Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Magistrale in
Lingue Moderne per la Comunicazione e la Cooperazione Internazionale
Classe LM-38

Tesi di Laurea

Corpora tradizionali e web corpora della lingua russa: un approccio comparativo

Relatore
Prof. Rosanna Benacchio

Laureando
Giorgia Ponso
n° matr.1184941 / LMLCC

Anno Accademico 2018 / 2019

INDICE

RINGRAZIAMENTI

Ringrazio innanzitutto la Professoressa Benacchio per l'aiuto e l'attenzione che mi ha dedicato nell'elaborazione di questo lavoro.

Il ringraziamento più grande va poi ai miei genitori che in questi cinque anni di studi universitari mi hanno sempre sostenuta e hanno sopportato le porte chiuse, la televisione a basso volume e le conversazioni a voce bassa.

Un pensiero va infine agli amici e ai colleghi che sono stati presenti e mi hanno incoraggiata durante questo percorso e hanno vissuto con me difficoltà e soddisfazioni.

INTRODUZIONE

L'oggetto di interesse di questa ricerca sono i corpora linguistici, in particolare della lingua russa. Dopo un'introduzione generale su questi strumenti linguistici, verrà ampiamente illustrato il repertorio di corpora di cui la lingua russa dispone, che oggi viene spesso ancora ignorato. Oltre al *Nacional'nyj Korpus Russkovo Jazyka*, che è il corpus russo per eccellenza, esiste infatti una serie di altri corpora delle più svariate tipologie, sia tradizionali che del web. Il focus principale del lavoro sarà proprio il web, attorno al quale ruota, nell'ambito della *corpus linguistics*, il concetto di *web as corpus*. La rete può essere infatti uno strumento prezioso per l'analisi linguistica se si dispone della conoscenza dei metodi e degli strumenti per sfruttarlo al meglio. Il nostro scopo sarà proprio quello di rendere noti tali metodi e strumenti. Esistono poi diverse accezioni del concetto di web come corpus in base al ruolo che il web gioca nella ricerca linguistica. Vedremo quali teorie esistono a riguardo, senza tralasciare quelle che invece si oppongono a questa idea. Passeremo poi al confronto tra corpora tradizionali e corpora del web analizzandone i concetti cardine. Lo scontro decisivo avverrà però a livello pratico. L'obiettivo finale sarà infatti la messa in pratica, in ottica comparativa, delle due tipologie di corpora così da evidenziarne somiglianze e differenza, pregi e difetti. L'intento non sarà quello di stabilire quale dei due sia lo strumento migliore ma di mostrare quali siano i più adatti in base alla tipologia di ricerca linguistica da condurre.

CAPITOLO 1
CORPORA E LINGUISTICA DEI CORPORA

Introduzione

Essendo questo lavoro di ricerca principalmente incentrato sui corpora e sulla linguistica dei corpora, risulta doveroso iniziare dalle basi con uno sguardo generale su questa materia. Vedremo quindi in questo primo capitolo la definizione di linguistica dei corpora, facendo poi un excursus storico della disciplina, fino ad arrivare a vedere nel dettaglio le sue caratteristiche e quelle dei suoi strumenti principali, i corpora. Infine, analizzeremo i numerosi campi in cui essi possono essere, e sono, un contributo prezioso.

1.1 La linguistica dei corpora

Numerose sono le definizioni che una gran quantità di autori ha formulato per tentare di concretizzare il concetto di "linguistica dei corpora", o meglio conosciuto in lingua inglese come corpus linguistics. Appunto perché sono molte, riportarne soltanto una sarebbe riduttivo. Ho deciso quindi di citare alcuni esempi più rappresentativi di autori più o meno conosciuti che hanno scritto su questa materia.

La linguistica dei corpora è stata definita da McEnery e Hardy, entrambi docenti di inglese e linguistica inglese, come "a set of procedures, or methods, for studying language" basati su "some set of machine-readable texts" (McEnery, Hardie 2012:1). Un'altra definizione è quella di McEnery insieme a Wilson, anch'esso docente di linguistica inglese, che hanno descritto la linguistica dei corpora come "the study of language based on examples of 'real life' language use" (McEnery, Wilson 2001:1). Ancora, viene definita come "empirical approach to the study of language based on the observation of authentic data", dati che sono "authentic texts in machine-readable format" da Maristella Gatto (2014: 7, 9), docente e ricercatrice italiana di lingua inglese e traduzione.

La conclusione che è possibile trarre leggendo queste diverse definizioni è sostanzialmente che la linguistica dei corpora è un particolare metodo che viene utilizzato per studiare ed analizzare una lingua e i suoi usi. I dati autentici e reali di cui gli autori parlano nelle loro definizioni, descritti anche come testi in formato elettronico, non sono altro che i corpora, il principale strumento utilizzato dalla linguistica dei corpora per lo studio della lingua. Cosa siano nello specifico i corpora

e quali siano le loro principali specificità lo vedremo tra poco. Prima ritengo opportuno fare un piccolo approfondimento sul tipo di metodologia adottata dalla linguistica dei corpora. Per le sue caratteristiche, essa è oggigiorno considerata un vero e proprio metodo scientifico. Partendo da questo presupposto, considerando quindi la linguistica dei corpora una scienza di per sé stessa, si è spesso discusso sulla possibilità di definirla, come a volte è stato fatto, una branca della linguistica. Alla questione tentano di dare una risposta McEnery e Wilson (2001), i quali ci dicono che la *corpus linguistics* non è una branca della linguistica nel senso in cui lo sono sintassi, semantica, sociolinguistica ecc. Queste sono infatti incentrate sul descrivere e spiegare alcuni aspetti dell'uso della lingua. Al contrario, più che un aspetto linguistico che necessiti di essere spiegato, la linguistica dei corpora è una metodologia che può essere a sua volta applicata a molti, se non a tutti, gli aspetti della lingua al fine di descriverli essa stessa. Non delimita quindi di per sé una vera e propria area della linguistica.

1.2 La linguistica dei corpora come metodo scientifico

Come abbiamo detto, la linguistica dei corpora come approccio di studio della lingua è stata spesso considerata un metodo scientifico[1], facendo sì che oggi sia una scienza vera e propria. Un contributo notevole a questo approccio alla materia è stato dato dal linguista inglese Geoffry Leech, il quale attestò che i corpora permettono di studiare la lingua con un vero e proprio metodo scientifico (Leech, 1992). Tale metodo si basa sull'applicazione alla linguistica dei corpora di tre criteri generalmente utilizzati dalla scienza per lo studio della realtà. Questi tre criteri sono attendibilità (*accountability*), falsificabilità (*falsifiability*) e replicabilità (*replicability*) (McEnery, Hardie 2012: 14).

Il metodo scientifico, e di conseguenza i tre criteri appena citati, vengono applicati, nell'ambito linguistico, nel momento in cui presupponiamo una teoria linguistica e ne vogliamo testare l'esattezza, come avviene nella maggior parte dei casi in cui ci si approccia all'utilizzo di un corpus. Il primo di questi criteri, ovvero una totale attendibilità dei dati raccolti, è il criterio per cui una teoria linguistica dev'essere testata basandosi sull'intera serie di dati a nostra disposizione, che sono attendibili nella loro totalità. I dati che dovremmo prendere in considerazione, infatti, non sono solamente quelli che la confermano, come sembrerebbe logico fare, ma anche quelli che la smentiscono. Di fatto, una teoria, per essere scientifica, non solo deve essere confermata ma deve anche poter essere confutata. Ed è qui che arriviamo al secondo criterio, la falsificabilità. Come

[1] Con metodo scientifico si intende qui quel procedimento adottato dalla scienza moderna mediante il quale si arriva ad una descrizione vera della realtà, cioè oggettiva e verificabile, chiamato anche metodo induttivo-sperimentale.

appunto sostengono McEnery e Hardie, nel momento in cui ci approcciamo ad una raccolta di dati con un'ipotesi in mente, questa ipotesi dev'essere testata non solamente con i dati che la confermano ma anche quelli che la confutano. Secondo questo criterio formulato dal filosofo Karl Popper, infatti, una teoria, per essere ritenuta scientifica, deve essere confutabile. In mancanza di questo elemento, sempre secondo il filosofo, si uscirebbe dall'ambito della scienza entrando in quello della metafisica.
Il terzo criterio è infine quello della replicabilità. Stando a questo criterio, il procedimento che ha portato alla conferma di una teoria deve poter essere replicato, e portare allo stesso risultato, perché la teoria possa considerarsi effettivamente valida.

Questo aspetto scientifico della linguistica dei corpora permette, in base ai risultati, di conservare o abbandonare l'ipotesi di partenza. Ecco perché la linguistica è stata paragonata ad una scienza fisica. Ci sono infatti interi campi delle scienze naturali, come l'astronomia, la geologia e la paleontologia, il cui studio si basa non su esperimenti in laboratorio ma sulla raccolta di grandi quantità di dati basati sull'osservazione (McEnery, Hardie 2012: 26). Questi dati saranno poi analizzati seguendo proprio il metodo scientifico.

1.2.1 Analisi qualitativa e quantitativa

I dati linguistici che i corpora ci permettono di ricavare, vengono in genere successivamente analizzati dal linguista. L'analisi dei dati può essere svolta mediante due diversi approcci. Il primo approccio è quello dell'analisi qualitativa. Mediante questo tipo di analisi, i dati vengono usati come base per identificare e descrivere determinati utilizzi della lingua e per fornire esempi di reale utilizzo di particolari fenomeni linguistici (McEnery, Wilson 2001: 76). Lo svantaggio di questo approccio è che i risultati ottenuti potrebbero non essere significativi in quanto dovuti al semplice caso.
Il secondo approccio è quello dell'analisi quantitativa. Con questo secondo tipo di analisi, i dati ricavati dal corpus vengono classificati, contati e analizzati secondo modelli statistici (McEnery, Wilson 2001: 76). In questo caso però, è possibile, mediante dei test chiamati *significance test*, determinare se i risultati siano dovuti al caso oppure siano effettivamente rilevanti.
Avendo visto quali sono i pro e i contro dei due approcci, possiamo dire che il miglior modo per ottenere risultati il più concreti possibile nell'analisi di un fenomeno linguistico, è quello di sfruttare la combinazione di entrambi.

Ma vediamo nel concreto a quali risultati possiamo arrivare con l'utilizzo dell'analisi qualitativa e quantitativa dei dati nell'ambito della linguistica dei corpora.
Come ci viene spiegato da McEnery e Hardie, la maggior parte degli studi compiuti dalla linguistica dei corpora, di fatto ha alla base quella branca della statistica chiamata *statistica descrittiva* (McEnery, Hardie 2012: 49). Lo scopo della statistica descrittiva è quello di analizzare e sintetizzare i dati raccolti in un esperimento, che nel nostro caso sono quelli ottenuti da una ricerca attraverso l'uso dei corpora. La misura più utilizzata è quella della frequenza, ovvero il conteggio del numero di occorrenze di un elemento linguistico. Grazie alla linguistica dei corpora, infatti, si è rafforzata la consapevolezza del fatto che comportamenti linguistici ripetuti da parte di molti parlanti sono significativi per lo studio di una lingua (Gatto 2014: 7). Una delle funzioni principali dei corpora moderni è proprio quella di fornire delle liste di frequenza delle parole contenute nei corpora, che indicano quante volte ogni singola parola compare all'interno del corpus intero. Prendiamo come esempio il caso in cui ci rivolgiamo ad un corpus per verificare se l'utilizzo di un dato lessema sia più o meno frequente in una lingua, ad esempio rispetto ad un altro suo sinonimo. Per fare un esempio, proviamo a verificare se sia più frequente nella lingua russa la parola *kniga* oppure *knižka*. Nel NKRJa la parola книга ricorre 18 202 volte, mentre *knižka* 3 160. Possiamo quindi trarre la conclusione che, nella lingua scritta, il sostantivo книга è più frequentemente utilizzato rispetto al sostantivo e suo diminutivo книжка.

Un'altra misura impiegata dalla statistica descrittiva è quella della percentuale. In base alle occorrenze di un determinato elemento linguistico all'interno di un corpus, possiamo stabilire in che percentuale esso sia presente rispetto all'intera quantità di dati del corpus. In questo caso non avremo come risultato la frequenza d'uso di quell'elemento ma quella che viene chiamata "frequenza relativa" (McEnery, Hardie 2012: 49). Prendiamo anche in questo caso come esempio il sostantivo *kniga*. Nel NKRJa, le cui occorrenze totali ammontano a 288 727 494 parole, se il sostantivo ricorre 18 202 volte questo significa che la percentuale di occorrenze di questa parola nell'intero corpus è del 0.0063 %.

Nf = (18 202/288 727 494) x 100 = 0.0063[2]

[2] Calcolo basato sulla formula riportata da McEnery e Hardie (2012:49), utilizzando però dati concernenti la lingua russa, ovvero *nf* = (number of examples of the word in the whole corpus ÷ size of corpus) x base of normalization

Per fare un confronto, utilizziamo il web-corpus GICR (General Internet-Corpus of Russian). Da una ricerca basata sulle pagine web della rete sociale russa VKontakte[3], il cui numero totale di parole è 9820 milioni, il sostantivo *kniga* ricorre qui 200 615 volte. Utilizzando la stessa formula risulta che la percentuale di occorrenza di questa parola all'interno di questo corpus è questa volta del 0.0020 %.

Nf = (200 615/9820 000 000) x 100 = 0.0020

Ma la statistica, se usata rimanendo nell'ambito puramente descrittivo, quindi con un approccio qualitativo, può portare fuori strada. Nella linguistica dei corpora, la descrizione dei dati ricavati è sì utile ma spesso insufficiente. Ecco perché è necessario fare un passo in più introducendo l'approccio quantitativo mediante il sopracitato *significance test*, che è appunto un metodo quantitativo (McEnery, Hardie 2012: 51). Questo test ci permette di distinguere un risultato ottenuto per semplice coincidenza, rischio che si corre nell'ambito di un'analisi puramente qualitativa, da un altro che è invece effettivamente rilevante. Affinché un risultato possa essere considerato rilevante, ci deve essere almeno il 95 % di possibilità che il risultato ottenuto non sia una mera coincidenza, altrimenti tale risultato è da considerarsi non attendibile. Il test in questione può essere principalmente di due tipi: uno è il calcolo delle parole chiave, che consiste nel testare la rilevanza di ogni parola che occorre in un corpus e confrontarne la frequenza con quella della stessa parola in un altro corpus. L'altro è il calcolo delle collocazioni, che consiste invece nel calcolare la frequenza della co-occorrenza di una parola e di tutto ciò che appare attorno a quella parola, una o più volte, nel corpus. Tra i molteplici metodi applicati nell'analisi linguistica quantitativa, questi sono solamente due tra i più importanti e più utilizzati, riportati spesso come gli esempi più rappresentativi. Molti altri sono quelli potenzialmente applicabili ma che non vedremo qui in quanto si tratta spesso di metodi molto complessi. Fortunatamente, "calcoli" di questo tipo avvengono in maniera automatica nella maggior parte dei moderni corpora elettronici. In questo modo l'utente otterrà risultati già di per sé attendibili.
Ecco quindi spiegato il perché la combinazione dei due approcci sia solitamente il modo migliore per svolgere una valida analisi dei dati linguistici. Perciò, riassumendo, mentre i risultati ottenuti

[3] Il social network numero uno in Russia, con 70 milioni di utenti. Creato nel 2006, gioca ora il ruolo del Facebook russo.

con la semplice statistica descrittiva sono considerati frutto di un'analisi quantitativa, quelli ottenuti mediante l'integrazione del *significance test* sono invece frutto di un'analisi qualitativa.

1.3 Excursus storico della disciplina

Nel delineare la storia della disciplina della *corpus linguistics*, è possibile individuare due fasi principali in cui essa si sviluppò (McEnery, Hardie 2012: 225-227). La prima è quella che va fino alla fine degli anni Ottanta del Novecento. In questa prima fase vediamo l'emergere della linguistica dei corpora come metodo, in particolare nell'ambito degli studi sulla lingua inglese. Una volta acquisita una certa importanza come metodologia di studio della lingua, dovette intraprendere una lotta per affermarsi dopo un periodo di dure critiche che indebolirono fortemente la sua credibilità. Queste critiche, che vedremo nello specifico tra poco, sono quelle mosse dal noto linguista statunitense Noam Chomsky e che fecero scivolare il metodo dei corpora da una posizione dominante ad una posizione di sostanziale marginalità. Vediamo infine, negli anni Ottanta, come questa disciplina si risollevò nonostante il periodo buio, riaffermandosi come metodo predominante di ricerca linguistica. La seconda fase, che da qui arriva fino ai giorni nostri, è una fase in cui assistiamo ad un vero e proprio cambiamento nella natura della *corpus linguistics*. Come osservato da McEnery e Hardie, dall'essere in pratica una branca della linguistica divenne una vera e propria "*enterprise*" (McEnery, Hardie 2012: 226) che è oggi una componente essenziale della metodologia di ricerca linguistica.

Oltre a questa suddivisione temporale proposta da McEnery e Hardie, si è soliti ritrovare nella letteratura un riferimento a due archi temporali distinti della storia della *corpus linguistics*. Lo snodo che sta nel mezzo è proprio Chomsky con la sua critica. Infatti, le due fasi di questa disciplina di cui spesso si parla descrivono il prima e il dopo l'avvento di Chomsky. La prima di queste due, l'era, quindi, pre-chomskiana, viene denominata da McEnery e Wilson "*early corpus linguistics*". Questa denominazione fu coniata a posteriori sulla base di quella moderna (*corpus linguistics*) in quanto ha con essa molte affinità. Il metodo adottato dalla linguistica pre-chomskiana prevedeva, in quanto basato sull'osservazione degli usi della lingua, l'utilizzo dei corpora, ed è quello che oggi definiamo metodo *corpus-based* (vedremo più precisamente la definizione di questo metodo tra qualche paragrafo). Riportando le parole di Harris, McEnery e Wilson ci spiegano che questo approccio iniziò con un'ampia collezione di "conversazioni" linguistiche in alcune lingue che venivano annotate, ovvero un corpus.

La *early corpus linguistics* vide i suoi albori nel IXX secolo e tra questo e l'inizio del secolo successivo veniva applicata in vari ambiti. Uno di questi è quello degli studi sull'acquisizione linguistica nei bambini. In questo caso si basava su diari accuratamente scritti dai genitori in cui annotavano le frasi pronunciate dai loro figli. Un'ampia raccolta di questi diari andava a formare il corpus da cui venivano poi ricavate le norme con cui l'acquisizione del linguaggio avveniva nei bambini. Sempre nel XX secolo vediamo impiegati i corpora linguistici in ambiti quali la didattica delle lingue straniere, la linguistica compartiva, e poi ancora altri settori specifici della linguistica come la sintassi e la semantica.

