: No.
ruTenTen11	**ПМСМ**: 1.371 volte.
ruTenTen17	**ПМСМ**: 756 volte.

ЛОЛ (LOL, Lot Of Laughs/Laughing Out Loud)	
NKRJa Osnovnoj korpus	**ЛОЛ**: 21 volte.
NKRJa Ustnyj korpus	**ЛОЛ**: 3 volte.
ruTenTen11	**ЛОЛ**: 17.198 volte.
ruTenTen17	**ЛОЛ**: 15.207 volte.

РОФЛ (*ROFL, Rolling On the Floor Laughing*)	
NKRJa Osnovnoj korpus	**РОФЛ**: 2 volte.
NKRJa Ustnyj korpus	**РОФЛ**: No.
ruTenTen11	**РОФЛ**: 73 volte.

ruTenTen17	**РОФЛ**: 60 volte.

СЗОТ (*SZOT, Sorri za offtopik* < *Sorry for off-topic*)	
NKRJa Osnovnoj korpus	**СЗОТ**: No.
NKRJa Ustnyj korpus	**СЗОТ**: No.
ruTenTen11	**СЗОТ**: 466 volte.
ruTenTen17	**СЗОТ**: 123 volte.

4.8.2 Le abbreviazioni di parole russe o di prestiti stranieri

Варик	
NKRJa Osnovnoj korpus	**Варик**: 1 volta.
NKRJa Ustnyj korpus	**Варик**: No.
ruTenTen11	**Варик**: 2.435 volte.
ruTenTen17	**Варик**: 2.402 volte.

Вписка	
NKRJa Osnovnoj korpus	**Вписка**: 24 volte.
NKRJa Ustnyj korpus	**Вписка**: 6 volte.
ruTenTen11	**Вписка**: 2.868 volte.
ruTenTen17	**Вписка**: 2.941 volte.

Днюха	
NKRJa Osnovnoj korpus	**Днюха**: 4 volte.
NKRJa Ustnyj korpus	**Днюха**: 32 volte.
ruTenTen11	**Днюха**: 4.749 volte.
ruTenTen17	**Днюха**: 3.873 volte.

Жиза, жиз	
NKRJa Osnovnoj korpus	**Жиза**: 3 volte.
NKRJa Ustnyj korpus	**Жиза**: No.
ruTenTen11	**Жиза**: 8 volte.
ruTenTen17	**Жиза**: 97 volte.

Кста	
NKRJa Osnovnoj korpus	**Кста**: 13 volte.
NKRJa Ustnyj korpus	**Кста**: 3 volte.
ruTenTen11	**Кста**: 10.012 volte.
ruTenTen17	**Кста**: 4.793 volte.

Музон	
NKRJa Osnovnoj korpus	**Музон**: 31 volte.
NKRJa Ustnyj korpus	**Музон**: 11 volte.
ruTenTen11	**Музон**: 4.283 volte.
ruTenTen17	**Музон**: 2.772 volte.

7Я	
NKRJa Osnovnoj korpus	**7Я**: No.
NKRJa Ustnyj korpus	**7Я**: No.
ruTenTen11	**7Я**: No.
ruTenTen17	**7Я**: 1.718 volte.

Кэп	
NKRJa Osnovnoj korpus	**Кэп**: 80 volte.
NKRJa Ustnyj korpus	**Кэп**: 1 volta.
ruTenTen11	**Кэп**: 13.446 volte.
ruTenTen17	**Кэп**: 7.760 volte.

4.8.3 Le interiezioni russe e le espressioni russificate

Ауф	
NKRJa Osnovnoj korpus	**Ауф**: 1 volta.
NKRJa Ustnyj korpus	**Ауф**: No.
ruTenTen11	**Ауф**: 744 volte.
ruTenTen17	**Ауф**: 405 volte.

По фану	
NKRJa Osnovnoj korpus	**По фану**: No.
NKRJa Ustnyj korpus	**По фану**: No.
ruTenTen11	**По фану**: 267 volte.
ruTenTen17	**По фану**: 249 volte.

Bibliografia

Achrenova N. A. (2013). "Teoretičeskie osnovy internet-lingvistiki", *Filologičeskie nauki. Voprosy teorii i praktiki*, 10:28 (2013), pp. 22-25.
Disponibile al link: https://www.gramota.net/articles/issn_1997-2911_2013_10_04.pdf.

Anderson L., Gavioli L., Zanettin F. (2018). *Translations and interpreting for language learners (TAIL). Lessons in honour of Guy Aston, Anna Ciliberti, Daniela Zorzi*, Milano, AItLA - Associazione Italiana di Linguistica Applicata, 2018, pp. 403-453.
Disponibile al link: http://www.aitla.it/pubblicazioni/studi-aitla/24-pubblicazioni/studi-aitla/608-studi-aitla-8.

Assunção C., Araújo C. (2019). "Entries on the History of Corpus Linguistics", *Linha D'Água*, 32:1 (2019), pp. 39-57.
Disponibile al link: https://www.revistas.usp.br/linhadagua/issue/view/11055.

Atkins S., Clear J., Osler N. (1991). 'Corpus Design Criteria' in *Literary and Linguistic Computing*, vol. 7:1 (1992), pp. 1-16.

Baker P., Hardie A., McEnery T. (2006). *A Glossary of Corpus Linguistics*, Edinburgo, Edinburgh University Press.

Baranov A. N. (2001). *Vvedenie v prikladnuju lingvistiku: Učebnoe posobie*, Editorial URSS.

Barbera M. (2009). *Schema e storia del "Corpus Taurinense*, *Linguistica dei corpora dell'italiano antico*, Alessandria, Edizioni dell'Orso.
Disponibile al link:
http://www.bmanuel.org/edocs/Barbera_CorpusTaurinense&ItalianoAntico=EB_2009.pdf.

Barbera M. (2013). *Linguistica dei corpora e linguistica dei corpora italiana*. *Un'introduzione*, Milano, Qu.A.S.A.R.
Disponibile al link:
http://www.bmanuel.org/man/Barbera_IntroduzioneCL_2013=Ver1-54.pdf.

Barbera M. (2015). *Quanto più la relazione è bella: Saggi di storia della lingua italiana 1999-2014*, Torino, bmanuel.org.

Barbera M., Corino E., Onesti C. (2007). *Corpora e Linguistica in rete*, Perugia, Guerra Edizioni.

Baroni M., Bernardini S. (2004). 'BootCat: Bootstrapping corpora and terms from the web', in *Proceedings of LREC 2004*, Lisbona, ELDA, pp. 1313-1316.

Baroni M., Bernardini S. (2006). *WaCky! Working Papers on the Web as Corpus*, Bologna, Gedit Edizioni.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009). 'The WaCky wide web: a collection of very large linguistically processed web-crawled corpora', in *Language Resources & Evaluation*, vol. 43:3 (2009), pp. 209-226.

Baroni, M., Kilgarriff, A. (2006). 'Large linguistically-processed web corpora for multiple languages' in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 3-7Aprile 2006, Trento, Italia, pp. 87-90.