Il corpus, come abbiamo detto, è alla base della metodologia linguistica *corpus-based* che è arrivata fino ai giorni nostri. Ma qual è l'origine di questo strumento che tanto viene oggi utilizzato? Barbera (2013:7) ci spiega che il capostipite di tutti i corpora attuali è il *Brown Corpus of American Written English*, compilato da Winthrop Nelson Francis ed Henry Kučera alla Brown University del Rhode Island e pubblicato nel 1964. Questo è infatti il primo corpus a soddisfare in tutto e per tutto la moderna definizione formale. Ad aver realmente inaugurato questa tradizione fu però Charles Carpenter Fries, che negli anni Cinquanta pubblicò una grammatica descrittiva della lingua inglese parlata basandosi sulla registrazione di 250 000 parole di conversazioni telefoniche. Se guardiamo invece al contesto italiano, il ruolo di Fries lo gioca Padre Roberto Busa SJ, autore dell'opera su Tommaso d'Aquino iniziata nel 1949 ma comunque già fondata su spogli elettronici. Ma se, come abbiamo detto, la *early corpus linguistics* nasce a partire dal IXX secolo, significa che lo strumento del corpus doveva esistere già a quell'epoca. A dire la verità, però, "secondo la definizione moderna, si possono considerare corpora solo gli 'oggetti' nati dagli anni Sessanta in poi, nell'era, cioè, dei computer, che, per la nostra disciplina, potremmo ben chiamare post-Brown". "Gli 'oggetti' precedenti, come quelli approntati dal Fries", che vengono qualificati come "'corpora preistorici' restano in effetti fuori dai paletti della nostra definizione di quei corpora che pure di essi sono i naturali discendenti" (Barbera, Corino, Onesti 2007: 33). Questi erano infatti oggetti non informatici, mentre nella definizione odierna di corpus l'elemento informatico, il cosiddetto *machine-readable format*, è fondamentale.

1.3.1 Chomsky e la sua critica alla *corpus linguistics*

La *corpus linguistics* si è affermata fino ai giorni nostri come metodologia linguistica, ma questo non senza ostacoli. Come abbiamo detto, è stata vittima di un periodo di impopolarità a causa delle critiche mosse da Chomsky negli anni Cinquanta del Novecento. Solamente dagli anni Ottanta

questa metodologia si riaffermerà più forte di prima confermandosi una fonte indispensabile di evidenza linguistica. Approfondire la questione e arrivare al cuore della critica è fondamentale, come sostengono McEnery e Wilson (2001:5), per comprendere a fondo le ragioni di un tale successo del metodo *corpus-based* e come esso si è sviluppato. È proprio questa critica, per l'appunto, ad aver suscitato la reazione da parte dei linguisti a sostegno dell'utilizzo dei corpora. Inoltre, è in questo contesto che si svilupperanno due dei concetti cardine di questa disciplina che sono, come vedremo nella sezione dedicata alle caratteristiche dei corpora, rappresentatività e bilanciamento. Ma veniamo al dunque. Le argomentazioni formulate da Noam Chomsky contro l'utilizzo dei corpora nell'ambito degli studi di linguistica, di cui ne danno un quadro generale ma chiaro McEnery e Wilson (2001: 5-24), hanno alla base il classico dibattito tra razionalisti ed empiristi. Questo dibattito, oltre al nostro caso, può riguardare ogni disciplina in cui è possibile condurre una ricerca dovendo scegliere se basarsi su elementi presenti in natura oppure su osservazioni prodotte "artificialmente", per utilizzare le parole dei due autori. Infatti, per chiarire meglio i due concetti, secondo la teoria razionalista, i dati su cui essa si basa devono essere dati comportamentali artificiali e giudizi introspettivi razionalmente formulati. Le teorie razionaliste hanno alla base, nell'ambito della linguistica, lo sviluppo di teorie mentali il cui principale obiettivo è l'essere attendibili dal punto cognitivo, razionale, sulla base delle conoscenze linguistiche che un soggetto già possiede. L'approccio empirico al linguaggio è dall'altro canto dominato dall'osservazione di dati reali di utilizzo della lingua, tipicamente con l'uso dei corpora. Per quanto riguarda la linguistica, quindi, questo dibattito riguarda la natura del dato linguistico preso in considerazione per formulare una teoria linguistica. Ovviamente, entrambi gli approcci hanno i loro pro e contro. L'approccio prediletto da Chomsky fu quello razionale, e, di conseguenza, cambiò l'oggetto dell'indagine linguistica dalle descrizioni astratte di uso della lingua, come avviene con il metodo *corpus-based*, a teorie che rispecchiano la realtà psicologica, ovvero modelli di linguaggio possibili da un punto di vista razionale. Così facendo, ha invalidato il corpus come fonte di evidenza scientifica nell'indagine linguistica. Egli suggerì invece che il corpus non potesse essere uno strumento utile per il linguista in quanto l'obiettivo di quest'ultimo è sia quello di mirare a modelli di competenza linguistica che a delle mere performance linguistiche. Difatti, Chomsky fa proprio una distinzione tra *competenza* e *performance*. La competenza viene descritta come la nostra conoscenza tacita, interiorizzata di una lingua, le regole linguistiche che noi conosciamo e che ci permettono di utilizzare la lingua in modo corretto. La performance, invece, è l'evidenza esterna della nostra competenza linguistica e il suo uso in particolari occasioni, che può essere influenzato ed alterato da altri fattori (l'esempio riportato da Chomsky è quello del parlante sotto effetto di alcolici, il cui corretto uso della lingua potrebbe essere compromesso). È per questo motivo, quindi,

che a mostrare realmente la conoscenza linguistica di un parlante non è la performance bensì la competenza. Dato che il linguista individuava come compito principale della linguistica quello di creare e definire un modello di competenza linguistica, e non quello di enumerare e descrivere fenomeni linguistici, un corpus non è un buono strumento per perseguire questo scopo. Ecco perché Chomsky sosteneva la necessità di passare da una prospettiva empirista ad una prospettiva razionalista.
Un'altra critica che Chomsky muove nei confronti dei corpora è quella di essere uno strumento incompleto e alterato. Un corpus, in quanto strumento di dimensioni limitate benché ampie, contiene molte ma non tutte le frasi di una lingua naturale. Alcune frasi, infatti, non sono presenti nei corpora in quanto ovvie, false o anche volgari, come sottolinea il linguista.

Per concludere, possiamo riassumere dicendo che l'approccio di Chomsky nei confronti della linguistica è razionalista e introspettivo, al contrario di quello che è l'approccio odierno, empirico e basato su dati reali di uso di una lingua. Come abbiamo visto, però, nonostante questa critica, i corpora si sono confermati una valida fonte di dati quantitativi e, come sostenuto da Leech, una valida metodologia del punto di vista del metodo scientifico poiché sono aperti ad un'oggettiva verifica dei risultati (Leech, 1992). Mentre la critica di Chomsky può risultarci oggi come totalmente errata, è possibile giustificare parte di essa considerando che, come ci spiegano McEnery-Wilson (2001:78), nel momento in cui la formulò i corpora erano entità davvero piccole rispetto ad oggi, che sono invece molto più rappresentativi di una lingua grazie alle dimensioni importanti che hanno raggiunto. È quindi comprensibile il fatto che screditasse i corpora come strumento per la ricerca linguistica. A sostegno dell'approccio odierno, ovvero quello empirico, possiamo invece dire che spesso, nel caso di una lingua, il fatto che ci sia una regola specifica di utilizzo non impedisce che essa venga ripetutamente infranta fino alla creazione di una nuova regola. Una regola è quindi da tenere in considerazione fino a che una nuova non subentri e la spazzi via. Basarsi solamente su regole linguistiche già esistenti, come avviene per l'approccio razionalista, spesso non è sufficiente. Ed è proprio qui ad entrare in gioco la necessità di osservare le performance linguistiche per verificare quanto una regola venga rispettata nell'utilizzo di una lingua e quando questa cominci a venire infranta. Si prenda ad esempio l'uso del congiuntivo nella lingua italiana. Sempre più si sta allargando la tendenza ad abbandonarlo in determinate situazioni, tanto che secondo alcuni linguisti potrebbe, tra molti anni, scomparire. Che questo sia vero o no, è così che le lingue funzionano, e se lo sappiamo è solo grazie alle performance di lingua scritta o parlata del passato di cui oggi disponiamo proprio grazie ai corpora.
Le critiche di Chomsky, alla fin fine, si sono rivelate non valide ed hanno contribuito a rafforzare l'utilizzo e lo sviluppo di questi strumenti che ancora oggi sono tanto, e sempre più, utilizzati. Ad

avere avuto un ruolo chiave nel loro sviluppo è stato il parallelo evolversi dell'informatica e dei computer. Negli ultimi cinquant'anni, infatti, i computer sono diventati strumenti sempre più economici, veloci e accurati nel processo di analisi dei dati testuali.

1.3.2 Sinclair e la teoria *Neo-Firthian*

Un altro importante contributo alla *corpus linguistics* proviene da un gruppo di studiosi chiamati collettivamente *neo-Firthian* (McEnery, Hardie 2012: 122), che emerse nella stessa epoca della critica di Chomsky, ovvero negli anni Cinquanta del Novecento. Il nome di questo gruppo deriva dal linguista che propose questo approccio al linguaggio, J. R. Firth. Come ci spiegano i due autori, tra i più fedeli sostenitori di questo approccio ritroviamo il linguista John Sinclair, pioniere della linguistica dei corpora, che fu uno dei primi a fare delle idee di Firth una vera e propria metodologia della linguistica dei corpora.
Un primo contributo da parte di Firth riguarda la sfera concettuale. Firth invitava i linguisti a studiare quella che definiva *attested language* (McEnery, Wilson 2001: 23). Si pose quindi dalla parte degli empiristi sostenendo la necessità di basarsi su esempi concreti di utilizzo della lingua di cui si ha testimonianza. Non solo, Firth è anche ricordato per la sua nozione fondamentale di *contextual language*. Secondo il linguista, quello linguistico è un atto fortemente contestuale, che dipende da fattori psicologici, accademici, sociali e personali, acquisiti dal parlante durante gli eventi della sua vita.

Da un punto di vista terminologico Firth è invece noto per aver introdotto un termine che è oggi uno dei punti cardine della linguistica dei corpora, le *collocations*, chiamate in italiano collocazioni. Questo concetto non era nuovo nell'ambito della linguistica, veniva infatti utilizzato già a partire dagli anni Trenta del Novecento. A quell'epoca però ci si riferiva a questo fenomeno con il termine *automation*, fino a che, circa vent'anni dopo, non subentrò il nuovo termine tutt'ora utilizzato. Essendo appunto una nozione cardine della linguistica dei corpora, il tema delle collocazioni sarà trattato più nel dettaglio in una sezione apposita di questo lavoro.
Infine è possibile sottolineare una differenza tra la tradizione *Neo-Firthian* e il tradizionale approccio *corpus-based* in voga a quel tempo per quanto riguarda i dati alla base degli studi linguistici. Mentre quest'ultimo approccio basava il suo lavoro su dei campioni di produzione linguistica che fossero rappresentativi di una determinata lingua, la tradizione *neo-Firthian* si basava sull'analisi di interi testi.

1.4 Approccio *corpus based* vs *corpus driven*

Fino ad ora abbiamo in generale parlato di approccio *corpus-based* come approccio della linguistica che fa dei corpora il suo strumento principale. Nello specifico però, possiamo più tecnicamente parlare di approccio *corpus-based* come metodo di ricerca linguistica che utilizza i dati dei corpora allo scopo di confermare o smentire una teoria linguistica (McEnery, Hardie 2012: 6). Siamo quindi, come abbiamo visto, nell'ambito del metodo scientifico, dove il corpus costituisce una fonte di dati empirici, qualitativi o quantitativi. A questo si oppone però un altro approccio, che è quello *corpus-driven*. In questo caso, il corpus non è visto come un metodo ma è invece il corpus stesso ad enunciare delle teorie linguistiche in base ai dati linguistici in esso contenuti, invece che esserne un elemento di conferma o smentita. Il corpus è quindi esso stesso una teoria. McEnery e Hardie parlano anche di *corpus-as-theory* nel caso dell'approccio *corpus-driven*, e di *corpus-as-method* per quanto riguarda invece l'approccio *corpus-based*.

Nel parlare di *corpus linguistics* come una vera e propria metodologia di ricerca siamo ovviamente nell'ambito dell'approccio *corpus-based.* Era però interessante mostrare anche una corrente della disciplina che utilizza il corpus non come strumento scientifico ma come strumento che si potrebbe definire "normativo".

1.5 Cos'è un corpus?

Ora che abbiamo inquadrato sia storicamente che metodologicamente la *corpus linguistics*, non ci resta che analizzare nel dettaglio lo strumento cardine utilizzato da questa disciplina, il corpus, partendo dalla sua definizione fino a vederne le sue svariate caratteristiche e tipologie.

La definizione generica di corpus, che ci viene fornita da un qualsiasi dizionario, è "raccolta completa di testi e di opere costituita secondo un particolare criterio"[4]. In effetti, come ci ricordano McEnery e Wilson (2001: 29), qualsiasi collezione di testi che ne contenga due o più è da considerarsi un corpus in base a quello che è il significato della parola latina *corpus*, ovvero corpo, in questo caso di testi. Nel gergo della linguistica, però, la parola che si riferisce a questo strumento ha assunto un significato differente, con connotazioni più specifiche. Sono vari i linguisti ad averne dato una definizione in questo ambito: "a collection of texts assumed to be representative of a given

[4] https://dizionari.corriere.it/dizionario_italiano/C/corpus.shtml

language, dialect or other subset of a language, to be used for linguistic analysis" è la definizione data da Francis (1982: 7); "Finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration" è quella data da McEnery e Wilson (2001: 32); poi abbiamo ancora la definizione di Sinclair: "a collection of naturally-occurring language chosen to characterize a state or variety of language" (Sinclair 1991: 171); e quella di Hunston: "a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study" (Hunston 2002: 2). Infine, trattandosi di un lavoro incentrato su una ricerca applicata alla lingua russa, mi sembrava interessante riportare la definizione russa di corpus linguistico data dal NKRJa: "Корпус — это информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме."[5] La definizione che tuttavia risulta essere la più completa nell'ambito della linguistica dei corpora, includendo molte delle caratteristiche dei corpora che vedremo tra poco, è quella proposta da Barbera. Un corpus viene da lui descritto come "una 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 per lo più scelti in modo da essere autentici e rappresentativi" (Barbera 2013: 14). Questo autore ha messo insieme, nella definizione da lui data, gli usi prevalenti che della parola *corpus* la comunità di linguisti ha fatto e fa, soprattutto nell'epoca moderna con l'avvento dei computer e il conseguente sviluppo della *corpus linguistics*. Secondo Barbera, infatti, in base a quella che è l'idea moderna di corpus, "in assenza delle due specifiche caratteristiche '*markup*' e 'tokenizzazione' […] non si può […] parlare di corpora ma solo, genericamente di (raccolte di) testi in formato elettronico" (Barbera, Corino, Onesti 2007: 26)

Passiamo ora ad illustrare quelle che sono le caratteristiche distintive dei corpora nell'ambito della linguistica. Per fare questo sono state raccolte le caratteristiche citate dagli autori più importanti, ma non solo, di lavori relativi alla linguistica dei corpora. Alla fine di questo lavoro di ricerca sono emersi in totale 11 elementi, tra cui la maggior parte imprescindibili per rientrare nella definizione di corpus che abbiamo visito, e pochi altri che non sono sempre citati tra le caratteristiche fondamentali dei corpora ma che sono state menzionate più volte da alcuni autori. Queste sono: autenticità, rappresentatività, formato elettronico, finitezza, grandi dimensioni, bilanciamento, campionamento, *markup* e annotazioni, ordinatezza finalizzata, standard, comparabilità.

[5] http://www.ruscorpora.ru/new/corpora-intro.html

1.5.1 Autenticità

Citata da gran parte degli autori, tra cui alcuni importantissimi quali Sinclair (1987 e 1991), Biber et al. (1998), Sampson-McCarthy (2004), McEnery (2003), Tognini-Bonelli (2001) e altri, è un principio guida dell'intera disciplina fin dagli inizi. L'attenzione, infatti, è sempre stata rivolta ad una raccolta di dati che fossero reali, ricavati da una lingua effettivamente prodotta e usata dai suoi parlanti. Come abbiamo detto all'inizio parlando della linguistica dei corpora, essa è una metodologia basata su esempi reali di utilizzo del linguaggio. Questo è anche ciò che Chomsky criticava dell'approccio empirista della linguistica a cui contrapponeva il suo approccio cognitivo, il punto di rottura tra l'era pre e post-Chomskiana. Come abbiamo visto, però, il dato reale ed autentico si è confermato il solo e unico preso in considerazione da questa disciplina per le sue ricerche, tanto che l'autenticità rientra sempre tra le prime caratteristiche citate nel descrivere i corpora linguistici.

1.5.2 Formato elettronico

Lo abbiamo già visto in precedenza dalle parole di Barbera, e a ricordarcelo sono anche altri autori tra cui McEnery-Wilson (2001), Bowker-Pearson (2002), Sinclair (2005), McEnery-Gabrielatos (2006) e altri autori secondari. Il formato elettronico "è il criterio più presente nelle definizioni" (Barbera, Corino, Onesti 2007: 54). È anche il requisito che segna il passaggio dall'era storica e preistorica della linguistica dei corpora a quella moderna. Dal momento che i computer ancora non esistevano, nelle prime definizioni di corpus questo elemento non era presente. Come ci spiegano Barbera-Corino-Onesti (2007: 54), la prima volta in cui è stato introdotto fu nella scuola sincleriana. Tuttavia, negli anni Novanta, Sinclair distingueva ancora tra *corpora* e *computer corpora*. Questo fattore è diventato poi centrale negli ultimi decenni, fino ad essere ora un requisito indispensabile per soddisfare la definizione moderna di corpus. McEnery-Wilson (2001: 31) osservano che al giorno d'oggi esistono ancora alcuni corpora pubblicati in formato cartaceo ma sono diventati ormai molto rari, se non quasi introvabili. Questi, più che dei veri corpora possono essere considerati una documentazione dei corpora, che non possono però essere usati come tale. C'è da aggiungere anche che il formato elettronico, seppur una condizione necessaria dei moderni corpora, non è di per sé sufficiente, in quanto, come abbiamo visto nel paragrafo dedicato alla definizione di corpus, il testo elettronico deve anche essere codificato tramite *markup* e tokenizzazione, processi che vedremo nel dettaglio più avanti. Il vantaggio di avere un testo in formato elettronico è infatti quello di poter

effettuare ricerche tali che in formato cartaceo non sarebbe possibile fare, specialmente con le dimensioni che i corpora attuali hanno raggiunto. Prendiamo il caso più semplice di dover ricercare tutte le occorrenze di una parola in un corpus di milioni di parole: questa azione, seppur ancora possibile manualmente, richiederebbe tempi lunghissimi di lavoro e margini di errore molto ampi. Ma il problema diventa insormontabile con ricerche più elaborate come ad esempio mostrare tutte le collocazioni di un termine o le molte altre ricerche, talvolta più matematiche che altro, che i corpora permettono si svolgere.

1.5.3 Grandi dimensioni

Nonostante la grande dimensione sia una caratteristica relativa e mai espressa in termini ben definiti, viene spesso citata nelle definizioni corpus, come ad esempio in Bowker-Pearson (2002), McEnery (2003), Baker-Hardie-McEnery-Gabrielatos (2006). Sinclair (2005), a questo proposito, sostiene che non ci sia una dimensione massima che un corpus può raggiungere. Essa, infatti, può variare a seconda della natura e dello scopo del corpus. La grande dimensione è però necessaria in quanto, come detto, il corpus dev'essere sufficientemente rappresentativo della lingua presa in esame. È comunque vero che nel caso si tratti di un corpus specializzato, che come vedremo più avanti è una tipologia di corpus che rappresenta solamente una varietà ristretta di una lingua, la dimensione può essere relativamente più piccola di uno che dall'altro canto rappresenta un'intera lingua naturale. C'è anche da dire che la percezione delle dimensioni di un corpus è cambiata col passare del tempo e l'evolversi della disciplina. Il primo corpus elettronico, il Brown Corpus, contiene 1 milione di parole. Mentre allora era considerata una dimensione notevole, oggigiorno è da considerarsi piuttosto piccolo, paragonandolo ad esempio al moderno British National Corpus (BNC) che contiene oggigiorno 100 milioni di parole. La dimensione è quindi una questione relativa. Se consideriamo che, come ci spiega Gatto (2014: 14), ci devono essere almeno 20 occorrenze di un fenomeno linguistico in un corpus perché si possa considerare un dato quantitativo valido, risulta chiaro come la dimensione possa influire sull'utilità di un determinato corpus. E non solo, più un corpus è grande, più può essere utile per effettuare ricerche basate non solo su parole ma anche su intere frasi, che si avrà più possibilità di svolgere con l'utilizzo di corpora di grandi dimensioni.

1.5.4 Rappresentatività

Anche questo è un punto cardine della caratterizzazione dei corpora, citato da numerosi autori importanti quali Sinclair (1987), Francis (1982), Biber et al. (1998), McEnery-Wilson (2001), Tognini-Bonelli (2001), Sampson-McCarthy (2004) e molti altri. I testi contenuti in un corpus devono costituire un campione della lingua studiata che ne riproduca tutte le caratteristiche. Analizzando poi tali dati linguistici deve essere possibile "risalire a conclusioni valide ad un livello più ampio e generalizzato dello studio linguistico" (Barbera 2013: 39). Prendendo in considerazione una lingua viva e parlata, il numero di enunciati sarebbe teoricamente infinito poiché essa è in continua crescita. Dal momento che analizzarli tutti sarebbe impossibile, nel creare un corpus l'intento è quello di ottenere un campione "which is maximally representative of the variety under examination, that is, which provides us with as accurate a picture as possible of the tendencies of that variety, including their proportions" (McEnery, Wilson 2001: 30)

1.5.5 Bilanciamento

Questa caratteristica, citata da autori come McEnery-Xiao-Tono (2006) e Gatto (2014), viene ritenuta fondamentale affinché si verifichi il secondo requisito di questo elenco, la rappresentatività. La rappresentatività di un corpus, infatti, specialmente quando parliamo di corpora generali di una lingua, come il British National Corpus, o il NKRJa, dipende dal fatto che esso sia ben bilanciato nella sua composizione. Affinché questo si verifichi, deve includere un'ampia gamma di categorie testuali che si suppone siano rappresentative della lingua che esso rappresenta. Quindi, nel creare un corpus, testi di diverse categorie vengono inseriti proporzionalmente di modo che "it offers a manageably small scale model of the linguistic material which the corpus builders wish to study" (Atkins et al. 1992: 6). Per determinare il bilanciamento di un corpus non ci sono misure scientifiche ma, come sostengono McEnery-Xiao-Tono (2006: 16), è "largely an act of faith rather than a statement of fact" e si basa per lo più su "intuition and best estimates". Sempre grazie a questi autori sappiamo che spesso, per determinare il bilanciamento di un corpus, i creatori si basano su corpora già esistenti che sono generalmente riconosciuti come corpora bilanciati, ad esempio il British National Corpus. Guardando ai criteri con cui un corpus come questo è stato costruito può essere d'aiuto per capire come costruire a loro volta un corpus che sia adeguatamente bilanciato.

1.5.6 Campionamento

Anche questa caratteristica, citata da McEnery-Wilson (2001), McEnery-Xiao-Tono (2006), Gatto (2014), è strettamente legata a quella della rappresentatività. Come abbiamo detto, perché un corpus sia rappresentativo di una lingua, i testi in esso contenuti devono costituire un campione della lingua studiata che ne riproduca tutte le caratteristiche, dal momento che descrivere una lingua naturale nella sua totalità è concretamente impossibile. Una volta constatato questo, però, bisogna decidere in che modo campionare i testi di un corpus di modo che esso risulti il più rappresentativo possibile. Come riportano McEnery-Xiao-Tono (2006: 20) dalle parole di Manning e Schütze (1999), "a corpus is a *sample* of a much larger *population*. A sample is assumed to be representative if what we find for the sample also holds for the general population". Ci spiegano poi che per ottenere un campione rappresentativo bisogna per prima cosa definire l'"unità di campionamento" (*sampling unit*) e i limiti di quella che chiamano "*population*". Prendendo l'esempio di un corpus di testi scritti, l'unità di campionamento potrebbe essere libri, periodici o quotidiani. Mentre la *population* è l'insieme di tutte le unità di campionamento, quella che viene chiamata *sampling frame* è la lista di unità di campionamento. Se guardiamo al NKRJa, la cosiddetta *population* da cui sono stati estratti i campioni di testo sono testi scritti pubblicati in Russia dall'inizio del XVIII secolo, come riporta il sito del corpus. La lista di unità di campionamento che compongono la *sampling frame* è invece "прозаические оригинальные тексты, представляющие русский литературный язык […], но также и (в меньшем объёме) переводные сочинения (параллельно с оригиналом), поэтические тексты, а также тексты, представляющие нелитературные формы современного русского языка: разговорную (записи устной речи, публичной и непубличной), диалектную[6]. Una volta stabiliti questi due elementi, rimane però la questione di quanto i campioni debbano essere grandi. Secondo Biber (1993), dal momento che le caratteristiche linguistiche ricorrenti si distribuiscono in maniera equilibrata, piccole porzioni di testo sono sufficienti a studiare tali caratteristiche, mentre quelle più rare necessitano di campioni più grandi. La dimensione e il numero di campioni linguistici necessari affinché il corpus possa essere considerato rappresentativo, dipendono quindi dalla loro frequenza o peso in quella determinata lingua. Per quanto riguarda invece la proporzione e il numero di campioni per ogni categoria di testo, McEnery-Xiao-Tono sostengono che questi debbano essere "proportional to their frequencies and/or weights in the target population in order for the resulting corpus to be considered as representative" (2006: 20).

[6] Ibidem

1.5.7 Finitezza

Nonostante questa caratteristica non si ritrovi in molti autori, uno di questi è McEnery-Wilson (2001), viene da essi considerata fondamentale per la metodologia impiegata dalla linguistica dei corpora, che si avvale, come abbiamo visto, della statistica. Infatti, come sottolineato da Barbera-Corino-Onesti (2007: 51), "se consideriamo […] l'uso della statistica come una caratteristica individuante da sempre la *corpus linguistics* rispetto ad altre discipline linguistiche […], è condizione matematicamente banale che gli insiemi di elementi su cui opera debbano essere finiti. Più in generale, inoltre, la finitezza di un corpus ne garantisce la possibilità di operare entro confini scientificamente ed univocamente 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 […] degli esperimenti". Il tema della finitezza, comunque, sarà trattato più approfonditamente successivamente, quando si parlerà del concetto di *web as corpus*, in cui questo è un punto alquanto spinoso.

1.5.8 *Markup* e annotazioni

Il *markup* e le annotazioni, ovvero informazioni aggiuntive applicate ai testi contenuti nei corpora, sono ciò che distinguono un corpus da una semplice raccolta di testi in formato elettronico. Come sostengono Barbera-Corino-Onesti (2007: 56), è notevole il fatto che un esplicito riferimento al *markup* sia presente nella definizione del NKRJa: «Разметка — главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов»[7]. Sono ormai un elemento immancabile nella moderna concezione di corpus e possiamo trovarne una descrizione nei lavori più recenti come McEnery-Wilson (2001), McEnery-Hardie (2012), McEnery-Xiao-Tono (2006). "La sola ad invocare negativamente il *markup*, avanzando un esplicito requisito di semplicità per i corpora" (Barbera-Corino-Onesti 2007: 57) fu la scuola sincleriana (e anche i *neo-firthian*). In questa sezione, ci limiteremo a definire i *markup* e le annotazioni come elemento distintivo dei moderni corpora senza scendere ulteriormente nel dettaglio. Ci sarà, in seguito, una sezione ad essi dedicata in quanto meritano di un ulteriore approfondimento.

[7] http://www.ruscorpora.ru/new/corpora-intro.html

1.5.9 Ordinatezza finalizzata

L'ordinatezza finalizzata è la caratteristica dei corpora di essere ordinati in base ad uno scopo preciso, nella maggior parte dei casi a scopo linguistico, ed è spesso citata come specifica dei corpora, ad esempio da autori come Johansson (1991), Biber et alii (1998), Tognini-Bonelli (2001). Come ci spiegano Barbera-Corino-Onesti (2007: 52), il fatto di essere costruito secondo espliciti criteri di progettazione è "l'elemento determinante per distinguere un corpus da una biblioteca di testi elettronici", così come lo sono *markup* e annotazioni. Infatti, mentre "a corpus designed for linguistic analysis is normally a systematic, planned and structured compilation of text", un semplice archivio elettronico di testi è "repository, often huge and opportunistically collected, and normally not structured" (Kennedy 1998: 3)

1.5.10 Standard reference

Questo elemento viene menzionato solamente dagli autori McEnery-Wilson (2001) ma sono gli stessi ad ammettere che non sia una parte essenziale della definizione di corpus. Infatti, non è una caratteristica normalmente riferita nelle definizioni di corpus. In ogni caso è interessante notare, come ci spiegano i due autori, che "there is also often tacit understanding that a corpus constitutes a standard reference for the language variety which it represents" (McEnery, Wilson 2001: 32). Un corpus, quindi, che sia esso rappresentativo di una lingua naturale o di una sua varietà, si pone come modello di riferimento della lingua o varietà che rappresenta.

1.5.11 Comparabilità

Quest'ultima caratteristica è anch'essa citata solamente in un'opera, ovvero quella di McEnery-Hardie (2012), ma è comunque degna di essere tenuta in considerazione. La comparabilità dei corpora ha valore nel momento in cui si decide di comparare due corpora oppure quando dei corpora vengono appositamente creati per essere comparati ad altri. Questo può avvenire per varie ragioni, come ad esempio esplorare la variazione diacronica di una lingua, comparare i risultati di un quesito oppure fare una traduzione con l'ausilio di corpora paralleli. La comparabilità è anche "un'unità di misura" che verrà tenuta in considerazione in questo lavoro di ricerca, che prevede per l'appunto la comparazione principalmente di due diverse tipologie di corpora, i corpora classici e i web-corpora.

1.6 Tipologie di corpora

Come abbiamo appena visto, sono molte le caratteristiche che contraddistinguono i corpora e che li rendono lo strumento tanto utile e ormai ampiamente utilizzato dal punto di vista della ricerca linguistica. Tante però sono anche le diverse tipologie in cui è possibile ritrovare questo strumento. Alcune sono già state accennate in precedenza, come ad esempio i corpora generali nazionali o quelli specializzati, ma, come anticipato all'inizio del capitolo, andremo a vederle più nel dettaglio una ad una.

È possibile incontrare, nei lavori che trattano i corpora e la *corpus linguistics*, varie categorie in cui le tipologie di corpora vengono generalmente suddivise, a seconda delle loro caratteristiche e del loro scopo primario.
Di ogni tipologia che vedremo ora, verrà solamente accennato un esempio per quanto riguarda i corrispettivi corpora russi, tema che verrà approfondito nel capitolo successivo. Una prima distinzione che possiamo fare, in base al medium che viene campionato, è quella tra corpora di lingua parlata e di lingua scritta. I primi consistono in una raccolta di interazioni orali che vengono successivamente trascritte, mentre i secondi consistono in una raccolta di testi scritti. Sempre in questa categoria sono da includere anche i corpora multimediali, o video corpora, che registrano invece caratteristiche paralinguistiche come ad esempio i gesti compiuti dal parlante durante un discorso. Per quanto riguarda i corpora di lingua scritta, possiamo citare, per la lingua russa, in primo luogo il più conosciuto NKRJa ma anche il BOKR (*BOl'šoj Korpus Russkogo Jazyka*), antenato del NKRJa, anche conosciuto come *Russian Reference Corpus*. Questo corpus ammonta a 100 milioni di parole ed include testi di vario genere scritti in russo moderno. Un corpora di lingua parlata è il *Korpus ustnoj reči*, uno dei sottocorpora del NKRJa, che include sia le registrazioni di conversazioni pubbliche e private che le trascrizioni di film russi, per un arco di tempo che va dal 1930 al 2007. Un corpora multimediale è invece il *Mul'timedijnyj korpus* (MURCO), che anche in questo caso fa parte del NKRJa. Il materiale di cui è composto consiste in frammenti di film russi usciti tra gli anni Trenta e gli anni Duemila.
Una seconda distinzione è invece quella tra corpora generici e corpora specializzati. I corpora generici, i cosiddetti grandi corpora nazionali, sono quei corpora il cui obiettivo è essere (il più possibile) rappresentativi di una lingua. Per questo includono testi di generi e registri diversi e le loro dimensioni sono notevoli. Il corpus generico per eccellenza della lingua russa è ovviamente il NKRJa. Solitamente questi corpora possono essere usati a diversi scopi, come produrre materiale di riferimento per traduzioni o per l'apprendimento linguistico. Dall'altro canto, i corpora specializzati

mirano a rappresentare solamente una certa varietà linguistica o un certo dominio, come possono essere quello medico, accademico o giuridico, giornalistico, e molti altri. Le loro dimensioni sono di solito decisamente più piccole rispetto ai corpora generici, che devono invece contenere tutte le varietà e tutti i domini della lingua in questione. Come esempio di corpus specializzato possiamo citare ancora una volta il *Regensburgskij diachroničeskij korpus russkogo jazyka*, in quanto contiene testi in russo antico.

Un'altra suddivisione può essere fatta in base al numero di lingua coinvolte in un corpus. Abbiamo in questo caso corpora monolingui e multilingui. I corpora monolingui contengono testi in una sola lingua e mirano quindi ad essere rappresentativi di questa lingua. Un esempio possono essere i grandi corpora nazionali ma non solo, lo sono anche corpora di altro tipo come quelli specializzati, purché contengano testi solamente in una lingua. I corpora multilingui invece contengono testi in più di una lingua. Anche all'interno dei corpora multilingui può essere fatta una distinzione. Se contengono testi in solamente due lingue abbiamo i corpora bilingui. I corpora multilingui possono essere di diverso tipo: si parla di corpora paralleli, in inglese chiamati talvolta *translation corpora* oltre che *parallel corpora*, quando essi comprendono testi sia nella loro lingua originale che in traduzione, in una o più lingue. Tali corpora sono generalmente "allineati", ovvero le unità linguistiche dei testi in lingua originale, che possono essere parole o frasi, sono collegate alle corrispondenti unità linguistiche del testo in traduzione. L'allineamento, che viene oggi fatto automaticamente grazie all'aiuto di appositi software, facilita molto il lavoro di traduzione. I corpora paralleli sono infatti generalmente utilizzati come supporto da parte dei traduttori, ma anche in altri campi, come ad esempio quello dell'insegnamento delle lingue straniere. Come ci spiegano McEnery-Wilson (2001), quest'idea non è moderna, anzi, era già presente in forma più arcaica a partire dal medioevo. All'epoca esistevano appunto delle bibbie, definite "poliglotte", che contenevano testi a fronte in ebraico, greco, latino e altre lingue. Lo stesso meccanismo è oggi applicato a livello informatico con i moderni corpora elettronici. Il NKRJa contiene al suo interno corpora paralleli in varie lingue, facilmente accessibili dal sito, come inglese, tedesco, spagnolo, bielorusso, polacco, ucraino, svedese e, tra queste, anche l'italiano. Ancora un esempio di corpus parallelo è il *Korpus Necoveršennyj Perevod* o *Russian Learner Translator Corpus*. Un'altra tipologia di corpora multilingui è quella dei corpora comparabili. Questi, al contrario dei corpora paralleli, non contengono testi in traduzione ma testi originali in lingue diverse. Con il loro utilizzo è possibile confrontare testi di due o più lingue diverse dello stesso genere o dominio. Trattandosi di testi originali, permettono di ricavare dati molto meno artificiali e talvolta anche privi di errori, come risultano spesso alcuni testi dopo il processo di traduzione. Si può considerare un corpus comparabile il progetto realizzato da A. Kutuzov dall'Università di Oslo, M. Kopotev

dall'Università di Helsinki, T. Sviridenko e L. Ivanova dall'Università NIU VŠE di Mosca nel 2016[8]. Il loro obiettivo fu quello di creare un corpus comparabile russo-ucraino raccogliendo testi di ambito accademico per paragonarli nelle due lingue.
L'ultima coppia è quella dei corpora sincronici e diacronici. I corpora sincronici contengono testi appartenenti ad un periodo di tempo limitato e il loro scopo è quello di ottenere un'istantanea di una determinata lingua o varietà linguistica in un determinato periodo per analizzarne le sue tipicità. I corpora diacronici invece, mirano a mostrare l'evoluzione di una lingua nel tempo, e per questo includono testi appartenenti ad un periodo di tempo molto ampio oppure di epoche lontane così da confrontare la lingua più antica con quella più moderna. Grazie a queste due tipologie di corpora è possibile svolgere analisi sincroniche e diacroniche di una lingua. Il NKRJa, così come ogni altro grande corpus moderno, può fungere sia da corpus sincronico, grazie al filtro degli anni di pubblicazione dei testi, che da corpus diacronico, per il fatto che contiene testi scritti a partire dal 1800, un periodo di tempo non lunghissimo ma che può comunque mostrare dei cambiamenti nell'uso della lingua. Di quest'ultima tipologia possiamo citare ancora una volta il *Regensburgskij diachroničeskij korpus russkogo jazyka.*

Un'altra tipologia di corpora che viene talvolta inclusa tra i corpora diacronici è quella dei corpora di monitoraggio, o *monitor corpora*, che sono una collezione aperta di testi che muta nel tempo in quanto in continuo aggiornamento grazie all'introduzione di nuovi testi. Questo approccio di monitoraggio sulla lingua fu per la prima volta proposto da Sinclair. Con corpora di questo tipo è possibile monitorare le dinamiche del lessico della lingua anche in archi di tempo molto brevi. C'è inoltre una tendenza sempre più diffusa da parte dei linguisti a riconoscere come un corpus di questa categoria il web-corpus. Un lato positivo dei corpora di monitoraggio, che è stato notato da alcuni linguisti come Gatto (2014), è il fatto che man mano che il corpus cresce, gli errori che conteneva, di cui i corpora non sono mai totalmente privi, si correggono in maniera spontanea.
C'è infine un'altra particolare tipologia di corpora, i *learner corpora*, che consistono in una raccolta di testi scritti o interazioni orali prodotte da apprendenti di una lingua straniera. Un esempio è il *Korpus Russkich Učebnych Tekstov* (KRUT) o *Corpus of Russian Student Texts* (CoRST), che include testi scritti da studenti di russo di diverse università che appartengono a diversi ambiti, economico, filosofico, linguistico, storico ecc. Esso può essere un punto di riferimento per diverse figure, da ricercatori a insegnanti ad altri studenti.