Baronova M. M. (2013). *Vse pravila russkij orfografii i punktuacii c priloženiem universal'nogo orfografičeskogo slovarja i kratkogo kursa grammatiki*, Mosca, Astrel'.

Benko V. (2014). 'Aranea: Yet Another Family of (Comparable) Web Corpora' in *Text, Speech and Dialogue*, a cura di Sojka P., Horák A., Kopeček I., Pala K., Brno, Repubblica Ceca, Springer International Publishing, pp. 257–264.
Disponibile al link: https://www.tsdconference.org/tsd2014/download/preprints/672.pdf.

Benko V., Zacharov V. P. (2016). 'Very large Russian corpora: new opportunities and new challenges', in *Kompjuternaja lingvistika i intellektualnye technologii: po materialam meždunarodnoj konferencii "Dialog"*, vol. 15:22 (2016), pp. 79-93.

Bergh G. (2005), "Min(d)ing English Language Data on the Web. What Can Google Tell us?", *ICAME Journal*, vol. 29, pp. 25-46.

Biber D. (1993). "Representativeness in Corpus Design" in *Literary and Linguistic Computing*, vol. 8:4 (1993), pp. 243–257.
Disponibile al link: http://otipl.philol.msu.ru/media/biber930.pdf.

Biber D., Conrad S., Leech G. (2002). *Student Grammar of Spoken and Written English*, Essex (Inghilterra), Pearson Education Limited.

Bowker L., Pearson J. (2002). *Working with Specialized Languages: A practical guide to using corpora*, London and New York, Routledge.

Cavaglià G., Kilgarriff A. (2000). "Corpora from the Web" in *Fourth Annual CLUCK Colloquium*, Gennaio 2001,Sheffield, UK.

Cevese C., Dobrovolskaja J., Magnanini E. (2000). *Grammatica russa. Morfologia: teorie ed esercizi*, Milano, Hoepli.

Chesi C. (2012). "Competenza e performance: una distinzione cognitivamente obsoleta", *Sistemi intelligenti*, 24:2 (2012), pp. 241-260.

Chiari I. (2007). *Introduzione alla linguistica computazionale*, Bari, Laterza, pp.40-83.

Chiari I. (2012). "Corpora e risorse linguistiche per l'italiano. Stato dell'arte, problemi e prospettive", *Italienisch*, vol. 68 (2012), pp. 90-105.

Chini M. (2016). “Elementi utili per una didattica dell’italiano L2 alla luce della ricerca acquisizionale”, Italiano LinguaDue, 8:2 (2016), pp. 1-18.
Disponibile al link: https://riviste.unimi.it/index.php/promoitals/article/view/8172.

Clear J. (1986). “Trawling the language: Monitor corpora”, in *ZüriLEX '86 Proceedings, Papers read at the EURALEX International Congress*, 9-14 Settembre 1986, Università di Zurigo, pp. 383-389.
Disponibile al link: https://euralex.org/category/publications/euralex-1986/.

Cohen K. B., Xia J., Zweigenbaum P., Callahan T. J., Hargraves O., Goss F., Ide N., Névéol A., Grouin C., Hunter L. E. (2018). “Three Dimensions of Reproducibility in Natural Language Processing”, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 7-12 Maggio 2018, Miyazaki, Giappone, pp. 156–165.

Corino E. (2014). “Didattica delle lingue corpus-based” in *EL.LE*, vol. 3:2 (2014), pp. 231-257.
Disponibile al link: https://edizionicafoscari.unive.it/media/pdf/article/elle/2014/2/art-10.14277-2280-6792-99p.pdf.

Crystal D. (2004). *Language and the Internet*, Cambridge, Cambridge University Press.

Di Maio A. (1989). “L’«informatica linguistica» di padre Roberto Busa come metodo investigativo e come approccio al Medioevo”, *Medioevo*, 15 (1989), p. 325-362.

Dračeva J. N., Zubova N. N. (2015). “Mul’timedijnyj korpus regional’nych tekstov «Žiznennyj krug»: k probleme architektoniki oboločki”, in *Vestnik Čerepoveckogo gosudarstvennogo universiteta*, Vol. 3: 64 (2015), pp. 72-75.

Dračeva J. N., Zubova N. N. (2015). “Mul’timedijnyj korpus dialektnych tekstov kak osnova izučenija jazykovoj ličnosti žitelja regiona”, in *Vestnik Čerepoveckogo gosudarstvennogo universiteta*, Vol. 5: 66 (2015), pp. 47-50.

Dračeva J. N., Zubova N. N. (2015). “Mul’timedijnyj korpus «Žiznennyj krug» kak osnova izučenija kognitivnoj sostavljajuščej dialektnoj jazykovoj ličnosti”, in *Vestnik Čerepoveckogo gosudarstvennogo universiteta*, Vol. 6: 67 (2015), pp. 58-61.

Endredy I., Novak A. (2013). “More Effective Boilerplate Removal—the GoldMiner Algorithm”, *Polibits*, vol. 48 (2013), pp. 79-83.
Disponibile al link:
https://www.polibits.cidetec.ipn.mx/ojs/index.php/polibits/article/view/48-10.

Ferraresi A. (2009). “Google and beyond: web as corpus methodologies for translators”, *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, vol. 7, pp. 1-8.
Disponibile al link: https://raco.cat/index.php/Tradumatica/article/view/154831/206725.

Fletcher W. (2004). “Making the web more useful as a source for linguistic corpora” in *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*, a cura di Connor U., Upton T., Amsterdam, Rodopi.

Freddi M. (2014). *Linguistica dei corpora*, Roma, Carocci editore.

Fries P. H. (2010). “Charles C. Fries, linguistics and corpus linguistics”, *ICAME Journal*, 34 (2010), pp. 89-119.

Galjamina J. E. (2014). “Lingvističeskij analiz cheštegov Tvittera” in *Sovremennyj russkij jazyk v internete* a cura di Achapkina J. E., Rachlina E. V., Jazyki slavjanskoj kul’tury, pp. 13-22.

Gandin S. (2005). Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli, in *AnnalSS V*, pp. 133-152.
Disponibile al link: https://core.ac.uk/download/pdf/11689686.pdf.

Gatijatullina G.M., Bereznikov D. (2017). “Osnovnye etapy stanovlenija korpusnoj lingvistiki” in *Informacionnye technologii v issledovatel’skom prostranstve raznostrukturnych jazykov*, 5 Dicembre 2016, Kazan’, pp. 18-20.

Gatto M. (2009). *From Body to Web. An Introduction to the web as corpus,* Roma-Bari, Laterza.

Gatto M. (2014). *Web as corpus. Theory and practice,* Londra, Bloomsbury.

Gorbov A. A. (2018). “Voprosy etimologii i rol’ enciklopedičeskich svedenij v leksikografičeskom opisanii zaimstvovanij v sovremennom russkom jazyke”, *Scando-Slavica*, 64:2 (2018), pp. 263-282.

Granger S., Dupont M., Meunier F., Naets H., Paquot M. (2020). *International Corpus of Learner English, Version 3,* Belgio, Presses universitaires de Louvain.