[8] Progetto intitolato Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints

Tutti quelli che abbiamo visto fino ad ora sono esempi di corpora che sono nati in quanto strumento "finalizzato alla ricerca linguistica" (Barbera, Corino, Onesti 2007: 46), come sostenne anche Sinclair. Esistono però anche corpora che sono basati su materiali linguistici ma che sono nati con altri scopi. Un esempio è uno di quei corpora che abbiamo in precedenza definito "preistorici", il *Corpus Iuris Civilis*, ovvero una "raccolta ufficiale dei testi normativi fondamentali, fatta eseguire dall'imperatore Giustiniano I"[9]. Vediamo quindi come i corpora non rientrino solamente nell'ambito della *corpus linguistics*, che è quello preso in analisi da questo lavoro, ma anche in altri ambiti.

1.7 Qualche aspetto tecnico dei corpora

1.7.1 La tokenizzazione

La tokenizzazione, come accennato in precedenza, è una delle operazioni che, assieme ad altre, è fondamentale eseguire affinché un corpus rientri in quella che è la sua più moderna definizione, come quella proposta da Barbera. Egli ci spiega infatti che la tokenizzazione è "il requisito minimo perché un insieme di testi si possa considerare un corpus" (Barbera 2013: 18). Per comprendere in cosa consista la tokenizzazione, è necessario prima definire due concetti solitamente visti in coppia, quelli di *token* e di *type*. Il *token* è l'unità minima in cui sono divisi i testi che compongono un corpus. È quindi l'insieme di tutte le "parole" di un testo. Dall'altro canto, il *type* è il "descrittore della classe di tutti i token identici" (Barbera, Corino, Onesti 2007: 35), ovvero le stesse "parole" che in un testo si ripetono. Una volta chiarito questo possiamo definire la tokenizzazione "l'operazione di individuazione dei token, ossia delle unità minime che il PC tratterà" (ibidem). Per fare questo è necessario eseguire una serie di operazioni che consistono solitamente nell'individuarli con uno spazio prima e uno dopo. La tokenizzazione è uno dei concetti cardine della linguistica computazionale, ovvero, in parole povere, la disciplina che si occupa di dare ai computer la capacità di elaborare il linguaggio naturale. La linguistica computazionale sta infatti alla base dei corpora elettronici, così come di servizi che fanno parte della nostra vita quotidiana quali Google Translate, i motori di ricerca e tutti quei servizi che sfruttano la sintesi vocale.

[9] https://www.simone.it/newdiz/newdiz.php?action=view&id=325&dizionario=2

1.7.2 Le annotazioni: metadata, *markup* e *tag*

Così come la tokenizzazione, anche le annotazioni sono elementi ormai ritenuti fondamentali nel moderno concetto di corpora elettronici. Viene ancora fatta la distinzione tra corpora annotati, che sono oggigiorno i più utili e utilizzati, e non annotati, utilizzati più che altro nel passato. I corpora annotati contengono informazioni linguistiche che sono già presenti nei testi non annotati ma in maniera implicita. Ad esempio, il fatto che il verbo "vado" sia la prima persona singolare, tempo presente del verbo andare sarà noto a qualunque parlante di lingua italiana che leggerà un testo che contiene questo verbo (meno scontato sarebbe invece per un apprendente di lingua italiana, per il quale può risultare un'informazione utile). Con le annotazioni, queste informazioni vengono semplicemente esplicitate, rendendo più facile ottenere ed analizzare le informazioni linguistiche contenute nei corpora.

Ci sono diversi modi in cui le annotazioni possono presentarsi in un testo: possono essere incluse nei testi, modificandoli, oppure archiviate separatamente ma collegate ai testi con la possibilità di consultarle. Come ci spiegano McEnery e Hardie (2012: 13) le annotazioni separate sono generalmente preferite dai linguisti perché non vanno a modificare il testo ma lo lasciano invece intatto.

Le tre tipologie di annotazioni che si possono generalmente trovare in un corpus elettronico sono: metadata, *markup* e *tag*[10]. I metadata sono informazioni che ci dicono qualcosa sul testo stesso, ad esempio l'autore, la data di pubblicazione, la lingua in cui è scritto ecc. Possono inoltre identificare i parlanti in un testo e darne informazioni quali l'età e il sesso. Si tratta in questo caso di informazioni extra-testuali. I *markup* forniscono invece informazioni di carattere testuale, come indicare la formattazione del testo, definire se si tratti di un testo in poesia o in prosa oppure dove un determinato parlante inizia e finisce di parlare. Abbiamo infine i *tag*, che forniscono informazioni di carattere linguistico alle diverse porzioni di testo. Questi sono forse i più utili per effettuare ricerche di carattere linguistico. I *tag*, anche chiamati annotazioni morfosintattiche, consistono solitamente in sigle che hanno lo scopo di individuare, in ogni frase, le parti del discorso, quindi le classi di parole. Indicano se di tratta di sostantivi singolari o plurali, aggettivi maschili o femminili, o ancora forme verbali presenti o passate, e così via. Ecco perché ci si riferisce spesso ai *tag* con la denominazione *part-of-speech tagging* o *POS-tagging*. I creatori di ogni corpus scelgono un *tagset* da adottare per tutto il corpus, in modo che, ad esempio, tutti i nomi

[10] Terminologia presa da Barbera M., Corino E., Onesti C. (2007), che distingue in maniera precisa le tre categorie di annotazioni.

propri siano etichettati come *np* e i nomi comuni come *nc*, la forma plurale indicata con la sigla *pl* e quelli singolare con *sg*.

I *tag* si trovano solitamente tra le parentesi angolari (<>) e stanno sia all'inizio che alla fine dell'unità linguistica a cui designano una classe. Un esempio di testo su cui è stato eseguito il *POS-tagging* può essere questo:

<Agg.poss>mio</Agg.poss><N>fratello</N>

in cui *Agg.poss* sta per *aggettivo possessivo* e *N* per *nome*.

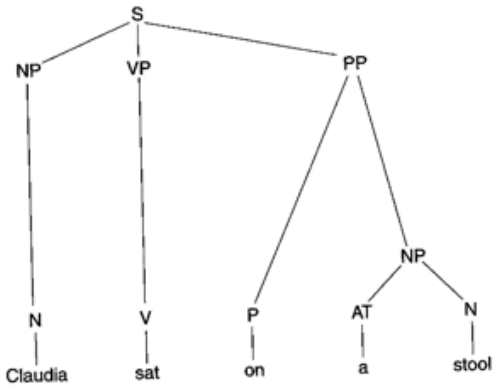
Mentre in passato l'etichettatura veniva eseguita manualmente, oggi è una procedura per lo più automatizzata che avviene tramite l'utilizzo di software appositi chiamati *tagger*. Questi programmi, però, non sono alla portata di tutti. Essendo molto complessi e richiedendo competenze informatiche avanzate non vengono utilizzati dai linguisti che creano i corpora ma saranno gli ingegneri informatici ad eseguire la procedura di etichettatura dei testi. Nonostante sia una procedura automatica, non mancano gli errori che possono essere commessi dai software. Per questo è talvolta previsto l'intervento dei linguisti che eseguono una correzione manuale.

Oltre al *POS-tagging* esistono altre forme di *tag*, che vedremo qui di seguito, come la lemmatizzazione, che individua i lemmi, e il parsing, che individua la categoria semantica.

1.7.3 La lemmatizzazione e il *parsing*

La lemmatizzazione è "l'operazione di ricondurre ogni *type* al proprio lemma, per cui canta, canteremo, canterò sono marcati tutti come *type* del lemma cantare" (Barbera 2013: 27). Il lemma è la forma delle parole che solitamente si trovano nei dizionari, la loro forma "neutra", non declinata in alcun modo. Grazie alla lemmatizzazione è possibile ricercare in un corpus un lessema e trovare anche tutte le sue varianti. Questo può essere utile quando si vogliono effettuare delle particolari ricerche linguistiche, come ad esempio quelle che riguardano la frequenza.
Abbiamo poi il *parsing*, ovvero il processo di analisi della struttura sintattica della frase. È una tipologia di annotazione che indica i costituenti di una frase e permette quindi di analizzarla sintatticamente. Per fare questo vengono assegnati alle frasi dei marcatori. Come ci spiega Barbera (2013), il *parsing* è un'operazione particolarmente diffusa nei corpora di lingua inglese e meno in quelli di altre lingue.

Questi sono due dei molteplici modi in cui può apparire un testo su cui è stato eseguito il *parsing* (Fig. 1):




```
[S
    [NP Claudia NP]
    [VP sat
        [PP on
            [NP a stool NP]
        PP]
    VP]
S]
```

Figura 1. immagini tratte da McEnery, Wilson 2001: 54-55

È anche interessante notare che in alcune particolari tipologie di corpora è possibile trovare *tag* speciali, come ad esempio l'*error tag,* che nei corpora di apprendenti indicano gli errori commessi dagli studenti di una lingua.

1.7.4 Le concordanze

Quella delle concordanze è una delle ricerche più tipiche che possono essere condotte tramite l'utilizzo di un corpus. Le concordanze sono liste di parole che vengono mostrate nel loro contesto, in cui per contesto si intende la porzione di testo, di misura diversa a seconda della volontà dell'utente, che sta prima e dopo la parola cercata. Questa funzionalità viene anche spesso chiamata KWIC (*Key Word In Context*), che permette la visione verticale delle parole nel loro contesto. Le concordanze possono essere un aiuto importante per i linguisti ma non solo, sono preziose anche per gli apprendenti di una lingua e per tutti coloro che desiderano effettuare un'analisi quantitativa, conoscere i diversi significati con cui una parola può essere utilizzata e le preposizioni, aggettivi, verbi che solitamente precedono o seguono una determinata parola. Mostrano anche la tendenza di alcuni elementi lessicali a co-occorrere, indicando le strutture tipiche e frequenti del linguaggio. Può essere inoltre utile a scoprire espressioni polirematiche, come ad esempio "alzare il gomito", "tagliare la corda" e molto altro. Per fare ciò esistono oggigiorno software dedicati detti *concordancer,* che permettono all'utente non solo di ricercare una parola da mostrare nel suo

contesto ma anche di scegliere la lunghezza del contesto, ovvero il numero di parole che si desidera vedere prima e dopo la parola cercata.

Essendo i *concordancer* uno strumento tanto prezioso, è anche interessante capire come si sono sviluppati fino a diventare quelli che sono i moderni strumenti informatici accessibili a chiunque. La storia dei *concordancer*, che ci è stata raccontata da McEnery e Hardie (2012) in maniera molto dettagliata, la andremo a vedere qui più in breve. I primi *concordancer*, che potremmo definire "preistorici", risalgono al 13° secolo. Il cardinale Hugh di St Cher, con l'aiuto di circa 500 monaci, compilò la prima *concordance* della bibbia in latino volgare nel 1230, fornendo, per ogni parola, un indice di dove ognuna di esse poteva essere trovata. Non fu quindi Roberto Busa, colui che costruì il primo corpus in formato elettronico nel 1951, ad inventare il concetto di concordanza, ma fu lui ad applicarlo ai testi elettronici. Fu in grado di trasformare questo lavoro certosino, da svolgere solamente su testi di particolare importanza culturale, in una procedura che può essere applicata con facilità a qualsiasi tipologia di testo. Con Busa nacque quindi la prima generazione moderna di *concordancer*, che non fu però l'unica. Questi si trovavano sui primi calcolatori e ciò che potevano fare era semplicemente fornire delle KWIC, ovvero mostrare semplicemente le parole nel loro contesto. Si incontravano però parecchie difficoltà con i caratteri che non fossero i semplici caratteri romani non accentati. Alcuni dei problemi riscontrati nell'utilizzo di questi *concordancer* vennero risolti negli anni Ottanta con l'avvento dei personal computer. Si arriva quindi ai *concordancer* di seconda generazione. Mentre i primi erano accessibili solamente ad un pubblico ristretto e specializzato, ovvero a team di informatici, questi si diffusero assieme ai computer e divennero accessibili ad un pubblico più ampio. Per utilizzarli bastava solamente essere in grado di accendere un computer. Mentre alcuni errori furono corretti rispetto alla prima generazione, ne subentrarono di nuovi a causa della potenza ridotta dei personal computer rispetto ai calcolatori.
Con l'aumento di PC sempre più potenti, si passò poi alla terza generazione di *concordancer*, alcuni dei quali sono i tuttora conosciuti WordSmith (1996), MonoConc (2000), AntConc (2005). Questi software includevano un'ampia gamma di strumenti che permettevano di svolgere analisi statistiche importanti, come ad esempio le liste di frequenza, oltre che ad avere una mera funzione descrittiva. Supportavano inoltre una serie di sistemi di scrittura grazie allo standard XML e al sistema di caratteri standardizzato Unicode, recentemente sviluppatosi, risolvendo gran parte dei problemi legati ai caratteri.
L'ultima generazione di *concordancer*, quella attuale, è molto simile alla precedente nelle sue funzionalità, ma con alcuni miglioramenti tecnici. Con l'ampliarsi delle possibilità di diffusione dei corpora sorge però un altro problema, quello del copyright. I corpora infatti non potevano essere accessibili a tutti i ricercatori che avrebbero voluto utilizzarli perché questo avrebbe violato le leggi

di copyright nei confronti degli autori dei testi contenuti nel corpus. Questo problema fu risolto grazie all'interfaccia dei moderni corpora, la quale, dopo una ricerca, mostra all'utente solamente una stringa di testo, e non il testo intero, in modo che i limiti imposti dalle leggi dei copyright non fossero superati. La caratteristica principale dei corpora di quarta generazione è quella di essere accessibili non come software su desktop ma attraverso la rete, come ormai quasi tutti i moderni corpora.

1.7.5 Le collocazioni

Come già anticipato, le collocazioni sono un concetto che fu introdotto da Firth e quindi dall'approccio *neo-Firthian*. Anche questa nozione, assieme a quella di concordanza, è centrale nella *corpus linguistics*. Con le parole di Gatto, descriviamo le collocazioni come "the tendency of words to co-occure. More specifically, the collocates of a word are words that frequently occur in its immediate co-text" (Gatto 2014: 26). Da qui capiamo come spesso il significato di una parola non sta nella parola in sé, presa singolarmente, ma si trova invece nelle tipiche associazioni di questa con altre parole o strutture linguistiche. Per identificare le collocazioni, i corpora si avvalgono di una formula, il *MI-score*. Sempre Gatto ci spiega che più questo valore è alto, più è solido il legame tra due parole. Se il valore del *MI-score* è 3 o più, si tratta di collocazione. Per parlare di collocazione però, è necessario stabilire anche un altro valore, in questo caso si tratta del *t-score*. Mentre il *MI-score* misura la forza della collocazione, il *t-score* misura la certezza con cui si può affermare che c'è un'associazione tra due parole, tenendo conto della frequenza di co-occorrenza della collocazione.

Un altro concetto legato a quello di collocazione è la colliganza, in inglese *colligation*, ovvero "the occurence of a grammatical class or structural pattern with another one, or with a word or phrase" (Sinclair 2003: 173). Queste relazioni tra parole sono più difficili da osservare rispetto alle precedenti. Un primo sguardo potrebbe essere infatti non sufficiente ad individuarle, essendo quindi necessaria un'analisi delle stringhe di concordanze per poterle osservare. Un esempio di colliganza è la tendenza di un lemma a trovarsi in forma passiva. Ecco perché possono essere fondamentali per capire il comportamento di una parola. Grazie alle colliganze è anche possibile stabilire il campo semantico in cui una parola è solita ricorrere, un fenomeno, che rientra tra quello delle colliganze, chiamato "*semantic preference*".

Ribadiamo, in conclusione, l'importanza del concetto delle concordanze, uno strumento fondamentale della linguistica dei corpora grazie al quale è possibile compiere ulteriori step nella

loro analisi e ricavare altri elementi rilevanti nella ricerca linguistica quali le collocazioni e le colliganze.

1.8 Ambiti di applicazione della *corpus linguistics*

Per concludere il capitolo vediamo in sintesi un quadro degli ambiti in cui la linguistica dei corpora può essere e viene applicata. Ciò che stupisce è il fatto che nonostante si tratti di una disciplina linguistica, essa può essere applicata non solo in ambito linguistico ma anche in molti altri. Solitamente, anzi, un approccio metodologico pluralista viene favorito così da incrementare i dati ottenuti grazie al metodo della *corpus linguistics* ad altri dati, il che permette di ampliare gli orizzonti della ricerca in molte aree. La metodologia della *corpus linguistics*, come abbiamo visto, ha iniziato ad espandersi a partire dagli anni Ottanta, epoca in cui il concetto di corpus si è fuso con l'informatica e sono nati i primi corpora elettronici moderni. Da qui in poi, il numero e le dimensioni dei corpora sono aumentati esponenzialmente e questa metodologia ha cominciato ad essere applicata, oltre che all'ambito linguistico, ad ambiti sempre più svariati.

Come ci spiegano McEnery e Hardie, i corpora si sono dimostrati utili in diverse aree della linguistica come la linguistica contrastiva, l'analisi del discorso, la semantica, la sociolinguistica e così via; è stata ed è tuttora un valido aiuto per i lessicografi e i grammatici. Ma ci sono anche molte discipline umanistiche e scienze sociali che, pur non avendo a che fare direttamente con la linguistica, si occupano però dello studio dei testi. Ne sono un esempio la letteratura, la religione, la storia e anche la psicologia. Grazie ai suoi metodi di ricerca, la *corpus linguistics* si è rivelata preziosa anche in questi ambiti. Ovviamente, sulla base degli aspetti tecnici che abbiamo analizzato in precedenza, questo è possibili solamente dopo aver trasformato i testi su cui si vogliono svolgere delle ricerche in formato elettronico.
Per quanto riguarda gli ambiti culturali, i testi appartenenti ad essi contengono elementi culturali dei parlanti che grazie alla linguistica dei corpora possono essere analizzati in modo da studiare le diverse culture, le loro caratteristiche e ciò che le differenzia. Analogamente, la psicologia sociale, che studia l'interazione tra individuo e gruppi sociali, può servirsi dei corpora per effettuare diversi studi, come capire in che modo gli esseri umani si rapportano con l'ambiente circostante, ad esempio osservando come le persone utilizzano il linguaggio per spiegare le cose. Per fare ciò possono essere utili corpora che raccolgono quotidiani, diari, e così via. Sono poi utili i corpora di

lingua parlata per analizzare gli scambi di conversazioni quotidiane tra le persone (McEnery, Wilson 2001: 129). Questi sono solo alcuni dei molti esempi dei modi in cui i corpora possono essere utilizzati in discipline non linguistiche. Le aree di studio in cui la lingua è invece oggetto centrale di studio sono altrettante, tant'è vero che i corpora hanno rivoluzionato praticamente ogni branca della linguistica. Con l'utilizzo dei corpora è possibile svolgere studi sulla lingua parlata in quanto ne forniscono esempi concreti. La lingua in essi contenuta è infatti naturale, spontanea, e non artificiale o sotto monitoraggio come lo sarebbe una lingua prodotta appositamente per essere usata come oggetto di studio. Molti sono anche gli studi sul lessico che è possibile condurre grazie alla linguistica dei corpora. I lessicografi, a questo scopo, fanno uso di dati oggettivi come quelli ricavabili dai corpora da ancor prima che questi fossero inventati. Oggigiorno, tuttavia, recuperare questo tipo di dati diventa sempre più facile e rapido grazie all'uso dei corpora elettronici e dei computer. Inoltre, rispetto al passato, la mole di materiale a disposizione dei lessicografi grazie a questi nuovi strumenti è molto più grande. L'utilizzo dei corpora è molto frequente anche negli studi di grammatica, uno dei campi in cui più vengono utilizzati. I corpora sono infatti rappresentativi della grammatica di una varietà linguistica e sono fonte, come abbiamo visto, di dati empirici, quantificabili e rappresentativi per testare ipotesi che derivano da teorie grammaticali. Per quanto riguarda la semantica, i corpora hanno la caratteristica di mostrare il contesto in cui le parole compaiono, determinando quindi l'area di utilizzo di specifici elementi linguistici ed essendone un indicatore oggettivo. I corpora possono inoltre essere fonte di discorsi prodotti dai parlanti, che permettono di effettuare studi di pragmatica e analisi del discorso. Abbiamo poi la linguistica storica, che si occupa di studiare le cosiddette lingue "morte" o varietà antiche di una lingua come il latino, lo slavo ecclesiastico, il *Middle English* e l'O*ld English*, grazie ai corpora che raccolgono testi in queste lingue. In questo modo è possibile anche studiare l'evoluzione di una lingua, così come i dialetti, in quanto i corpora sono fonte di diverse varietà linguistiche, sia geografiche che temporali. Ed infine vediamo l'ambito dell'insegnamento delle lingue e della linguistica. Alcuni libri di testo impiegati nell'insegnamento delle lingue si basano su esempi inventati di uso della lingua. Altri, invece, si affidano a testi prodotti naturalmente ottenuti dai corpora, che sono fonte di esempi reali di utilizzo del linguaggio. Questi ultimi si sono dimostrati molto più efficaci e veritieri, contenendo esempi di lingua prodotta direttamente dai suoi parlanti nativi. È infatti più probabile che uno studente, durante il suo percorso di apprendimento, incappi in questo tipo di lingua, rispetto ad una lingua più artificiale e semplificata, nel momento in cui si trova a confrontarsi con testi in lingua originale o situazioni comunicative reali. Grazie ai *learner corpora* è poi possibile vedere quali sono i problemi più tipici che gli apprendenti incontrano durante lo studio di una lingua. I corpora, in questo caso di tipo specializzato, sono anche utili per l'insegnamento delle lingue

speciali, come quelle di ambito scientifico o giuridico. In questa tipologia di corpora, infatti, rispetto a quelli generici, è contenuto più materiale relativo a questi ambiti specifici. Ultimo, ma non per importanza, è l'ambito della traduzione, in cui i corpora sono sempre più utilizzati grazie soprattutto all'aumento dei corpora multilingui paralleli. Con questi il traduttore può sia vedere come certe parole, espressioni, frasi, vengono tradotto ma anche confrontare le proprie traduzioni con quelle già esistenti per affinare il proprio stile traduttivo.

Si conclude così questo primo capitolo introduttivo sui corpora e la *corpus linguistics*, lasciando spazio ad un capitolo più specifico sui corpora di lingua russa, che potremo analizzare sulla base delle nozioni teoriche qui apprese.

CAPITOLO 2
I CORPORA RUSSI

Introduzione

Dopo aver chiarito il concetto di corpus, averne esaminato le caratteristiche e le diverse tipologie, in questo secondo capitolo andremo nello specifico a vedere quali sono i corpora di cui la lingua russa dispone come strumenti di analisi linguistica sia per utenti russi che stranieri. Il primo da cui partiremo sarà, indubbiamente, il corpus russo per eccellenza, ovvero il *Nacional'nyj Korpus Russkovo Jazyka* (NKRJa). Lo analizzeremo a fondo nelle sue caratteristiche ma anche nell'ampia serie di sub-corpora e corpora paralleli che esso offre ai suoi utenti. Ma non parleremo soltanto dei corpora più tradizionali come è appunto il NKRJa. Daremo anche un ampio sguardo ai corpora più nuovi e sperimentali, come il *General'nyj Internet-Korpus Russkovo Jazyka* (GIKRJa), tuttora in fase di sviluppo, e assieme a questo altri mega-corpora la cui caratteristica è quella di essere basati su materiale presente sul Web. Vedremo poi anche una serie di corpora russi minori e più specializzati, per arrivare poi a completare il cerchio con alcuni strumenti che non sono dei veri e proprio corpora ma che si basano su materiale ricavato da essi per svolgere analisi specifiche.

2.1 Il *Nacional'nyj Korpus Russkovo Jazyka*

2.1.1 La storia del corpus

Come abbiamo visto nel capitolo precedente, nel corso dell'excursus storico, i primi corpora moderni, ovvero quelli elettronici, nascono a partire dagli anni Sessanta del Novecento negli Stati Uniti con la realizzazione del *Brown Corpus*. Nei decenni successivi, i corpora elettronici si svilupperanno fino a coprire, alcune prima e altre dopo, un gran numero di lingue, tra cui anche l'italiano. Ci spiega Noseda (2017: 21) che per la nostra lingua si considerano gli anni Novanta come l'inizio ufficiale della diffusione dei corpora elettronici, che avvenne per la prima volta con il PIXI Corpus, un corpus di lingua parlata in italiano e in inglese. Lo stesso vale per l'Inghilterra, in cui la creazione del British National Corpus avviene nel 1991. Per quanto riguarda invece la Russia,

la cosiddetta *corpus revolution* si concretizza solamente con il nuovo millennio grazie all'apertura del *Nacional'nyj Korpus Russkovo Jazyka*. La linguistica computazionale russa, infatti, era una delle più arretrate rispetto al resto del mondo dal momento che, come ci dice Šarov (2003:2), padre del progetto del Corpus Nazionale Russo, fino a quel momento il russo era una delle poche lingue principali per cui mancava un corpus rappresentativo della lingua moderna. È solamente negli ultimi 15 anni, con lo sviluppo e ampliamento del Corpus Nazionale, che la Russia ha raggiunto gli altri paesi recuperando alcune tappe importanti nella disciplina: sempre più dizionari vengono compilati sulla base del NKRJa dando vita alla *korpusnaja grammatika*, che ha lo scopo di descrivere parti della grammatica russa in base ai dati tratti dal corpus. Alla fine di questo percorso, sebbene il *Nacional'nyj Korpus* rimanga lo strumento preferito dai russisti per svolgere ricerche *corpus-based*, nuovi web-corpora sono comparsi nel panorama nazionale russo, come il GIKRJa, l'Araneum Russicum, il ruTenTen, il ruWac e molti altri che vedremo nel corso del capitolo.

Tornando alla storia del NKRJa, prima di arrivare a quello che esso è oggi ci furono tre tentativi di realizzare un corpus che fosse rappresentativo della lingua russa (Šarov 2003:1). Il primo fu quello da parte di Zasorina, negli anni Settanta. Il suo intento era quello di creare un corpus di un milione di parole che includesse quattro generi testuali, ovvero informazione, narrativa, scienza e dramma. Quello che ne risultò fu però un semplice dizionario di frequenza, il *Častotnyj slovar' russkogo jazyka* del 1977, tra le altre cose non accessibile al pubblico online.

Il secondo tentativo fu l'Uppsala Corpus, nato in Svezia negli anni Ottanta, che consisteva in una raccolta di un milione di testi tra narrativa e altri generi testuali. Tra i predecessori del NKRJa, questo fu il più conosciuto grazie anche alla sua accessibilità tramite il web. Nonostante ciò, rimaneva comunque un corpus troppo piccolo rispetto agli standard dell'epoca e inoltre mancava di annotazione morfosintattica e lemmatizzazione. Anche in questo caso il risultato fu un dizionario di frequenza.

Il terzo e ultimo tentativo avvenne negli anni Ottanta in Unione Sovietica. Si tratta del *Computer Fund of Russian Language* (CFRL), il cui obiettivo era quello che fu poi raggiunto pochi anni dopo dai creatori del *British National Corpus*, ovvero creare un grande corpus nazionale con inclusi sub-corpora di vario genere. Il progetto però fallì e si dovette attendere fino all'inizio del Duemila per un nuovo progetto, che sarebbe diventato poi quello definitivo.

Il vero e proprio antenato del NKRJa fu il *Bol'šoj Korpus Russkogo Jazyka* (BOKR), noto in inglese come *Russian Reference Corpus*, ovvero il progetto, presentato nel 2003 da Sergej Šarov (Šarov 2003), di un corpus che fosse l'equivalente russo del *British National Corpus*. Questo, che

può essere considerato la versione pilota dell'odierno NKRJa, voleva essere un corpus di 100 milioni di parole a libero accesso tramite il web, che coprisse le principali tipologie di testo della Russia contemporanea e che fosse dotato di annotazione morfosintattica e lemmatizzazione. Secondo il progetto, esso doveva includere un sub-corpus di 10 milioni di parole, chiamato *Russian Standard*, contenente al suo interno testi di narrativa moderna che fossero rappresentativi della lingua letteraria russa standard. Questa sezione voleva essere una fonte di informazione per lo sviluppo di grammatiche russe *corpus-based* a scopo accademico e di insegnamento. Il corpus generale, invece, sarebbe stato una fonte complementare di dati grammaticali e una fonte principale di dati lessicali. Al contrario del BNC, in cui il genere letterario copriva solamente una parte del quadro generale dei generi testuali, il BOKR prevedeva che quello letterario fosse il genere prediletto. Da qui vediamo come la lingua letteraria abbia in Russia uno status culturale superiore, essendo considerata la fonte autorevole che va a definire la lingua utilizzata dai suoi parlanti. Al di là di questo, il numero più ridotto di testi di genere diversi dalla *fiction* e anche dall'informazione era anche dovuto al fatto che all'epoca la loro reperibilità in formato elettronico era molto più limitata.
Il progetto prevedeva come data di rilascio della versione finale del corpus il 2004, anno in cui il NKRJa fu poi effettivamente aperto alla consultazione, rispettando quindi l'obiettivo prefissato.

Il corpus è stato realizzato presso l'Accademia Russa delle Scienze di Mosca. Alla sua creazione ha partecipato un grande numero di linguisti provenienti da diverse parti della Russia come Mosca, San Pietroburgo, Kazan, Voronež, Saratov e altri centri scientifici del paese. A sostegno del progetto vediamo in prima linea il più importante motore di ricerca russo Yandex, che fornì il suo supporto tecnico per la programmazione del corpus, ma anche vari istituti nell'ambito della ricerca linguistica, filologica, umanistica, che hanno contribuito alla realizzazione del corpus con fondi finanziari.

2.1.2 Le caratteristiche

La data ufficiale di apertura del NKRJa alla consultazione fu il 29 aprile del 2004. Il corpus, rappresentativo della lingua russa moderna, contiene attualmente (gennaio 2020) 700 milioni di parole, arrivando a superare di gran lunga quello che era il suo modello iniziale, il BNC[11]. Gli scopi principali che esso si propone sono quelli di facilitare la ricerca accademica nell'ambito del lessico,

[11] Contiene oggi 100 milioni di parole

della grammatica e più in generale, negli studi sul processo di cambiamento della lingua russa. Inoltre vuole essere un punto di riferimento per la soluzione di problemi lessicali, grammaticali, accentologici e sulla storia della lingua. Gli utenti a cui si rivolge rientrano sotto diversi profili, a partire dal linguista ma non solo: il corpus può risultare utile anche a figure legate alla letteratura, alla storia e ad altre discipline umanistiche per svolgere ricerche di diverso tipo; è uno strumento ormai fondamentale anche e soprattutto per apprendenti della lingua russa e insegnanti sia russofoni che stranieri. Inoltre, sempre nell'ambito accademico, il corpus viene utilizzato nella creazione di testi scolastici e per la strutturazione di programmi di insegnamento linguistico.

I testi contenuti nel NKRJa coprono un arco temporale compreso tra il XVIII e il XXI secolo (ad eccezione del sub-corpus storico che include testi precedenti che vanno dal XI al XVIII secolo). Essendo i suoi punti cardine la rappresentatività e il bilanciamento, esso possiede, nelle sue varie sezioni, tutte le tipologie di testi scritti e orali presenti nella lingua. Ci sono testi di narrativa, che vanno dalla prosa, al dramma e alla poesia, ma non solo: abbiamo anche testi giornalistici, accademici, economici, di divulgazione scientifica, memorie, saggi, lettere e diari personali, per poi passare a testi in lingua colloquiale e anche dialettale. Il corpus si dimostra così uno strumento completo, rappresentativo della lingua russa a 360°.

Il corpus, lo abbiamo già anticipato, è ovviamente dotato di annotazioni, come ormai ogni corpus elettronico che si rispetti. Nel nostro caso abbiamo cinque/sei tipologie di annotazione, che sono: meta-testuale, morfologica e semantica applicate a tutte le sezioni del corpus, e sintattica, accentuale e poetica applicate solamente ad alcuni sub-corpora che vedremo in seguito.

L'annotazione meta-testuale comprende informazioni testuali come l'autore, il genere, l'anno di produzione e così via. Questo permette, selezionando i vari parametri, di effettuare ricerche più approfondite e in ambiti diversi da quello puramente linguistico. È grazie a questa funzione, così come ad altre, che lo strumento del corpus può essere utile per svolgere ricerche in ambiti umanistici o sociologici, come può essere ad esempio confrontare le caratteristiche della scrittura di autrici donne rispetto agli uomini, ricercando elementi linguistici in testi scritti specificatamente dagli uni o dagli altri.

L'annotazione morfologica comprende invece le informazioni linguistiche relative ad ogni singolo lessema (appartenenza alla classe, genere, transitività o intransitività per i verbi, aspetto ecc.) e ad ogni forma di parola specifica (caso, numero, tempo, persona in cui occorre) (Noseda 2017: 40). È basata su un *tagset* in lingua inglese pensato anche e soprattutto per utenti stranieri ma è presente anche un *tagset* russo per utenti esclusivamente russofoni. Per quanto riguarda l'annotazione semantica, essa aggiunge informazioni semantiche ai lemmi del corpus. Ad ogni parola sono quindi

abbinati uno o più tratti semantici a seconda dei suoi diversi utilizzi. Passando invece alle tipologie di annotazione tipiche dei singoli sub-corpora, sempre affiancate alle tre tipologie appena viste, abbiamo quella sintattica, applicata solamente al corpus sintattico, quella accentuale, applicata al corpus della storia dell'accento russo, e quella poetica, applicata al corpus di testi poetici, tutti sub-corpora che analizzeremo singolarmente a breve.
L'annotazione in un corpus è utile non solo ad ottenere informazioni aggiuntive sull'oggetto di ricerca, ma anche, come permettono il NKRJa e altri grandi corpora nazionali, a creare sub-corpora *ad hoc* ed affinare la ricerca su un oggetto specifico.

Mi sembra poi interessante segnalare alcune altre caratteristiche tipiche dei corpora nazionali, presenti anche in quello russo, che possono essere di grande utilità e fare la differenza rispetto a corpora in cui queste funzionalità sono assenti. Una di queste è la possibilità di visualizzare i risultati della ricerca in formato KWIC (*Key Word In Context*), fondamentale, come abbiamo visto nel primo capitolo, nell'ambito delle concordanze. Abbiamo poi il processo di disambiguazione dell'omonimia grammaticale, utile nel caso in cui due parole uguali abbiano significati diversi in diversi contesti, che è stata effettuata attualmente solo in alcune sezioni del corpus (corpus educativo, sintattico e poetico), essendo un processo lungo e molto spesso svolto manualmente. Infine è da sottolineare la possibilità data all'utente di segnalare errori, qualora venissero riscontrati, cliccando sulla voce *soobscit' ob oščibke* ("segnalare un errore"), in modo da contribuire al miglioramento delle funzionalità del corpus.

2.1.3 I sub-corpora

Il NKRJa suddiviso in dodici sezioni, o meglio sub-corpora, ognuno dei quali contiene diverse tipologie di testi e che permettono di svolgere specifiche ricerche linguistiche e non. Questi sono: *Osnovoj korpus, Sintaktičeskij korpus, Gazetnyj korpus, Parallelnye korpusa, Korpus dialektnych tekstov, Korpus poetičeskich tekstov, Obučajuščij korpus, Korpous ustnoj reči, Akcentologičeskij korpus, Mul'timedijnyj korpus, Istoriceskij korpus, Multimedijnyj parallel'nyj korpus Mul'tipark*.
Nel momento in cui l'utente si appresta a svolgere una ricerca, potrà selezionare il corpus con il quale preferisce effettuare tale ricerca.
Per comprendere meglio le tipologie di testi che essi contengono e le loro peculiari caratteristiche, spenderemo qualche parola per ognuno di loro. Dal momento che ogni sub-corpus, se preso singolarmente, ha tutte le caratteristiche per essere un corpus a sé, per praticità li chiameremo nel corso di questa parte di lavoro semplicemente "corpus".