Gries S. T., Berez A. L. (2017). “Linguistic Annotation in/for Corpus Linguistics”, in *Handbook of Linguistic Annotation,* a cura di Ide N., Pustejovsky J., Springer Science e Business Media Dordrecht, pp.379-409.

Grišina E. A., Savčuk S. O. (2009). “Korpus ustnych tekstov v NKRJa: sostav i struktura”, in *Nacional’nyj korpus russkogo jazyka: 2006-2008. Novye rezul’taty I perspektivy*, Nestor-Istorija, pp. 129-149.
Disponibile al link: https://ruscorpora.ru/new/sbornik2008/07.pdf.

Grišina E. A. (2011). “Mul’timedijnyj russkij korpus: sovremennoe sostojanie i perspektivy razvitija”, in *Trudy meždunarodnoj konferencii “Korpusnaja Lingvistika – 2011”*, San Pietroburgo, 27-29 Giugno 2011, pp. 138-144.

Grišina E. A. (2015). “Mul’timodal’nyj modul’ v sostave Nazional’nogo Korpusa Russkogo Jazyka”, *Trudy Instituta russkogo jazyka im. V. V. Vinogradova*, vol. 6 (2015), pp. 65-87.

Disponibile al link: https://ruscorpora.ru/new/multimodule-module.pdf.

Gunraj D. N., Drumm-Hewitt A. M., Dashow E. M., Upadhyay S. N., Klin C. (2016). "Texting insincerely: The role of the period in text messaging", in *Computers in Human Behavior*, vol. 55 B, 2016, pp. 1067-1075.

Gusmani R. (1993). *Aspetti del prestito linguistico*, Napoli, Libreria Scientifica Editrice.

Heid U. (2007). "*Il corpus WorkBench come strumento per la linguistica dei corpora. Principi ed applicazioni*", in Corpora e Linguistica in rete, a cura di Barbera M., Corino E., Onesti C., Perugia, Guerra edizioni, 2007, pp. 89-108.

Henzinger M., Lawrence S. (2004). "Extracting Knowledge from the World Wide Web", in *Proceedings of the National Academy of Sciences*, vol. 101:1, pp. 5186-5191. Disponibile al link: https://www.pnas.org/content/pnas/101/suppl_1/5186.full.pdf.

Higdon D. L. (2003). "The Concordance: Mere Index or Needful Census?", *Text*, Vol. 15 (2003), Indiana University Press, pp. 51-68.

Hundt M., Nesselhauf N., Biewer C. (2007). *Corpus Linguistics and the Web*, Amsterdam - New York, Rodopi.

Iannaccaro G. (2013). *La linguistica italiana all'alba del terzo millennio (1997-2010)*, Roma, Bulzoni, 2013, "Pubblicazioni della Società di linguistica italiana [SLI]".

Irgizova K. V. (2019). "Korpusnaja lingvistika v otečestvennom i zarubežnom jazykoznanii na sovremennom etape", in *Огарёв-Online*, vol. 6:127 (2019). Disponibile al link: http://journal.mrsu.ru/arts/korpusnaya-lingvistika-v-otechestvennom-i-zarubezhnom-yazykoznanii-na-sovremennom-etape.

Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V. (2013). “The TenTen Corpus Family”, in *7th International Corpus Linguistics Conference CL 2013*, Lancaster, pp. 125-127.

Johannessen J. B., Guevara E. R. (2011). "What kind of corpus is a web corpus?", in *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA)*, Riga, Latvia, Northern European Association for Language Technology (NEALT), pp. 122-129.
Disponibile al link: https://aclanthology.org/W11-4617.

Johansson S. (2003). “Reflection on Corpora and their Uses in Cross-linguistic Research” in F. Zanettin, S. Bernardini and D. Stewart (eds) Corpora in Translator, St. Jerome, Manchster: 138-140.

Jussila J., Alkhammash E., Saleh Alghamdi N., Madhala P., Ayoub Khan M. (2022), “A Netnographic-Based Semantic Analysis of Tweet Contents for Stress Management”, *CMC-Computers, Materials & Continua*, Vol.70:1 (2022), pp. 1845-1856.
Disponibilie al link: https://www.techscience.com/cmc/v70n1.

Kačkovskaja T., Kočarov D., Skrelin P., Vol’skaja N. (2016). “CoRuSS - a New Prosodically Annotated Corpus of Russian Spontaneous Speech”, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, pp. 1949-1954.

Kačkovskaja T. *et alii* (2020). “SibLing Corpus of Russian Dialogue Speech Designed for Research on Speech Entrainment”, in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marsiglia, 11-16 Maggio 2020, pp. 6556–6561.
Disponibile al link: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.807.pdf.

Kamšilova O. N. (2017). “Overuse in learner language: frequency and accuracy”, *Russian Linguistic Bulletin*, 3: 11 (2017), pp. 28-31.
Disponibile al link:
https://rulb.org/wp-content/uploads/wpem/pdf_compilations/3(11)/28-31.pdf

Kilgarriff, A. (2007). “Googleology is bad science” in *Computational linguistics*, 33:1, pp. 147-151.

Kilgarriff A., Grefenstette G. (2003). “Introduction to the Special Issue on the Web as Corpus” in *Computational Linguistics*, 29:3, pp. 333-347.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. (2004). “The Sketch Engine”, in *Proceedings of Euralex*, a cura di Williams G., Vessier S., Lorient, Francia, Université de Bretagne-Sud, pp. 105-116.

Koller, V., Hardie, A., Rayson, P., Semino, E. (2008). “Using a semantic annotation tool for the analysis of metaphor in discourse”, *Metaphorik.de*, Vol. 15 (2008), pp. 141-160.
Disponibile al link: www.metaphorik.de/15/.

Kotov A. A., Gopkalo O. C. (2011). “Russkojazyčnyj emocional’nyj korpus: kommunikativnoe vzaimodejstvie v real’nych emocional’nych situacijach” in *Trudy meždunarodnoj konferencii «Korpusnaja lingvistika-2011»*, San Pietroburgo, pp. 211-216.

Kovalev V. (2014). *Dizionario Russo-Italiano, Italiano-Russo*, Bologna, Zanichelli, 2496 p.

Krjučkova O. J., Gol’din V. E. (2011). “Korpus russkoj dialektnoj reči: koncepcija i parametri ocenki”, in *Meždunarodnaja konferencija “Dialog-2011” «Komp’juternaja lingvistika i intellektual’nye technologii»*, Bekasovo, 25-29 Maggio 2011, pp. 359-367.
Disponibile al link: http://www.dialog-21.ru/dialogue2011/results/presentations/.

Krjučkova O. J., Gol'din V. E. (2015). "Parametry obrabotki tekstov dlja russkogo dialektnogo korpusa", in *Meždunarodnaja konferencija «Korpusnaja lingvistica -2015»*, San Pietroburgo, 22-26 Giugno 2015, pp. 307-3014.
Disponibile al link: https://events.spbu.ru/events/archive/corpora-2015/dokladyi.html.