Osnovoj korpus - Corpus principale

Questo è il corpus più voluminoso tra tutti quelli contenuti all'interno del NKRJa (289 milioni di parole)[12]. I testi in esso raccolti, scritti tra il XVIII e il XXI secolo, sono testi in prosa che rappresentano la lingua russa standard. La quantità di questi testi, che sono di genere letterario, è proporzionata alla loro presenza nella vita quotidiana in Russia, ovvero il 40%, contro il 60% di testi di altre tipologie

Questo sub-corpus contiene testi di tre tipologie. Testi scritti contemporanei, pubblicati tra la metà del XX secolo e l'inizio del XXI, rappresentativi della lingua letteraria russa contemporanea. I generi testuali a cui appartengono sono molto vari: narrativa di vario genere, dramma, memorie e biografie, testi giornalistici e critiche letterarie, testi di divulgazione scientifica ed educativi, religiosi e filosofici, tecnici, pubblicitari, giuridico-amministrativi, testi privati non destinati alla pubblicazione come diari e lettere personali.

Ci sono poi testi di lingua russa contemporanea parlata e spontanea.

Ed infine abbiamo testi precedenti che vanno dalla metà del XVIII secolo alla metà del XX, appartenenti sempre a vari generi testuali. La quantità di testi di questa tipologia è inferiore rispetto a quelli della prima a causa della limitata reperibilità in formato elettronico di questi testi meno recenti. Viene inoltre sottolineato dai creatori stessi del corpus che i testi precedenti al 1918. sono scritti con la moderna ortografia e le peculiarità dell'ortografia originale, conservate nelle edizioni accademiche, sono conservate anche nel corpus.

Sintaktičeskij korpus - Corpus sintattico

Chiamato anche *Gluboko-annotirovannyj korpus* (1 milione di parole), questo corpus "presenta informazioni sulle relazioni sintattiche instaurate dalla parola ricercata con le altre unità linguistiche della frase" (Noseda 2017:41). La sua peculiarità è quella di essere dotato, oltre che di annotazione meta-testuale, morfosintattica e semantica, anche di annotazione sintattica. Questo tipo di annotazione consiste in grafici con struttura ad albero che mostrano la struttura sintattica della frase, come in questa immagine (fig. 2):

[12] Cifre che fanno riferimento a gennaio 2020

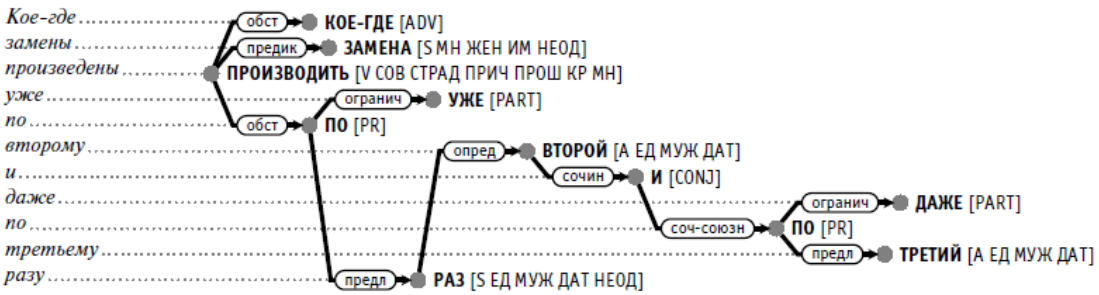


Figura 2. Risultato della ricerca della parola разу all'interno del corpus sintattico del NKRJa.

Gazetnyj korpus – Corpus giornalistico

Il corpus giornalistico fu aperto nel 2010 e contiene documenti tratti dai mezzi di comunicazione di massa tra gli anni Novanta e Duemila. Le sue dimensioni sono notevoli (240 milioni di parole), tanto che non si scostano di molto da quelle del corpus principale. I testi, che vengono raccolti da sette periodici russi sia cartacei (*Izvestija*, *Sovetskij Sport*, *Trud*, *Komsomol'skaja Pravda*) che online (*RIA Novosti*, *RBK*, *Novyj Region*), sono aggiornati annualmente.

Il corpus include al suo interno un'ulteriore sezione, il *Korpus regional'noj i zarubežnoj pressy*, molto più piccola rispetto al corpus totale (14 milioni di parole, comprese nella cifra totale). Questa sezione fu aperta qualche anno dopo, nel 2015, ed include testi tratti da giornali locali di diverso livello, regionali, provinciali e cittadini, provenienti da un ampio spazio geografico: oltre agli stati della Federazione Russa ci sono anche alcuni paesi della Comunità degli Stati Indipendenti (tra cui Bielorussia, Moldavia, Kirghizistan) e alcuni Paesi Baltici. L'arco temporale ricoperto da questi testi va dal 1996 al 2013. Al suo interno sono incluse quattro collezioni autonome: *Lingvističeskij illjustrativnyj korpus SMI Grodeninščiny*, che contiene testi del giornale della regione Grodeninščina in lingua russa e bielorussa; due collezioni di giornali regionali della Russia, una che va dal 1990 al 2000 e l'altra che comprende il primo decennio degli anni 2000; e infine la collezione di pubblicazioni regionali della *Komsomol'skaja Pravda*.

Parallelnye korpusa – Corpora paralleli

I corpora paralleli del NKRJa (98 milioni di parole totali) includono la traduzione da e verso il russo per un totale di diciannove lingue, che sono: armeno, baschiro, bielorusso, bulgaro, buriato, ceco, cinese, estone, finlandese, francese, inglese, italiano, lettone, lituano, polacco, spagnolo, svedese,

tedesco. Oltre a questi è presente anche un corpus parallelo multilingue (5 milioni di parole oltre al totale degli altri corpora paralleli) che permette di ricercare una parola in una delle 19 lingue e vedere contemporaneamente come questa sia stata tradotta in tutte le altre. Come avviene per la maggior parte dei corpora di questa tipologia, ogni unità di testo è allineata alla sua traduzione per quanto riguarda la coppia di lingue scelta. In questa parte ci limiteremo solamente a queste informazioni sui corpora paralleli del NKRJa in quanto ne verrà dedicata più avanti una sezione apposita, in cui verrà trattato nel dettaglio anche e soprattutto il corpus parallelo russo-italiano, recentemente ampliato.

Korpus dialektnych tekstov - Corpus dialettale

Questo piccolo corpus (285 mila parole), aperto nel 2005, include testi redatti nei dialetti delle diverse regioni della Federazione Russa. La sua peculiarità è quella di mettere in luce le caratteristiche morfologiche, sintattiche e lessicali dei vari dialetti. Per fare questo sono presenti delle speciali annotazioni per alcuni elementi caratteristici dialettali. Il corpus include inoltre un commentario dedito a spiegare il significato e l'utilizzo di lessemi puramente dialettali.

Korpus poetičeskich tekstov – Corpus poetico

Il corpus poetico (12 milioni di parole), comprende testi poetici che vanno dalla metà del XVIII secolo fino al XX secolo. Aperto nel 2006, permette la ricerca non solo di elementi grammaticali e lessicali ma anche di elementi tipicamente poetici, come la metrica e il ritmo. Per la segnalazione di questi elementi è presente una speciale annotazione, come sempre affiancata a quella meta-testuale, morfosintattica e semantica.

Obučajuščij korpus – Corpus educativo

Il corpus educativo è un piccolo corpus (664 mila parole), aperto nel 2006, orientato all'insegnamento della lingua russa nelle scuole. In esso sono infatti inclusi testi di narrativa attinenti ai programmi scolastici.

Korpous ustnoj reči - Corpus della lingua parlata

Questo corpus (12 milioni di parole) esiste come corpus indipendente dal 2007. Esso include le trascrizioni di discorsi spontanei sia pubblici che privati in lingua russa, pronunciati nell'arco di tempo che va dal Novecento al Duemila. Assieme a questi sono presenti anche le trascrizioni di film russi usciti tra il 1930 e il 2007. I testi sono di varie tipologie e con parlate tipiche delle diverse zone della Russia quali Mosca, San Pietroburgo, Saratov, Ulyanovsk, Taganrog, Ekaterinburg ecc.

Akcentologičeskij korpus – Corpus della storia dell'accento russo

Detto anche *Korpus istorii russkovo udarenija*, questo corpus (33 milioni di parole), aperto nel 2008, include testi graficamente accentati presi da alcuni degli altri sub-corpora, ovvero tutti i testi del corpus poetico e alcuni testi del corpus orale, film inclusi. L'utente avrà così la possibilità di svolgere ricerche in base alla posizione dell'accento e alla struttura prosodica delle parole. Per fare questo il corpus è dotato di speciali annotazioni contenenti informazioni sulla posizione dell'accento nelle parole.

Mul'timedijnyj korpus (MURKO) - Corpus multimediale

Il corpus multimediale (4,5 milioni di parole), che abbiamo già introdotto nel primo capitolo, è stato aperto nel 2010 e contiene la trascrizione di frammenti audio-video di film russi usciti tra il 1930 e il 2000. Ad ogni clip, di una durata che va dai 5 ai 20 secondi, è allineata la corrispondente trascrizione. Il materiale trascritto, essendo un vero e proprio testo, è dotato di annotazione standard come tutto il resto del corpus, ovvero quella meta-testuale, morfologica e semantica. Questo sub-corpus, però, è dotato altresì di due speciali tipologie di annotazioni. Si tratta di annotazione ortoepica per quanto riguarda il suono, e di annotazione gestuale per la parte video (Grišina 2010: 2). Questo permette di effettuare ricerche non solo sui testi ma anche sui gesti (accenno col capo, pacca sulla spalla ecc.) e sugli atti discorsivi (approvazione, ironia, ecc.).

Istoričeskij korpus – Corpus storico

Il corpus storico (14 milioni di parole totali) include testi in russo antico scritti in diverse epoche storiche. Per questo è diviso in quattro sezioni, ognuna dedicata ad un'epoca diversa, che sono: *Drevnerusskij korpus* (500 mila parole), *Berestjanye gramoty* (19 mila parole), *Starorusskij korpus* (8 milioni di parole), *Cerkovslavjanskij korpus* (4,5 milioni di parole) (Mitrenina 2014).

Il *Drevnerusskij korpus* contiene testi originali in russo antico scritti tra il VII e il XV secolo, di cui tutte le cronache e gli scritti di Kirill di Turov, e una serie di testi tradotti, incluso l'*Izbornik* (1076) in bulgaro antico. Di questi testi tradotti è possibile vedere anche l'originale in Greco; il *Berestjanye gramoty* contiene i documenti scritti su corteccia di betulla tra l'XI e il XV secolo. Esso include non solo testi ma anche immagini e frammenti non leggibili. Cliccando sul documento, infatti, all'utente sarà possibile vedere i dettagli del documento e con un link verrà rimandato alla fotografia della corteccia dalla quale è stato tratto; lo *Starorusskij korpus* include testi che sono stati scritti tra il XIV e l'inizio del XVIII secolo; lo *Cerkovslavjanskij korpus*, infine, sezione più voluminosa della quattro, contiene testi scritti tra il XVII e il XX secolo (Dobrušina, Poljakov 2013) e rappresenta l'*izvod* slavo-orientale in slavo ecclesiastico, nella sua forma arcaica, standard e moderna. La maggior parte dei testi in esso inclusi, che sono sia originali che tradotti, sono testi liturgici.
Con i corpora *Drevnerusskij, Cerkovslavjanskij, Berestjanye gramoty* è possibile svolgere ricerche che mostrano la distanza tra le parole o le forme grammaticali, mentre, per quanto riguarda lo *Starorusskij korpus* è possibile ricercare solamente parole o frasi specifiche. I primi tre corpora prevedono una tastiera virtuale che, al momento della ricerca, permette di inserire caratteri in cirillico antico. Il corpus storico del NKRJa è attualmente in espansione così da arrivare ad includere un numero maggiore di testi, data la sua dimensione ancora piuttosto ridotta. Inoltre, i linguisti sono al lavoro per lo sviluppo di una ricerca tramite il corpus anche dal punto di vista sintattico, che non è ancora stata implementata.

Multimedijnyj parallel'nyj korpus Mul'tipark

Sul sito del NKRJa è da poco comparso un'ulteriore sezione, ovvero il corpus multimediale parallelo chiamato *Mul'tipark* (Grišina 2011). Il corpus appare nel sito come un sub-corpus a sé, anche se alcuni, come Grišina, lo ritengono parte del corpus multimediale MURKO. Il suo obiettivo è quello di confrontare le diverse interpretazioni artistiche e teatrali di una stessa opera. Il corpus è diviso in tre parti. La prima include le diverse interpretazioni o letture orali di testi poetici e prosaici, da parte sia di attori professionisti che non. Un esempio può essere la declamazione della poesia *Odin idet prijamym putem…* della Achmatova da parte di Svetlana Krjučkova, Alla Demidova e altre attrici. La seconda riguarda invece le diverse messe in scena e adattamenti cinematografici di opere teatrali come può essere Il Revisore di Gogol, tanto famoso da essere stato adattato innumerevoli volte. La terza riguarda invece materiali multilingue. La sezione include infatti gli adattamenti cinematografici di una stessa opera in diverse lingue, come ad esempio le

versioni americana, inglese e russa di Guerra e Pace o le versioni russa e francese di Anna Karenina.

2.1.4 I corpora paralleli del NKRJa

Come anticipato, ai corpora paralleli del NKRJa dedichiamo interamente questa sezione con lo scopo di illustrare in maniera più dettagliata le loro caratteristiche e il loro funzionamento, in particolare per quanto riguarda il corpus parallelo italiano-russo.
Come ci spiega Noseda (2017:43) la sezione del Corpus Nazionale Russo dedicata ai corpora paralleli fu creata nel 2005 con il nome di KoParT, che comprendeva testi in russo e in inglese. Questo fu, in generale, uno dei primi corpora paralleli apparsi nella russistica. I testi inizialmente presenti al suo interno erano solamente opere letterarie. Poi, con il tempo, la sezione ha iniziato ad essere ampliata per quanto riguarda la quantità di testi, le tipologie testuali e anche il numero di lingue. Il numero complessivo di parole di cui attualmente conta è 98 milioni e le lingue incluse sono diciannove, ovvero: armeno (2 milioni), baschiro, bielorusso (8 milioni), bulgaro (3 milioni), buriato (75 mila), ceco, cinese (55 mila), estone (408 mila), finlandese, francese (3 milioni), inglese (24 milioni), italiano (4,5 milioni), lettone (777 mila), lituano, polacco (6 milioni), spagnolo (320 mila), svedese (409 mila), tedesco (9 milioni) (a cui si aggiunge un sub-corpus composto esclusivamente da grandi classici russi tradotti in tedesco chiamato *Russkaja klassika v nemeckich perevodach*) e ucraino (9 milioni)[13]. I più recenti tra questi sono quelli in lingua baschira, ceca, finlandese e lituana[14]. Tra questi vi è anche un corpus multilingue di 5 milioni di parole. Si tratta in tutti i casi di corpora bi-direzionali, che permettono quindi di svolgere ricerche sia da che verso il russo. Come ogni altro sub-corpus, anche quelli paralleli sono dotati di annotazione meta-testuale, morfologica e semantica.

[13] Cifre risalenti all'anno 2017, non aggiornate al 2020 in quanto non presenti sul sito nel NKRJa. La scelta di riportare comunque le cifre è dovuta al fatto di voler dare un'idea della dimensione dei singoli corpora in proporzione a quelli delle altre lingue. Le loro dimensioni potrebbero quindi essere aumentate vista la velocità e sistematicità con cui il NKRJa viene aggiornato.
[14] Il materiale più recente utilizzato per la sezione dedicata ai corpora paralleli arriva fino all'anno 2017. In questo materiale queste quattro lingue non erano ancora presenti. Sono invece presenti oggi (2020) sul sito del NKRJa. Mentre per ogni sub-corpus viene mostrato il numero di parole che contiene, questo non avviene per ogni corpus parallelo. Non sono quindi al momento reperibili informazioni sul numero di parole i nuovi corpora paralleli contengono.

2.1.4.1 Il corpus parallelo italiano-russo

Le informazioni più complete che abbiamo oggi sul corpus parallelo italiano-russo ci vengono date dai lavori di Valentina Noseda (2017, 2018), ricercatrice italiana che ha partecipato al progetto di ampliamento del corpus. Il corpus parallelo italiano-russo è stato infatti sottoposto ad un importante processo di ampliamento nell'anno 2015. Prima di allora era uno dei più piccoli, trovandosi al terzultimo posto per numero di parole. Il corpus era nato pochi anni dopo la nascita del NKRJa come esperimento pilota. Includeva in tutto cinque testi, di cui tre originali italiani con la traduzione russa e due originali russi con la traduzione italiana, per un totale di sole 700 mila parole. Oltre al numero ridotto di parole, presentava anche evidenti lacune che lo rendevano inadatto al ruolo di valido strumento di ricerca, tra cui il suo carattere non sistematico, ovvero la mancanza di chiari criteri di progettazione e costruzione. In assenza di altri corpora paralleli per la coppia di lingue italiano-russo, divenne evidente la necessità di ampliarlo. Questo progetto di ampliamento iniziò nel 2014 grazie alla collaborazione tra l'Istituto di Lingua Russa di Mosca (*Institut Russkovo Jazyka imeni V. V. Vinogradova*), la cattedra di Lingua e Letteratura Russa dell'Università Cattolica di Milano e la Scuola Superiore di Lingue Moderne per Interpreti e Traduttori di Forlì (Università di Bologna). A dicembre del 2015 il progetto viene così realizzato, e grazie ad esso vennero anche poste le basi per assicurarne uno sviluppo che fosse sistematico. Da questo momento in poi verrà infatti costantemente ampliato e monitorato.

I criteri che sono stati seguiti nel corso della sua espansione si rifanno al concetto di rappresentatività, essendo un corpus vasto e variegato, e bilanciamento, in quanto contiene in egual misura testi di generi differenti. Viene poi stabilito che il corpus sarebbe stato bidirezionale, così da consentire un maggior numero di ricerche e per rendere il corpus comparabile, oltre che parallelo. In questo modo è possibile confrontare gli originali in russo o italiano con le rispettive traduzioni. Dove presenti, il corpus contiene anche diverse traduzioni di uno stesso testo, il cui confronto può essere utile in diversi ambiti di ricerca, come ad esempio linguistica contrastiva, studi traduttivi o ricerca letteraria.

I testi presenti nel corpus appartengono ad una gamma di generi limitata rispetto ad un corpus monolingue, in quanto non tutti i generi vengono tradotti. In un corpus parallelo il genere dominante è solitamente quello letterario, seguito dalla saggistica. Per quest'ultima si è scelto, per differenziare il più possibile i testi, di includere brani che trattassero argomenti diversi.

Per quanto riguarda le dimensioni, i creatori del corpus si erano posti l'obiettivo di raggiungere almeno tre milioni di parole, pari all'*English-Norwegian Parallel Corpus*, adottato spesso come modello di progettazione di corpora paralleli per via della sua sistematicità e bilanciamento.