Krjučkova O. J. (2021). "O vozmožnostjach korpusnogo izučenija kartiny mira nositelej tradicionnoj narodnoj kul'tury", in *Izvestija Saratovskogo universiteta. Novaja serija: Filologija. Žurnalistika. 2021*, Vol. 21: 2 (2021), pp. 126-130.

Kuzmenko E., Kutuzov A. (2014). "Russian Error-Annotated Learner English Corpus: a Tool for Computer-Assisted Language Learning", in *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014*, Linköping University Electronic Press, Uppsala University, pp. 87-97.

Lawrence S., Giles C. L. (1999). "Accessibility of information on the Web", *Nature*, 400 (1999), pp. 107-109.

Lew R. (2009). "The Web as corpus versus traditional corpora: their relative utility for linguists and language learners", in *Contemporary Corpus Linguistics*, a cura di P. Baker, London, Continuum, pp. 289-300.

Lyding V. *et alii* (2014). "The PAISÀ Corpus of ItalianWeb Texts", in *Proceedings of the 9th Web as Corpus Workshop (WaC-9), EACL 2014*, a cura di Bildhauer F. e Schäfer R., Gothenburg, Svezia, 26 Aprile 2014, pp. 36-43.

Lüdeling A., Evert S., Baroni M. (2007). "Assessing the web as corpus", in *Corpus Linguistics and the Web*, a cura di M. Hundt, N. Nesselhauf, C. Biewer, Amsterdam, Rodopi, pp. 7-24.

Lüdeling A., Kytö M. (2008), *Corpus Linguistics, An International Handbook Volume 1*, Berlino, New York, Walter de Gruyter.

Lukaševič N. J., Klyšinskij E. S., Kobozeva I. M. (2016). 'Lexical research in Russian: are modern corpora flexible enough?', in *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"*, Mosca, pp. 427-440.

Majorova A. D. (2017). "Korpusnaja lingvistika: istoričeskij i lingvodidaktičeskij aspekty", *Meždunarodnyj naučno-issledovatel'skij žurnal*, 5:59 (2017), pt. II, pp. 42-46.

Makarova V., Petrušin V. A. (2002). "RUSLANA: A database of Russian emotional utterances", in *7th International Conference on Spoken Language Processing*, Denver, Colorado, 16-20 Settembre 2002, pp. 2041-2044.

Malyševa E. G., Rogaleva O. S. (2012). *Sovremennyj russkij jazyk. Fonetika. Orfoepija*, Omsk, Omsk Izdatel'stvo.

Manca E. (2015). "Un approccio corpus-driven al linguaggio dell'immigrazione", *Lingue e Linguaggi*, 16 (2015), pp. 485-507.

Martynjuk O. A. (2013). "Korpusnaja lingvistika i novye vozmožnosti lingvističeskogo issledovanija", in *Naukovyj visnik mižnarodnogo gumanitarnogo universitetu*, vol. 7 (2013), pp. 27-33.
Disponibile al link: http://www.vestnik-philology.mgu.od.ua/index.php/arkhiv-nomeriv?id=42.

McEnery T., Hardie A. (2012). *Corpus Linguistics: Method, Theory and Practice*, Cambridge, Cambridge University Press, pp. 1-70.

McEnery T., Wilson A. (2001). *Corpus Linguistics: An Introduction*, Edinburgh, Edinburgh University Press, pp. 1-27.

Mitrenina O. (2014). “The corpora of old and middle Russian texts as an advanced tool for exploring an extinguished language”, *Scrinium* 10:1 (2014), pp. 455-461.

Moneglia M. (2019), “Lablita-Suite. Risorse Per L’acquisizione Dell’italiano L2”, *Kwartalnik Neofilologiczny*, LXVI, 2 (2019), pp. 407-421.
Disponibile al link: https://journals.pan.pl/kn/129431#tabs.

Moneglia M., Panunzi A. (2010). *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*, Firenze, Firenze University Press.

Nyhan J., Passarotti M. (2019). *One Origin of Digital Humanities. Fr Roberto Busa in His Own Words*, Springer, pp. 1-17.
Disponibile al link: https://doi.org/10.1007/978-3-030-18313-4.

O’Keeffe A., McCarthy M. (2010). *The Routledge Handbook of Corpus Linguistics*, Londra e New York, Routledge.

Ol’chovskaja A. I. (2019).”Korpusnoe prepodavanie pusskogo jazyka”, *Vestnik tomskogo gosudarstvennogo pedagogičeskogo universiteta*, 2:199 (2019), pp. 98-107.

Olohan M. (2004). *Introducing Corpora in Translation Studies*, Routledge, London & New York, pp. 24-25.

Panunzi A., Cresti E., Gregori L. (2014). *RIDIRE. Corpus and Tools for the Acquisition of Italian L2*, in Proceedings of the XVI EURALEX International Congress: The User in Focus, Bolzano, 15-19 Luglio 2014, a cura di Abel A., Vettori C., Ralli N., Bolzano, EURAC research.

Paracchini L. (2017). “La lingua di Internet in Russia: stato della ricerca” in *L'analisi linguistica e letteraria*, vol. 25:1, pp. 45-98.
Disponibile al link:
https://www.analisilinguisticaeletteraria.eu/index.php/ojs/article/view/181/143.

Paracchini L. (2019). “I meccanismi di suffissazione relativi alla formazione dei verbi nella lingua russa di Internet”, in *Studi di linguistica slava. Nuove prospettive e metodologie di ricerca*, a cura di Iliyana Krapova, Svetlana Nistratova, Luisa Ruvoletto, Venezia, Edizioni Ca' Foscari, 2019, pp. 389-409.
Link: https://edizionicafoscari.unive.it/it/edizioni4/libri/978-88-6969-369-4/i-meccanismi-di-suffissazione-relativi-alla-formaz/.

Piperski A. Č. (2013). “General’nyj internet-korpus russkogo jazyka i ponjatie reprezentativnosti v korpusnoj lingvistike”, *Sovremennye problemy nauki i obrazovanija*, vol. 5 (2013).
Disponibile al link: https://s.science-education.ru/pdf/2013/5/14.pdf.

Piperski A. Č. (2020). *Russkij jazyk i korpusnoe raznoobrazie*, in Komp’juternaja lingvistika i intellektyal’nye technologii: po materialam meždunarodnoj konferencii «Dialog 2020», Mosca, 17-20 giugno 2020.
Disponibile al link: http://www.dialog-21.ru/media/5114/piperskiach-087.pdf.

Prada M. (2010). “LIPSI. Il lessico di frequenza dell’italiano parlato in Svizzera”, *Italiano LinguaDue*, vol. 2:1 (2010), pp. 182-205.

Pearsall R. (1971). “Cruden of the Concordance 1701-1770”, *New Blackfriars*, Vol. 52:609 (1971), pp. 88-90.

Piccolo G., Di Maio A. (2014). “Roberto Busa: tra «cervello meccanico» e «cervello spirituale»”, *La Civiltà Cattolica*, 3(2014), pp. 67-78.

Pomikalek J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*, PhD thesis, Masaryk University, Brno.