La selezione dei testi è avvenuta sulla base dei database delle maggiori biblioteche nazionali,

dell'*Index Translationum*[15] e dei cataloghi delle case editrici, iniziando ad ampliare prima la prosa letteraria per poi passare alla saggistica.
I testi che troviamo nel corpus non sono limitati ad un solo periodo di pubblicazione, in modo da consentire studi anche di tipo diacronico. Sono infatti state scelte opere russe a partire dal XVIII secolo e pertanto anche i testi italiani con la traduzione russa non sono anteriori a questo periodo. Le traduzioni sono state scelte in base ad alcuni criteri, ovvero i traduttori dei testi selezionati devono essere noti (non anonimi) e parlanti nativi della lingua del testo d'arrivo. In caso di testi tradotti più volte sono state scelte le traduzioni più autorevoli. I testi, che possono essere in versione integrale o ridotta, si trovano in numero pressoché pari in lingua 1 e 2 (tranne per alcune eccezioni) così da rispettare il bilanciamento, trattandosi di un corpus bidirezionale. Per evitare di inserire troppi testi dello stesso stile si è pensato di scegliere inizialmente solamente un'opera per autore. Per quanto riguarda la ricerca, il suo funzionamento è analogo a quello del resto del NKRJa.

Ma quali sono i risultati a cui ha portato questo processo di ampliamento? Prima di tutto sono aumentate le dimensioni, passando da 700 mila a 4,5 milioni di parole. In secondo luogo, parlando di equilibrio tra i testi, il corpus risulta purtroppo ancora piuttosto sbilanciato: i testi letterari russi sono 33 contro i 21 italiani, i testi di saggistica russi sono 5 mentre non ne sono presenti di italiani, la sezione russo-italiano include 3,5 milioni di parole mentre quella italiano-russo poco più di 700 mila, ed infine gli estratti e racconti brevi sono 42 mentre le opere in versione integrale sono solo 17. Al di là di questo aspetto, però, abbiamo ora uno strumento di ricerca linguistica affidabile e con base statisticamente rilevante. Per il futuro i creatori si pongono come obiettivo quello di equilibrare questi dati e aumentare le opere di saggistica per avere uno strumento ancora più affidabile di quanto lo sia oggi.

Per chiudere questa sezione è interessante citare altri corpora paralleli disponibili online per quanto riguarda la coppia di lingue russo-italiano. Uno di questi è InterCorp, nato con l'intento di creare un ampio corpus parallelo comprendente tutte le lingue studiate alla Facoltà di Arte dell'Università Carolina di Praga, tra cui appunto l'italiano e il russo. L'altro è invece OPUS, un corpus che raccoglie dati provenienti da diversi settori e che ricopre più di 90 lingue. I testi che include sono per lo più legislativi o amministrativi, reperiti dai database delle Nazioni Unite o di altre istituzioni come l'Unione Europea. Ci sono però anche sottotitoli di film, articoli di giornale, e altri testi tratti da varie risorse online. Con Opus è possibile selezionare una coppia di lingue e verranno proposti

[15]Un indice delle opere tradotte in tutto il mondo. Creato dalla Società delle Nazioni e gestito dall'UNESCO, la banca dati contiene opere di tutte le discipline: letteratura, scienza sociale, scienze naturali, arte, storia, ecc.

all'utente tutti i corpora paralleli disponibili per tali lingue (Noseda 2016: 48).

2.2 Il *General'nyj Internet-Korpus Russkovo Jazyka*

Dopo aver analizzato a fondo il più importante corpus tradizionale russo, ci spostiamo ora a uno di quei corpora definibili, in maniera semplice ma efficace, con due aggettivi: "mega" e "*internet-based*". Si tratta di corpora di generazione ancora più nuova rispetto ai normali corpora elettronici, già di per sé considerati di nuova generazione. L'aggettivo "mega" viene utilizzato per descriverne le dimensioni, che sono di gran lunga superiori rispetto ai corpora tradizionali; se per questi ultimi si parla di milioni di parole, per i mega corpora si parla addirittura di miliardi. L'aggettivo "*internet-based*" sta invece ad indicare il fatto che questa tipologia di corpora si basa interamente sul web come fonte di raccolta dei suoi testi. Non andiamo oltre, in questo momento, con l'analisi di questa tipologia di corpora, necessaria qui per introdurre il corpus di cui stiamo per parlare, assieme ad alcuni altri sempre nell'ambito della russistica. Il tema dei corpora dell'internet verrà poi trattato più approfonditamente nel prossimo capitolo, dedicato interamente al web inteso come corpus.

Il *General'nyj Internet-Korpus Russkovo Jazyka* (GIKRJa) è un corpus il cui materiale è tratto interamente dal web, ma con un criterio particolare, che vedremo a breve. Ci troviamo di fronte ad un vero e proprio mega corpus il cui numero totale di parole è di ben 20 miliardi, molto distante dai 700 milioni del NKRJa, visto che parliamo di una dimensione di quasi 30 volte superiore. Il corpus, aperto nel 2012, si trova tuttora in fase di sviluppo e non è ancora aperto a tutti. L'accesso, infatti, è possibile tramite richiesta inviando una mail al contatto presente sul sito e spiegando le ragioni della volontà ad utilizzare il corpus. Dopo aver analizzato e ritenuto idonea la nostra domanda, i creatori ci forniranno delle credenziali con cui potremo accedere al corpus. Il progetto dello sviluppo del corpus vede la partecipazione di diverse figure, quali gli studenti del Dipartimento di Linguistica Computazionale dell'Università Statale di Studi Umanistici e l'Istituto di Fisica e Tecnologia, entrambe di Mosca, l'Istituto di Scienza e Tecnologia Skolkovo, così come gli esperti dell'Università di Leeds in Inghilterra e di altre società. Il progetto si dichiara inoltre aperto a ricercatori esterni qualora volessero partecipare alla sua realizzazione.

Il GIKRJa, come accennato prima, è stato creato con un obiettivo particolare, ovvero quello di rappresentare la lingua russa dell'internet, una lingua cioè colloquiale e spontanea. Il corpus nasce per rispondere alle inadeguatezze dei corpora tradizionali al raggiungimento di questo scopo,

registrando solo sporadicamente forme di linguaggio colloquiale. Si prenda ad esempio il NKRJa, in cui un linguaggio orale, ma non totalmente spontaneo, lo si ritrova solamente nel sub-corpus *Korpus ustnoj reč*, che su 700 milioni di parole totali del corpus, ne copre solamente 12 milioni. Inoltre si tratta infatti per lo più di discorsi non spontanei, vedasi i discorsi pubblici, che sono il 52%, e le trascrizioni dei film che sono il 38%. Da ultimo, abbiamo che, in questo sub-corpus, la parte che riguarda il linguaggio puramente colloquiale e spontaneo comprende solamente il 10% (Grišina, Savčuk 2009:134). Anche il corpus di lingua russa orale *Odin rečevoj den*, che vedremo nella sezione dedicata agli altri corpora di lingua russa, contiene materiale orale spontaneo. Il numero limitato del materiale che contiene non lo rende però uno strumento adatto a studi di carattere generale.

Anche per quanto riguarda la lingua scritta, i corpora tradizionali come il NKRJa sono soliti focalizzarsi in primo luogo su testi letterari e di narrativa, che di colloquiale hanno molto poco. Il GIKRJa vuole invece recuperare la lingua vera, spontanea, con i suoi slang e tratti regionali. Ed è qui che entra in gioco la lingua del web. Il motivo per cui è stata scelta come valida alternativa al linguaggio orale è la sua caratteristica di trovarsi a metà tra la lingua scritta e quella parlata, ovvero avvicinarsi, per le sue peculiarità, al linguaggio orale, ma garantendo allo stesso tempo i vantaggi della ricerca su materiale fissato nello scritto (Riti 2017: 39). L'utente, nel web, scrive ciò che direbbe a parole in una situazione comunicativa orale. La lingua che ne risulta contiene quindi forme tipiche del russo parlato che non rientrano nella norma scritta. Alcune di queste forme del parlato o violazioni della norma scritta, tipiche della lingua russa del web possono essere (Che Chen 2013, Trofimova 2011): errori di ortografia o punteggiatura, elementi del *prostorečie*, violazione di alcune norme stilistiche e norme dell'ordine delle parole, semplificazione delle frasi, costruzione del testo come sequenza di frasi slegate, scrittura fonetica delle parole, coesistenza di alfabeto cirillico e latino, adattamento di termini stranieri alle strutture morfosintattiche russe, ed assieme a queste ne esistono ancora molte altre.

Nonostante sia oggi possibile interrogare la lingua di internet direttamente da un motore di ricerca per svolgere delle analisi linguistiche, questo porta ad una serie di problemi ed ha inoltre dei limiti, che vedremo nel prossimo capitolo. Il GIKRJa è quindi lo strumento di cui abbiamo bisogno in questo caso. Nei corpora tradizionali, è comunque presente una parte di materiale tratto dal web, ma questo non è sufficiente. Si tratta infatti, nel caso del NKRJa, del 20%, che è per di più materiale non aggiornato risalente all'anno 2007.

Un'altra particolarità del GIKRJa è quella di essere un compromesso tra un corpus chiuso, come lo sono di solito i corpora tradizionali, e aperto, come i mega-corpora. Il materiale che si trova al suo

interno consiste in una raccolta di pagine web dei più importanti social media di lingua russa, ovvero la piattaforma blog Livejournal (*Živoj Žurnal*) e la rete sociale *Vkontakte*, che contano rispettivamente 8720 e 9820 milioni di parole. Inoltre sono presenti anche gli articoli di Novosti.ru, sito che raccoglie notizie di quotidiani online e agenzie stampa, per un totale di 851 milioni di parole, e *Žurnal'nyj zal*, una raccolta di giornali a carattere letterario, per un totale di 313 milioni di parole. Questi ultimi rappresentano solamente il 5% dell'intero materiale del corpus. Nel caso di entrambe le reti sociali gli utenti sono giovani al di sotto dei 34 anni. La scelta di raccogliere il materiale da queste fonti è dovuta al fatto che proprio nei blog e post dei social network si ritrova il particolare linguaggio "ibrido" di cui abbiamo parlato. Il materiale proviene quindi da una sezione specifica dei contenuti web in lingua russa, che essendo per l'appunto limitata si allontana da quello che è il meccanismo di raccolta dei testi dei corpora puramente aperti. Questi ultimi, infatti, recuperano il materiale presente nel web in maniera del tutto casuale secondo metodi statistici.

Il GIKRJa si differenzia ancora una volta dagli altri mega-corpora del web per la presenza dei dati meta-testuali come età, sesso, regione di residenza, nome o username degli utenti autori dei testi. Grazie a questi dati si possono svolgere, con l'utilizzo di questo corpus, ricerche di tipo sociolinguistico ma anche studi di variazione linguistica su territorio russofono e ricerche dialettologiche. Tra le sue peculiarità c'è infatti quella di includere 16 sub-corpora di carattere regionale, di cui due per l'Ucraina (Oblast' di Donec'k e Kiev) e 14 per la Russia (Krasnodar, Krasnojarsk, Mosca, Novosibirsk, Omsk, Perm', Baschiria, Tatarstan, Rostov, Samara, San Pietroburgo, Saratov, Sverdlovsk, Čeljabinsk) (Piperski 2013: 4).

Non è previsto invece, con il GIKRJa, uno studio di tipo diacronico della lingua dal momento che il corpus viene continuamente aggiornato eliminando i materiali più vecchi di cinque anni. Il corpus si pone quindi volutamente come sincronico, rendendo una fotografia della lingua in un momento ben preciso. Sono possibili, grazie al suo utilizzo, ricerche di tipo sociologico, come possono essere l'influenza di genere, età e altri fattori sulla lingua, e l'analisi dei social network, così come ricerche di tipo linguistico, come gli studi sulla distribuzione delle parole, sulla lingua dei social network, sulla frequenza delle parole, studi sulle espressioni e costruzioni linguistiche, sulle caratteristiche stilistiche dei testi nelle diverse sezioni del web e molti altri sulla stessa scia.
Per quanto riguarda gli studi di tipo sociologico, questo corpus, rispetto a quelli tradizionali, risulta essere uno strumento più valido se si considera il fatto che il linguaggio che contiene appartiene a categorie sociali più ampie. Nei corpora tradizionali invece, trattandosi per lo più di testi letterari, alcune categorie sociali sono totalmente escluse.

Oltre all'annotazione meta-testuale, il corpus è dotato anche di annotazione morfologica e lemmatizzazione. Assieme a tutti i vantaggi che questo corpus offre, ci sono però anche alcuni svantaggi. Manca infatti la possibilità di filtrare i risultati delle ricerche sulla base del contesto precedente o seguente la parola ricercata. Un altro problema è quello delle annotazioni. Dal momento che i testi tratti dal web possono essere non lemmatizzati correttamente, questo può portare ad errori di tag, in particolare quando si tratta di termini gergali e slang. Annotazioni errate possono però derivare anche da errori di battitura o dalla scrittura fonetica di alcune parole (Selegej et al 2016: 9).

Nonostante ciò, il GIKRJa rimane uno degli strumenti più innovativi per l'analisi della lingua russa, sia per le sue dimensioni, che garantiscono risultati che con altri corpora è difficile raggiungere, sia per la particolarità del genere testuale che si è scelto di includere, sia, infine, per l'attenta selezione del materiale raccolto e del suo monitoraggio.

2.3 Altri corpora della lingua russa

Oltre ai due corpora della lingua russa appena visti, che possiamo considerare uno il corpus tradizionale per eccellenza, e l'altro uno tra i web-corpora più notevoli, vediamo ora una serie di altri corpora, alcuni tradizionali, altri di internet. Di questi, alcuni sono mega e altri più contenuti, alcuni generici altri specializzati, alcuni di lingua scritta e altri di lingua parlata. Questo, per dare una visione generale dei corpora di cui la lingua russa dispone, che gli utenti possono scegliere in base al diverso utilizzo che ne dovranno fare. Alcuni di questi corpora di lingua russa fanno parte di quelle che vengono spesso chiamate "famiglie" di corpora, ovvero software online che includono molteplici corpora di diverse lingue ma anche di diverse tipologie (ad esempio Sketch Engine, Aranea)

2.3.1 Mega-corpora del web

Aranea. Aranea è un progetto realizzato dall'Accademia Slovacca delle Scienze di Bratislava nato dalla mancanza, in questa, di corpora adeguati che potessero essere utilizzati dagli studenti di lingue straniere e studi traduttivi. I creatori lo descrivono come un progetto slovacco-centrico dal momento che è stato da principio pensato per includere le lingue utilizzate in Slovacchia e nei paesi

circostanti, come Repubblica Ceca, Germania, Ungheria, Polonia, Ucraina, ma anche quelle insegnate nelle scuole e università di questi paesi come l'inglese, il francese, lo spagnolo, l'italiano, il russo (Benko 2014). L'idea era quindi quella di creare una famiglia di corpora comparabili e che includesse per l'appunto molteplici lingue. Nonostante esistessero altre famiglie di corpora, come ad esempio Sketch Engine, i creatori di Aranea ritenevano che in questi mancasse la possibilità di effettuare il download dei corpora e che fossero troppo grandi per un utilizzo in classe, che era proprio l'obiettivo principale del loro progetto. Ecco quindi che nel 2013 il progetto viene realizzato, e, grazie al suo continuo sviluppo, è arrivato oggi a contenere un totale di 18 lingue. Trattandosi di corpora comparabili, i testi delle diverse lingue sono di tipologie, generi, registri e dimensioni simili ed accessibili da un unico luogo, ovvero il software NoSketch Engine[16]. I testi sono inoltre dotati di annotazione morfosintattica. Ogni corpus ha un nome latino diverso in base alla lingua, ovvero Araneum Russicum per il russo, Araneum Anglicum per l'inglese, Araneum Germanicum per il tedesco e così via. Ognuno di questi corpora è composto da due varianti, che si differenziano per il loro volume: Maius è la versione medio-grande, mentre Minus è quella piccola (include il 10% del materiale della versione Maius estratto in maniera casuale). Per alcune lingue è anche presente la versione Maximum, ovvero quella più voluminosa. Araneum Russicum Maius contiene più di 850 milioni di parole mentre Araneum Russicum Minus ne contiene 90 milioni[17]. Per il russo esiste anche la versione Maximum[18] che contiene 11 miliardi di parole (Kutuzov, Kunilovskaja 2017: 1), e può essere per questo considerato il più grande web-corpus russo dopo il GIKRJa. Oltre a quelli appena citati, per la lingua russa esistono corpora anche di altre versioni: Araneum Russicum Russicum, che contiene testi tratti dai domini .ru e. rf; Araneum Russicum Externum, che contiene testi in lingua russa ma "esteri", ovvero da domini diversi da .ru e .rf.; Araneum Russicum, che contiene testi russi ricavati da diversi siti, indipendentemente dal loro dominio.

Sketch Engine[19]. Altra famiglia di corpora, una tra le più grandi esistenti al momento. Le lingue comprese sono ben 143, tra cui le principali lingue del mondo, lingue maggioritarie e minoritarie, lingue morte, varietà regionali e locali[20]. Una delle funzioni principali di questo software, che è anche una sua peculiarità proprio come dice il nome, è quella dei *word sketches*. Si tratta di una sintesi del comportamento grammaticale e delle collocazioni di una parola. Ad esempio, ricercando,

[16] Una versione limitata del software Sketch Engine, accessibile dal sito https://nlp.fi.muni.cz/trac/noske
[17] Dati ricavati da: http://ucts.uniba.sk/aranea_about/_russicum.html
[18] Questa versione, oltre che da NoSketch Engine, può essere consultata anche da Sketch Engine
[19] http://www.sketchengine.eu
[20] La lista completa delle lingue incluse è presente a questo link: https://www.sketchengine.eu/corpora-and-languages/ Per ogni lingua viene indicato se sono disponibili l'annotazione morfosintattica, la lemmatizzazione, i *word sketches* e la ricerca delle parole chiave

nella sezione *word sketches* un aggettivo, apparirà una schermata in cui vengono mostrati, in sezioni apposite: i sostantivi a cui viene solitamente accostato, i suoi sinonimi o contrari, avverbi che lo accompagnano, preposizioni con cui viene utilizzato, costruzioni verbali in cui si può incontrare e tutte le sue collocazioni. Nel caso in cui si ricerchi un verbo, di questo verranno mostrati: possibili soggetti e complementi oggetto, preposizioni che lo possono seguire, avverbi e aggettivi che lo accompagnano e i suoi sinonimi e contrari. Si ha così un quadro completo di tutti gli utilizzi e le caratteristiche grammaticali della parola oggetto di ricerca. Questa funzione, che purtroppo non è ancora disponibile per tutte le lingue ma per quelle principali, fu utilizzata per la prima volta nella realizzazione del dizionario di lingua inglese Macmillan e, una volta che il software fu presentato alla Euralex 2002[21] per la lingua inglese, i rappresentanti delle altre lingue si chiesero quando avrebbero potuto avere uno strumento simile per la propria (Kilgarriff et al 2004). Da qui è iniziato un importante processo di ampliamento ad altre lingue. Tra le altre funzioni di Sketch Engine ci sono anche: *sketch differences*, che permette di specificare, per due parole semanticamente vicine, quali comportamenti condividono e quali no; thesaurus, che permette di mostrare le parole che ruotano attorno allo stesso campo semantico di quella ricercata; le concordanze e la loro visualizzazione parallela; liste di parole; termini e parole chiave; estrazione bilingue delle parole; tendenze e liste di frequenza; dizionari. È inoltre presente l'annotazione morfosintattica, anche se non per tutte le lingue, così come la lemmatizzazione. L'annotazione meta-testuale è presente ma è ridotta rispetto ai corpora tradizionali. Trattandosi di pagine web, è difficile recuperare dati come la data in cui è stata scritta o l'autore, a meno che non si tratti di articoli di giornali o blog.