Prat Zagrenelsky M. T. (2007). “L'introduzione della corpus linguistics o linguistica dei corpora, nelle università italiane: una ricostruzione personale dagli anni '60 a oggi”, in

Lessicologia e lessicografia nella storia degli insegnamenti linguistici, atti delle giornate di Bologna, quaderni del CIRSIL-4 (2005).
Disponibile al link: http://amsacta.unibo.it/2716/1/Prat_sito.pdf.

Rockwell, G., Passarotti, M. (2019). "The Index Thomisticus as a Big Data Project". *Umanistica Digitale* 5(2019), pp. 13-34.

Rossini Favretti R. (2001). 'La linguistica dei "corpora" in Europa: prospettive di analisi' in *Lingua e Stile, Rivista di storia della lingua italiana*, vol.2/2001, pp. 367-382.

Rozental' D. E. (2010). *Russkij Jazyk. Učebnoe posobie dlja škol'nikov staršich klassov i postupajušich v vuzy*, Mosca, Oniks – Mir i obrazovanie.

Rundell M. (2008). "The corpus revolution revisited", *English Today*, Vol. 24: 1 (2008), pp. 23-27.
Disponibile al link: https://doi.org/10.1017/S0266078408000060

Sabatini F. (2006). "La storia dell'italiano nella prospettiva della corpus linguistics" in *EURALEX* 2006, pp. 31-37.
Disponibile al link: https://euralex.org/category/publications/euralex-2006/.

Šaroff S. (2006). "Methods and tools for development of the Russian Reference Corpus", in *Corpus linguistics around the world*, vol. 56 (2006), pp. 167-180.

Saženin I. I. (2013). "Slovarnyj korpus kak element optimizacii issledovatel'skogo processa", *Vestnik Novosibirskogo gosudarstvennogo pedagogičeskogo universiteta*, vol. 2:12 (2013), pp. 120-127.

Signorini S., Tucci I. (2004). *Il restauro e l'archiviazione elettronica del primo corpus di italiano parlato, il corpus Stammerjohann*. In *Costituzione, gestione e restauro di corpora vocali*. Atti delle XIV Giornate del GFS Viterbo, 4-6 dicembre 2003, a cura di

De Dominicis A., Mori L. e Stefani M., Collana degli Atti dell'Associazione Italiana di Acustica, vol. XXXI, pp. 119-124.

Solnyškina M. I., Gatijatullina G. M. (2020). Istorija razvitija korpusnoj lingvistiki (na primere anglojazyčnych korpusov), Vestnik Tomskogo gosudarstvennogo universiteta. Filologija, vol. 63 (2020), pp. 132-160.

Starko V. (2020). *Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools*, in Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), Lviv, Ucraina, 23-24 Aprile 2020, Vol. I, pp. 239-248.
Disponibile al link: http://ceur-ws.org/Vol-2604/.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*, Language Science Press, Berlin, pp. 21-59.
Disponibile al link: https://langsci-press.org/catalog/book/148.

Suchomel V., Pomikálek J. (2012). "Efficient Web Crawling for Large Text Corpora" in *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, Lione, 2012 a cura di Adam Kilgarriff e Serge Sharoff, pp. 39-43.

Tavosanis M. (2019). "Variazione linguistica nei commenti su Facebook", in *Italiano LinguaDue*, 11:1, pp. 112-125.
Disponibile al link:
https://riviste.unimi.it/index.php/promoitals/article/view/12205/11354.

Tognini-Bonelli E. (2001). *Corpus Linguistics at Work*, Amsterdam, Philadelphia, John Benjamins Publishing Company, pp. 65-100.

Volk, M. (2002). 'Using the web as corpus for linguistic research', in *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, a cura di Pajusalu R., Hennoste T., Tartu, University of Tartu, pp. 355-369.

Wynne, M. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*, a cura di Wynne M., Oxford, Oxbow Books.
Disponibile al link: http://ota.ox.ac.uk/documents/creating/dlc/.

Zacharov V. P. (2013). "Corpora of the Russian Language", in *Text, Speech, and Dialogue: 16th International Conference, TSD 2013,* Pilsen, Repubblica Ceca, Settembre 2013, a cura di Habernal I. e Matoušek V.,Springer, 2013, pp. 1-13.

Zacharov V. P. (2014). "Korpusnaja lingvistika v Rossii", in *CrossLingua'2014, III Meždisciplinarnyj naučnyj forum*, Sinferopoli, Crimea, 27-30 Maggio 2014.
Disponibile al link: http://crosslingua.cfuv.ru/publications/2014_2_zakharov.pdf.

Zacharov V. P., Bogdanova S. J. (2013). *Korpusnaja lingvistika yčebnoe posobie*, Sankt-Peterburgskij gosudarstvennyj universitet, San Pietroburgo, 2013.

Zong Z., Hong C. (2019). "Research on Alignment in the Construction of Parallel Corpus", in *Journal of Physics: Conference Series*, vol. 1213 (2019).
Disponibile al link: https://iopscience.iop.org/article/10.1088/1742-6596/1213/4/042003.

Sitografia

Descrizione (§ capitoli e paragrafi di riferimento): URL del Sito Web	**Ultima consultazione (dd/mm/yyyy)**
"About the BNC" nel sito del *British National Corpus* (§ 2.2.4): http://www.natcorp.ox.ac.uk/.	18/10/2021
Annotazione ortoepica in Treccani (§ 2.3.5): https://www.treccani.it/vocabolario/ortoepico/.	1/11/2021
Articolo de *Il Post* sulla pronuncia delle parole in inglese (§ 4.4): https://www.ilpost.it/2021/09/09/inglese-pronuncia-trascrizione/.	6/12/2021
Articolo sulla pronuncia delle parole in inglese (§ 4.4): https://officinamagazine.it/perche-linglese-non-si-pronuncia-come-si-scrive/.	6/12/2021
Asia Pacific Corpus Linguistics Association (APCLA) (§ 1.4): https://apcla.net/.	31/9/2021
Asociación Española de Lingüística de Corpus (AELINCO) (§ 1.4): http://www.aelinco.es/es.	31/9/2021
Associazione Italiana di Linguistica Computazionale (AILC) (§ 1.4): https://www.ai-lc.it/.	1/10/2021
Častotnyj slovar' russkogo jazyka (§ 1.5): http://project.phil.spbu.ru/lib/data/slovari/zasorina/zasorina.html.	1/10/2021
Chel'sinskij annotirovannyj korpus (§ 1.5): http://h248.it.helsinki.fi/hanco/index.html.	7/10/2021
Choždenie Bogorodicy po mukam, testo integrale (§ 2.3.3): http://drevne-rus-lit.niv.ru/drevne-rus-lit/text/hozhdenie-bogorodicy-po-mukam/hozhdenie-bogorodicy-po-mukam-original.htm.	30/10/2021
Corpus of Russian Student Texts (*CoRST*) (§ 2.3.3): http://web-corpora.net/learner_corpus.	29/10/2021

Definizione di campione statistico (§ 2.2.5): https://www.treccani.it/enciclopedia/campione-statistico_%28Dizionario-di-Economia-e-Finanza%29/. 19/10/2021

Definizione di *device* in Garzanti Linguistica (§ 3.2.4): https://www.garzantilinguistica.it/ricerca/?q=device. 9/11/2021

Definizione di *fonema* nel dizionario *Treccani* (§ 4.4.1): https://www.treccani.it/vocabolario/fonema/. 8/12/2021

Definizione di *fono* nel dizionario *Treccani* (§ 4.4.1): https://www.treccani.it/vocabolario/fono_res-03c72dc1-001d-11de-9d89-0016357eee51/ 8/12/2021

Diachroničeskie korpusa russkogo jazyka e Staroslavjanskie korpusa (§ 2.3.4): https://ruscorpora.ru/new/corpora-other.html. 31/10/2021

Dipartimento di Lessicografia Sperimentale dell'Istituto di Lingua Russa dell'Accademia Russa delle Scienze (§ 2.3.4): https://www.ruslang.ru/node/251. 31/10/2021

EURALEX (European Association for Lexicography) (§ 1.4): https://euralex.org/. 31/9/2021

I corpora *Aranea* (§ 3.3): http://ucts.uniba.sk/aranea_about/index.html. 19/11/2021

I due significati della parola russa varik in Acadmic.ru (§ 4.8.2): https://dic.academic.ru/dic.nsf/lastnames/1908, https://dic.academic.ru/dic.nsf/ushakov/765570. 1/01/2022

I significati della parola *Kep* in russo secondo *Academic.ru* (§ 4.8.2): https://sokrasheniya.academic.ru/807, https://dic.academic.ru/dic.nsf/business/7101, https://technical_translator_dictionary.academic.ru/105579. 2/01/2022

I sotto-corpora dell'*NKRJa* (§ 2.3.2): https://ruscorpora.ru/new/corpora-intro.html. 28/10/2021

I *TenTen corpora* su *Sketch Engine* (§ 3.3): https://www.sketchengine.eu/documentation/tenten-corpora/. 17/11/2021

I testi contenuti nella sezione *Drevnerusskij korpus* (*NKRJa*) (§ 2.3.3): https://ruscorpora.ru/new/search-old_rus.html. 30/10/2021

I *WaCky* corpora (§ 3.3): https://wacky.sslmit.unibo.it/doku.php?id=start. 19/11/2021

Il *Brexit Corpus* (§ 2.2.1): https://www.sketchengine.eu/brexit-corpus/#toggle-id-1. 13/10/2021

Il *British National Corpus* (§ 2.2.4): http://www.natcorp.ox.ac.uk/corpus/index.xml. 15/10/2021

Il contributo online di D. Crystal "The scope of Internet linguistics" (§ 3.1): https://www.davidcrystal.com/GBR/Books-and-Articles?itemId=807. 6/11/2021

Il corpus *CiT* (§ 1.4): http://www.culturitalia.info/ARCHIVIO/s_spina/cit/demo.htm. 27/09/2021

Il corpus *CoLFIS* (§ 1.4): http://www.ge.ilc.cnr.it/strumenti.php 27/09/2021

Il corpus *CORIS/CODIS* (§ 1.4): http://corpora.dslo.unibo.it/coris_ita.html. 27/09/2021

Il *Corpus La Repubblica* (§ 1.4): https://corpora.dipintra.it/. 27/09/2021

Il corpus manoscritto *Barberiniano latino 3953* originale (§ 1.4): https://spotlight.vatlib.it/dante/feature/il-canzoniere-di-nicolo-de-rossi. 21/9/2021

Il corpus *OPUS2* nella piattaforma *Sketch Engine* (§ 2.3.6): https://www.sketchengine.eu/opus-parallel-corpora/#toggle-id-1. 3/11/2021

Il *Corpus Taurinense* (§ 2.2.5): http://www.bmanuel.org/projects/ct-HOME.html. 23/10/2021

I criteri di progettazione utilizzati nella creazione del *BNC* (§ 2.2.4): http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html. 14/10/2021

Il *Curden's Concordance*, una ristampa originale (§ 1.3): 10/09/2021

https://www.unz.com/print/CrudenAlexander-1858.

Il *Domostroj*, opera integrale (§ 2.3.3): 30/10/2021
https://azbyka.ru/otechnik/Silvestr/domostroj/.

Il *Longman Learners' Corpus* (§ 2.3.7): 3/11/2021
http://global.longmandictionaries.com/longman/corpus#aa.

Il *Komp'juternyj korpus tekstov russkich gazet konca XX veka* (§ 1.5): 5/10/2021
http://www.philol.msu.ru/~lex/corpus/.

Il *Korpus russkogo žestovogo jazyka* (§ 2.3.5): 31/10/2021
http://rsl.nstu.ru/site/project.

Il manuale del *Brown Corpus* (§ 1.3): 15/9/2021
http://icame.uib.no/brown/bcm.html#tc.

Il manuale del CHAT (§ 1.4): 29/9/2021
https://talkbank.org/manuals/CHAT.pdf.

Il manuale del *Kolhapur Corpus* (§ 2.2.1): 13/10/2021
http://korpus.uib.no/icame/manuals/KOLHAPUR/INDEX.HTM.

Il manuale del LOB Corpus (§ 1.3): 18/9/2021
http://korpus.uib.no/icame/manuals/LOB/INDEX.HTM.

Il manuale del *SUSANNE Corpus* (§ 2.2.5): 21/10/2021
https://www.grsampson.net/SueDoc.html.

Il manuale del *Wellington Corpus* (§ 2.2.1): 12/10/2021
http://korpus.uib.no/icame/wsc/INDEX.HTM.

Il manuale dell'*Australian Corpus* (§ 2.2.1): 14/10/2021
http://korpus.uib.no/icame/manuals/ACE/INDEX.HTM.

Il *Manuscript Corpus* (§ 2.3.6): 2/11/2021
http://mns.udsu.ru/.

Il motore di ricerca russo *Yandex* (§ 4.3): 6/12/2021
https://yandex.com/.

Il *phrasal verb chill out* nel *Collins Dictionary* (§ 4.4.3): 11/12/2021

https://www.collinsdictionary.com/it/dizionario/inglese/chill-out.

Il *phrasal verb chill out* nell'*Oxford Learner's Dictionaries* (§ 4.4.3): https://www.oxfordlearnersdictionaries.com/definition/english/chill-out_2. 11/12/2021

Il progetto *C-ORAL-ROM* (§ 1.4): http://www.elda.org/en/proj/coralrom.html. 30/9/2021

Il progetto *CorDIC LABLITA* (§ 1.4): http://corporadidattici.lablita.it/. 30/9/2021

Il programma di de-duplicazione *Onion* (*ONe Instance ONly*) (§ 3.3): http://corpus.tools/wiki/Onion. 19/11/2021

Il programma di tokenizzazione *Unitok* (§ 3.3): http://corpus.tools/wiki/Unitok. 20/11/2021

Il programma di web crawling *Heritrix* (§ 3.3): https://webarchive.jira.com/wiki/spaces/Heritrix/overview. 19/11/2021

Il programma di web crawling *SpiderLing* (§ 3.3): http://corpus.tools/wiki/SpiderLing. 15/11/2021

Il programma per l'annotazione morfosintattica *TreeTagger* (§ 3.3): https://www.cis.lmu.de/~schmid/tools/TreeTagger/. 20/11/2021

Il *Russian Error-Annotated Learner English Corpus* (§ 2.3.1): https://realec.org/. 25/10/2021

Il *Russkij Učebnyj Korpus* (§ 2.3.7): http://web-corpora.net/RLC. 3/11/2021

Il significato della parola russa *muzon* in *Academic.ru* (§ 4.8.2): https://dic.academic.ru/dic.nsf/ruwiki/1834562. 1/01/2022

Il significato della parola russa *vpiska* in *Academic.ru* (§ 4.8.2): https://argo.academic.ru/845. 1/01/2022

Il sito del *Collins COBUILD Corpus* (§ 1.3): https://collins.co.uk/pages/elt-cobuild-reference-the-history-of-cobuild. 20/09/2021

Il sito del *GIKRJa* (§ 2.3.2): 28/10/2021
http://www.webcorpora.ru/.

Il sito dell'*International Business Machines Corporation* (§ 1.4): 24/09/2021
https://www.ibm.com/ibm/history/history/history_intro.html.

Il sito dell'Università statale di Mosca (MGU) (§ 1.5): 6/10/2021
http://www.philol.msu.ru/~lex/corpus/corp_descr.html.

Il sito della *Higher School of Economics (HSE)* (§ 2.3.2): 26/10/2021
https://www.hse.ru/en/.

Il sito di KWiCFinder (§ 3.3): 13/11/2021
https://www.kwicfinder.com/KWiCFinder.html.

Il sito ufficiale del *BNC* (§ 1.5): 8/10/2021
https://www.english-corpora.org/bnc/.

Il sito ufficiale del *COCA* (§ 1.5): 10/10/2021
https://www.english-corpora.org/coca/.

Il sito ufficiale del Commissariato di Pubblica Sicurezza (§ 3.2.1): 8/11/2021
https://www.commissariatodips.it/approfondimenti/social-network/approfondimenti-normativi/index.html.

Il sito ufficiale del *Corpus Thomisticus* (§ 1.4): 27/09/2021
http://www.corpusthomisticum.org/.

Il sito ufficiale del *Mašinnyj Fond Russkogo Jazyka* (§ 1.5): 2/10/2021
http://cfrl.ruslang.ru/.

Il sito ufficiale del sistema operativo *GNU* (§ 3.4.1): 21/11/2021
https://www.gnu.org/home.it.html.

Il sito ufficiale dell'*NKRJa* (§ 2.3.2): 28/10/2021
https://ruscorpora.ru/new/corpora-intro.html.

Il sito ufficiale dell'Università di Ratisbona (§ 2.3.3): 29/10/2021
https://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/netzwerke/regensburger-korpora/index.html.

Il sito ufficiale dell'Università di Saratov (§ 2.3.5): https://www.sgu.ru/structure/philological/narrech. 31/10/2021

Il sito ufficiale dello *SKAT* (§ 2.3.6): http://project.phil.spbu.ru/scat/page.php?page=project. 2/11/2021

Il sito ufficiale di *Academic.ru* (§ 4.1): https://academic.ru/. 3/12/2021

Il sito ufficiale di BootCat (§ 3.3): https://bootcat.dipintra.it/. 11/11/2021

Il sito ufficiale di Creative Commons (§ 3.4.1): https://creativecommons.org/. 21/11/2021

Il sito ufficiale di David Crystal (§ 3.1) https://www.davidcrystal.com/GBR/Books-and-Articles?itemId=807 9/11/2021

Il sito ufficiale di Erika Darics (§ 3.4.2): https://scholar.google.co.uk/citations?user=M1YG-2kAAAAJ&hl=en. 22/11/2021

Il sito ufficiale di Gretchen McCulloch (§ 3.4.2): https://gretchenmcculloch.com/. 22/11/2021

Il sito ufficiale di Noam Chomsky (§ 1.3): https://chomsky.info/. 13/9/2021

Il sito ufficiale di *Sketch Engine* (§ 1.5): https://www.sketchengine.eu/#blue. 11/10/2021

Il sito ufficiale di *Vikislovar'* (§ 4.1): https://ru.wiktionary.org/wiki. 3/12/2021

Il sito *World Wide Web Size* (§ 2.2.5): https://www.worldwidewebsize.com/. 20/10/2021

Il sostantivo *chill-out* nell'*Oxford Learner's Dictionaries* (§ 4.4.3): https://www.oxfordlearnersdictionaries.com/definition/english/chill-out_1. 11/12/2021

Il sostantivo *troll* nel *Cambridge Dictionary* (§ 4.7.2): https://dictionary.cambridge.org/it/dizionario/inglese/troll. 29/12/2021

Il sostantivo *trolling* nel *Cambridge Dictionary* (§ 4.7.2): https://dictionary.cambridge.org/it/dizionario/inglese/trolling. 29/12/2021

Il sotto-corpus *Mul'tipark* (*NKRJa*) (§ 2.3.6): https://ruscorpora.ru/new/search-multiparc_rus.html. 3/11/2021

Il *Timestamped JSI web corpus* su *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/jozef-stefan-institute-newsfeed-corpus/. 1/12/2021

Il *Timestamped Russian corpus* su *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/timestamped-russian-corpus/. 1/12/2021

Japan Association for English Corpus Studies (*JAECS*) (§ 1.4): https://www.jaecs2020.org/. 31/9/2021

Kirill Turovskij o Cirillo da Turov nel sito *Akademic* (§ 2.3.3): https://dic.academic.ru/dic.nsf/ruwiki/107271. 29/10/2021

Korean Association for Corpus Linguistics (*KACL*) (§ 1.4): http://kacl.or.kr/. 31/9/2021

L'aggettivo *chill-out* nell'*Oxford Learner's Dictionaries* (§ 4.4.3): https://www.oxfordlearnersdictionaries.com/definition/english/chill-out_3. 11/12/2021

L'etimologia della parola inglese *trash* (§ 4.3): https://www.etymonline.com/word/trash. 5/11/2021

L'*International Corpus of Learner English*, Versione 2 (§ 2.3.7): https://uclouvain.be/en/research-institutes/ilc/cecl/iclev2.html. 3/11/2021

L'*International Corpus of Learner English*, Versione 3 (§ 2.3.7): https://corpora.uclouvain.be/cecl/icle/trial/. 3/11/2021

L'*International Journal of Corpus Linguistics* (*IJCL*) (§ 1.4): https://benjamins.com/catalog/ijcl. 31/9/2021

L'*Učebnyj Mul'timodal'nyj Korpus* (§ 2.3.7): https://studbooks.net/2148397/literatura/uchebnyy_multimodalnyy_korpus. 3/11/2021

La definizione di *flash mob* nel *Collins Dictionary* (§ 4.4.3): https://www.collinsdictionary.com/it/dizionario/inglese/flash-mob. 14/12/2021

La definizione di *phrasal verb* nel *Cambridge Dictionary* (§ 4.4.3): https://dictionary.cambridge.org/it/dizionario/inglese/phrasal-verb. 12/12/2021

La funzione *N-grams* di *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/guide/n-grams-multiword-expressions/. 3/12/2021

La funzione *Parallel concordance* di *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/guide/parallel-concordance-searching-translations/. 3/12/2021

La funzione per ricercare neologismi e fare analisi diacroniche su Sketch Engine (§ 4.2.2): https://www.sketchengine.eu/guide/trends/. 3/12/2021

La funzione *Thesaurus* di *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/guide/thesaurus-synonyms-antonyms-similar-words/. 3/12/2021

La funzione *Wordlist* di *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/guide/wordlist-frequency-lists/. 3/12/2021

La funzione *Word Sketch* di *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/. 2/12/2021

La funzione *Word Sketch Differences* di *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/guide/word-sketch-difference-compare-words/. 3/12/2021

La parola *body art* nel *Collins Dictionary* (§ 4.4.3): https://www.collinsdictionary.com/it/dizionario/inglese/body-art. 9/12/2021

La parola *body art* nel *Merriam-Webster* (§ 4.4.3): https://www.merriam-webster.com/dictionary/body%20art. 9/12/2021

La parola *body positive* nel *Cambridge Dictionary* (§ 4.4.3): https://dictionary.cambridge.org/it/dizionario/inglese/body-positive. 9/12/2021

La parola *body shaming* nel *Cambridge Dictionary* (§ 4.4.3): 9/12/2021

https://dictionary.cambridge.org/it/dizionario/inglese/body-shaming.

La parola *body shaming* nel *Merriam-Webster* (§ 4.4.3): 9/12/2021
https://www.merriam-webster.com/dictionary/body-shaming.

La parola *cringe* nel sito dell'*Accademia della Crusca* (§ 4.4.7): 23/12/2021
https://accademiadellacrusca.it/it/parole-nuove/cringe/18487.

La parola *cyberbulling* nel *Collins Dictionary* (§ 4.4.3): 9/12/2021
https://www.collinsdictionary.com/it/dizionario/inglese/cyberbullying.

La parola *cyberbulling* nell'*Oxford Learner's Dictionaries* (§ 4.4.3): 9/12/2021
https://www.oxfordlearnersdictionaries.com/definition/english/cyberbullying?q=cyberbullying.

La parola *flash mob* nel *Cambridge Dictionary* (§ 4.4.3): 11/12/2021
https://dictionary.cambridge.org/it/dizionario/inglese/flashmob.

La parola *flash mob* nel *Collins Dictionary* (§ 4.4.3): 11/12/2021
https://www.collinsdictionary.com/it/dizionario/inglese/flash-mob.

La parola *flash mob* nel *Merriam-Webster* (§ 4.4.3): 11/12/2021
https://www.merriam-webster.com/dictionary/flash%20mob.

La parola *flash mob* nell'*Oxford Learner's Dictionaries* (§ 4.4.3): 11/12/2021
https://www.oxfordlearnersdictionaries.com/definition/english/flash-mob?q=flash+mob.

La parola *friend zone* nel *Cambridge Dictionary* (§ 4.4.3): 9/12/2021
https://dictionary.cambridge.org/it/dizionario/inglese/friendzone?q=friend+zone.

La parola *friend zone* nel *Collins Dictionary* (§ 4.4.3): 9/12/2021
https://www.collinsdictionary.com/it/dizionario/inglese/friend-zone.

La parola *off topic* nel *Collins Dictionary* (§ 4.4.3): 9/12/2021
https://www.collinsdictionary.com/it/dizionario/inglese/off-topic.

La parola *streaming* nel *Cambridge Dictionary* (§ 4.6.1): 27/12/2021
https://dictionary.cambridge.org/it/dizionario/inglese/streaming.

La parola *styling* nel *Collins Dictionary* (§ 4.6.1): 27/12/2021

https://www.collinsdictionary.com/it/dizionario/inglese/styling.

La pronuncia di *dislike* secondo il *Cambridge Dictionary* (§ 4.4.5): https://dictionary.cambridge.org/it/dizionario/inglese/dislike. 12/12/2021

La sezione *Glavnaja* dell'*NKRJa* (§ 2.3.4): https://ruscorpora.ru/new/. 31/10/2021

La sezione *Osnovnoj korpus tekstov* dell'*NKRJa* al link (§ 4.2.1): https://ruscorpora.ru/new/corpora-structure.html. 1/12/2021

La sezione *statistika korpusa* dell'*NKRJa* (§ 2.3.5 e § 4.2.1): https://ruscorpora.ru/new/corpora-stat.html. 3/12/2021

Le etimologie delle parole di origine inglese (§ 4.3): www.etimonline.com. 5/12/2021

Le funzioni di ricerca della parole chiave e di estrazione di terminologia specifica su *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/guide/keywords-and-term-extraction/. 3/12/2021

Lista delle lingue presenti in *Sketch Engine* (§ 4.2.2): https://www.sketchengine.eu/corpora-and-languages/. 3/12/2021

Lista di corpora dell'Accademia della Crusca (§ 1.4): http://old.accademiadellacrusca.org/it/link-utili/banche-dati-dellitaliano-scritto-parlato.html. 27/9/2021

Lista di corpora di M. Barbera (§ 1.4): http://www.bmanuel.org/clr/clr3_fi.html#Italian. 27/9/2021

Lista di parole contenenti -бью- (§ 4.4.2): https://wordhelp.ru/contains/. 10/01/2022

Per consultare l'*Australian Corpus* (§ 2.2.1): https://www.ausnc.org.au/corpora/ace. 12/10/2021

Povest' vremennych let, opera integrale (§ 2.3.3): https://azbyka.ru/otechnik/Nestor_Letopisets/povest-vremennyh-let/. 30/10/2021

Sito del Governo Italiano "Comunicazione Pubblica in Rete" (§ 1.1): http://qualitapa.gov.it/sitoarcheologico/www.urp.it/sito- 10/9/2021

storico/www.urp.it/Sezione.jsp-idSezione=1874.html.

Società di Linguistica Italiana (*SLI*) (§ 1.4): https://www.societadilinguisticaitaliana.net/. 31/09/2021

Swedish National Service Data (§ 1.5): https://snd.gu.se/en/catalogue/study/ext0071. 4/10/2021

The periphrastic future with shall and will in Modern English (§ 1.3): https://www.jstor.org/stable/457534?seq=43#metadata_info_tab_contents. 12/9/2021

Variazione diacronica, diafasica, diamesica, diatopica (§ 2.2.4): https://www.treccani.it/enciclopedia/variazione-linguistica_(Enciclopedia-dell'Italiano)/. 17/10/2021

VKontakte, ufficio stampa (§ 2.3.2): https://vk.com/press/q1-2021-results. 28/10/2021

Young's Analytical Concordance to the Bible (§ 1.3): https://archive.org/details/analyticalconcor00younuoft/page/n5/mode/2up?view=theater. 11/9/2021

Živoj Žurnal (§ 2.3.2): https://www.livejournal.com/. 28/10/2021

Žurnalnyj Zal (§ 2.3.2): https://magazines.gorky.media/. 29/10/2021