Oltre a corpora di diverse lingue, sono presenti anche corpora di diverse tipologie e con diverse tipologie di testi, realizzati da altri sviluppatori. Molti di questi sono web-corpora ma non sono però gli unici. Il nome dei corpora propri di Sketch Engine è composto dalla sigla della lingua seguita da "TenTen", ad esempio ruTenTen per il russo, itTenTen per l'italiano, enTenTen per l'inglese. Di questi sono presenti diverse versioni a seconda dell'anno in cui sono stati raccolti. Ecco quindi che avremo ruTenTen11 creato nel 2011 e itaTenTen16 creato nel 2016. Il corpus russo più importante presente in Sketch Engine è proprio ruTenTen, che conta 15 miliardi di parole. Questo include sia l'annotazione morfosintattica, che la funzione *word sketches*, che la lemmatizzazione. Oltre a ruTenTen, per la lingua russa sono accessibili da qui anche altri corpora come Aranum Russicum Maius (i corpora di Aranea sono disponibili anche per le altre lingue), CHILDES Russian Corpus, OPUS Russian e altri.

[21] la più importante associazione europea per linguisti, lessicografi e altre figure simili. Sono un luogo di scambio di idee in questi ambiti grazie all'organizzazione di forum e congressi

Sketch Engine ha raggiunto oggi una grande rilevanza nella linguistica computazionale, combinando gli approcci della linguistica tradizionale e della statistica (Chochlova, Zacharov 2010:1). È uno strumento spesso utilizzato anche per la compilazione di grammatiche e dizionari.

WaCky[22]**.** Anche WaCky è una famiglia di corpora di diverse lingue. Questi, similarmente a Sketch Engine, sono denominati con la sigla della lingua seguita da "WaC". Avremmo quindi itWaC per il corpus di lingua italiana, ruWaC per quello di lingua russa e così via. Le loro dimensioni vanno dai 2 milioni ai 20 miliardi di parole (Zacharov, Benko 2016: 81). La maggior parte di questi corpora sono accessibili liberamente dal software NoSketch Engine[23], tra cui quello inglese, italiano, francese, tedesco e alcuni altri. Altri sono invece ma privati o accessibili su licenza. Questo è purtroppo il caso del corpus di lingua russa, che non risulta al momento accessibile. Si tratta comunque di un utile strumento linguistico, anche grazie al fatto di essere dotato di annotazione morfosintattica.

2.3.2 *Learner corpora*

Child Language Data Exchange System (CHILDES)[24]. È un database che contiene le produzioni linguistiche dei bambini, in diverse lingue tra cui anche il russo, per segnare le tappe della loro acquisizione linguistica. Il materiale in esso contenuto consiste sia in trascrizioni che in clip audio e video di linguaggio spontaneo prodotto dai bambini. Il corpus, che è accessibile anche da Sketch Engine, si presta altresì a studi di carattere morfologico, oltre che di altri tipi.

Korpus Nesoveršennogo perevoda[25]**.** È un corpus parallelo di traduzioni della coppia di lingue inglese-russo, in cui ogni traduzione è allineata con il testo di partenza. Il corpus è dotato di annotazione meta-testuale, con informazioni sia sul testo che sul traduttore, e morfosintattica. I testi tradotti sono di diverso genere: accademici, informativi, saggi, interviste, narrativa, educativi, lettere, pubblicità ecc.

Korpus Russkich Učebnych Tekstov **(KRUT)**[26]. È una collezione di testi scritti da studenti di diverse università e che appartengono a diversi corsi di laurea come economia, sociologia, scienze politiche, giurisprudenza, psicologia, giornalismo, linguistica, storia, filologia, matematica,

[22] https://wacky.sslmit.unibo.it/doku.php
[23] https://corpora.dipintra.it/public/run.cgi/first_form
[24] https://childes.talkbank.org/
[25] https://rus-ltc.org/search
[26] http://web-corpora.net/CoRST/search/?interface_language=ru

filosofia. Le tipologie principali di testi sono: esami, tesi, saggi, riassunti, relazioni, autobiografie ecc., per un totale di 3.1 milioni di *token*. I testi sono annotati dal punto di vista meta-testuale, indicando informazioni sugli studenti come sesso, età, corso di laurea e anno e per alcuni testi anche la loro provenienza ed eventuale bilinguismo, e sul testo, come la materia e l'ambito di cui tratta. Sono anche presenti l'annotazione morfologica ma soprattutto una tipologia particolare di annotazione ovvero l'*error tag*. Questo segnala il tipo di errore (lessicale, stilistico, grammaticale) commesso dallo studente e la possibile causa che lo ha portato a commettere tale errore.
Il corpus, che permette di mettere in luce i principali errori commessi dagli apprendenti di lingua russa, è pensato principalmente per ricerche di tipo linguistico, essendo utile sia a linguisti (ricercatori, insegnanti di lingua russa), che a studenti per il loro apprendimento, ma può essere utile anche a livello sociologico.

Russkij Učebnyj Korpus[27]. È una collezione di testi prodotti da due categorie di persone: apprendenti di lingua russa come lingua straniera e parlanti di lingua russa come seconda lingua con prime lingue diverse. Il corpus contiene produzione linguistica sia scritta che orale, che consiste in testi sia accademici che non, come descrizioni di film e immagini, riassunti di libri, saggi ecc. Dotato di annotazione morfologica, permette di mettere in luce un uso della lingua che si discosta dalla lingua standard a causa di errori ortografici o grammaticali.

2.3.3 Corpora storici e dialettali

Oltre ai corpora storici del NKRJa che abbiamo visto nella sezione dedicata, per la lingua russa sono presenti anche altri corpora, sia storici che dialettali, che vediamo di seguito.

Corpus del *Velikie Minei Čet'i*[28] (Mitrenina 2014). È un corpus elettronico del più grande menologio in slavo ecclesiastico compilato da Macario, metropolita di Mosca nel XVI secolo. Il corpus contiene i testi de gli Atti degli Apostoli, le Lettere di Paolo, le lettere Ecumeniche, per un totale di quasi 300 mila parole. Non è dotato di annotazione grammaticale, quindi non permette ricerche da questo punto di vista. Permette invece una ricerca di tipo lessicale.

[27] http://web-corpora.net/RLC
[28] http://www.vmc.uni-freiburg.de/Mens/

Corpus parallelo delle traduzioni del Canto della schiera di Igor'[29]. Contiene le traduzioni dell'opera russa in diverse lingue e realizzate da diversi traduttori. Selezionando le versioni da confrontare, è possibile visualizzare i diversi frammenti dell'opera allineati.

Manuscript'[30]. È una raccolta delle copie in formato elettronico di manoscritti in *drevnerusskij* e *starorusskij* per un totale di 3,5 milioni di parole.

Regensburgskij diachroničeskij korpus[31] (Mitrenina 2014). È una raccolta di testi in russo antico. Si trova in due versioni, una vecchia e una nuova. Quella vecchia include 11 testi in *drevnerusskij* e *starorusskij*. Quella nuova, che contiene più di 100 mila parole, contiene i seguenti testi: le opere di Kirill di Turov, *Domostroj*, *Choždene Bogorodicy po mukam*, *Povest' vremennych let*, *Mineja Janvar* e le Cronache di Novgorod. Tutti i testi contengono annotazione morfosintattica e questo lo rende l'unico corpus che permette di effettuare ricerche dal punto di vista sintattico per testi in russo antico.

***Sankt-Peterburgskij korpus agiografičeskich tekstov* (SKAT)**[32]. Contiene testi della letteratura agiografica antica appartenenti all'epoca tra il XV e il XVII secolo, per un totale di 200 mila parole. Al momento non contiene l'annotazione grammaticale e sintattica, che è prevista invece per il futuro.

Ustya River Basin corpus[33] (von Waldenfels et al 2014). È un corpus creato nel 2013 da un gruppo di studenti universitari russi e svedesi durante una ricerca sulla variazione linguistica e i dialetti. Il gruppo, al momento della ricerca, era stanziato nella regione Ustya dell'Archangelskaja Oblast', nel villaggio di Michalevskaja. Gli studenti, muovendosi tra questo e i villaggi circostanti, intervistarono gli abitanti chiedendo loro di raccontare delle loro vite e altre storie. Il materiale orale raccolto, per un totale di 40 ore di conversazione, è stato poi trascritto in russo standard. Questo è stato poi sottoposto a lemmatizzazione e annotazione morfosintattica.

A questi vanno ovviamente aggiunti i corpora storici del NKRJa che abbiamo visto

[29] http://nevmenandr.net/slovo/
[30] http://manuscripts.ru/index_en.html
[31] Accessibile su richiesta
[32] http://project.phil.spbu.ru/scat/page.php?page=project
[33] http://www.parasolcorpus.org/Pushkino/login.php#

2.3.4 Corpora della lingua russa parlata

CORPRES[34] (Skrelin et al 2010). Questo corpus di lingua russa orale è stato sviluppato dal dipartimento di fonetica dell'Università Statale di San Pietroburgo. Contiene 60 ore di parlato da parte di 8 parlanti, 4 uomini e 4 donne, di diverse aree di San Pietroburgo. La lingua orale presente nel corpus CORPRES non è spontanea ma si tratta invece della lettura di testi di diverse tipologie: narrativa con dialoghi, narrativa più descrittiva, drammi con dialoghi espressivi dal punto di vista emotivo, testi neutri puramente informativi ecc. Le registrazioni, che sono di alta qualità tecnica, sono state poi trascritte dal punto di vista fonetico ed è stata applicata l'annotazione prosodica, di intonazione e ortografica. Il corpus, che contiene un totale di più di 500 mila parole, vuole essere un campione della lingua russa standard nella variante di San Pietroburgo, con le sue diverse pronunce.

CoRuSS[35] (Kachkovskaja et al 2016). È un corpus di lingua russa parlata spontanea basato su situazioni comunicative (dialoghi, monologhi, lettura di un breve testo) registrate da 60 russofoni, che sono uomini e donne di età dai 16 ai 77 anni. Le registrazioni, che contengono in totale 30 ore di parlato, sono state poi trascritte e annotate dal punto di vista fonetico e prosodico.

Odin rečevoj den[36] (Asinovskij et al 2009, Šerstinova 2009). È un corpus sviluppato dall'Istituto di Ricerca Filologica dell'Università Statale di San Pietroburgo. Il significato letterale del nome di questo corpus è "un giorno di conversazioni". Questo, infatti, contiene le registrazioni delle conversazioni quotidiane del russo medio. Per realizzarlo è stato scelto un gruppo di persone che fosse demograficamente equilibrato e che rappresentasse varie fasce sociali e d'età della popolazione di San Pietroburgo. Questi individui hanno trascorso un'intera giornata con un registratore al collo che registrava le loro conversazioni nei vari momenti della giornata, dalla colazione, al lavoro, alle conversazioni telefoniche, al pranzo, ai momenti di svago. Il corpus contiene oggi più di mille ore di conversazione (Kachkovskaja et al 2016). Queste sono state divise in una serie di episodi comunicativi poi trascritti nel dettaglio. Un corpus di questo tipo è una preziosa fonte di dati per ricerche di diverso tipo, sia linguistico, in quanto permette di avere un linguaggio orale e spontaneo che difficilmente si trova in altri corpora, che sociologico, grazie anche ai metadati che contengono informazioni sui parlanti come nome, sesso, età, luogo di nascita, gruppo sociale, educazione, qualifiche, occupazione, nazionalità ecc.

[34] Ibidem
[35] Ibidem
[36] Non sono presenti nella letteratura informazioni sulle modalità di accesso al corpus. Durante questo lavoro di ricerca non è stato trovato alcun sito che permettesse di accedervi.

***Učebnyj Mul'timodal'nyj Korpus* (UMKO)** (Zacharov 2013: 13). È un corpus che contiene dialoghi di studenti di lingua cinese, russa e tedesca sottoforma di 25 video clip della durata di 1,5-3 minuti. Sono presenti al suo interno una serie di sub-corpora paralleli in cui le conversazioni da parte dei nativi di una lingua sono allineati alle conversazioni nella stessa lingua ma da parte di apprendenti.

2.3.5 Corpora di comunicazione non verbale

MURCO[37]. È da inserire anche qui il corpus multimediale di lingua russa, sub-corpus del NKRJa già analizzato in precedenza.

Russian emotional corpus (Kotov, Budyanskaya 2012). È un corpus che contiene registrazioni video di comunicazione spontanea ed emozionale. Il materiale proviene da 3 fonti che sono mass media, performance di attori, dati sperimentali in cui stimoli emozionali sono indotti in situazioni sperimentali. Si tratta di situazioni in cui i soggetti hanno una forte motivazione nel comunicare e raggiungere il proprio obiettivo. Il corpus è dotato di annotazione che descrive linguaggio, espressioni facciali, gesti delle mani, rapidi cambiamenti di espressione, strategie comunicative come umore, educazione, emozioni, e sistema dialogico.

2.3.6 Altri corpora

***Chel'sinskij annotirovannyj korpus* (CHANKO)**[38] (Kopotev, Mustajoki 2003). È un progetto iniziato nel 2001 dal Dipartimento di Lingue e Letterature Slave e Baltiche dell'Università di Helsinki. Si tratta di un corpus di lingua russa non molto grande dotato però di annotazione morfologica, sintattica e semantica. I testi in esso inclusi sono per lo più di natura pubblicistica: materiale analitico, interviste, recensioni ecc.

Corpus della biblioteca Moškov[39] (Zacharov 2013). Corpus di 680 milioni di parole basato sul materiale della biblioteca online Moškov[40]. Si tratta di un corpus di dimensioni notevoli, pur essendo aggiornato solo all'anno 2006.

[37] http://www.ruscorpora.ru/new/search-murco.html
[38] http://h248.it.helsinki.fi/hanco/
[39] http://aot.ru/search1.html
[40] http://lib.ru/

Integrum World Wide[41]. Si tratta di un database che raccoglie testi in lingua russa di natura informativa, ricavati principalmente dai mass media di Russia, Ucraina e altri paesi dell'ex Unione Sovietica. I documenti contenuti sono oltre 40 mila ed è possibile svolgere anche ricerche di tipo morfologico. Il sito permette una prova gratuita di questo database, che, al termine di questa, sarà accessibile a pagamento.

Komp'juternyj korpus tekstov russkich gazet konca 20-go veka[42]**.** Creato dalla facoltà di filologia dell'Università Statale di Mosca, è un corpus di lingua russa giornalistica contemporanea per un totale di 11 milioni di parole. Il materiale giornalistico su cui è basato è stato ricavato da 13 giornali russi di diverse tipologie (quotidiani e non, letterari ecc.) pubblicati tra il 1994 e il 1997. È dotato di annotazione meta-testuale, grammaticale e lessicale.

Korpus Sintaktičeskich Kombinacij **(CoSyCo)**[43]. Corpus di co-occorrenze sintattiche che contiene informazioni sulle relazioni sintattiche delle parole russe.

Leeds University Corpora[44] (Zacharov 2013). È una raccolta di corpora di diverse lingue, tra cui il russo, basati su materiale web. Permette ricerche grammaticali e lessicali e anche di ottenere liste di collocazioni.

OpenCorpora[45] (Bocharov et al 2011). È un corpus di lingua russa liberamente accessibile, dotato di annotazione morfologica, sintattica e semantica. Il corpus è stato realizzato sulla base del concetto di *crowdsourcing*, ovvero è stato sviluppato collettivamente su invito dei creatori del progetto. Coloro che lo desiderassero hanno potuto prendere parte alla creazione del corpus nella raccolta del materiale. Sono contenuti in esso testi raccolti da fonti che ne permettono la libera ridistribuzione, come Wikipedia, WikiNews e Chaskor.ru, ma anche classici della letteratura russa disponibili in pubblico dominio su WikiSource.

OPUS[46]. È un corpus parallelo di testi tratti da 3 fonti, che sono i siti web di organizzazioni internazionali come l'Unione Europa o l'ONU, i manuali tecnici e la documentazione di software opensource e i corpora del web. Esso comprende inoltre diversi sub-corpora per testi raccolti da Wikipedia, i TED Talks e numerose altre fonti. Il corpus, dotato di annotazione linguistica, include

[41] http://www.integrumworld.com/index.html
[42] http://www.philol.msu.ru/~lex/corpus/
[43] http://cosyco.ru/
[44] http://corpus.leeds.ac.uk/internet.html
[45] http://opencorpora.org/
[46] http://opus.nlpl.eu/

più di 100 lingue tra cui il russo, ed è accessibile, oltre che dal sito ufficiale, anche da Sketch Engine.

RuTweetCorp. È il corpus della lingua russa di Twitter. Include 17,6 milioni di tweet, ovvero brevi testi che con poche parole devono essere in grado di esprimere un sentimento, che sia positivo e negativo (Rubtsova, Zagorulko 2014).

2.3.7 Altre risorse *corpus-based*

Častotnyj russkij slovar. Dizionario di frequenza creato da Zasorina tra gli anni Sessanta e Settanta.

Častotnyj slovar' sovremennogo russkovo jazyka (Kopotev et al 2016: 11). Dizionario basato sul materiale del NKRJa che riporta una lista dei 50 mila lemmi più frequenti della lingua russa. Permette di tenere traccia dei cambiamenti, nel tempo e nei registri, nella distribuzione delle parole. È inoltre possibile, con il suo utilizzo, ottenere liste di frequenza.

CoCoCo (Kopotev et al 2015). È un estrattore di espressioni multi-parola di vario tipo (lessemi multi-parola, collocazioni, colligazioni ecc.) dai corpora di lingua russa.

Dizionario di frequenza basato sul materiale di ruWac. Dizionario che contiene oltre 5000 lemmi.

Dizionario di frequenza lessico-grammaticale (Ljaševskaja 2013). Mostra la distribuzione delle forme grammaticali nel paradigma flessivo dei nomi, aggettivi e verbi russi.

FrameBank (Ljaševskaja, Kaškin 2015). Progetto spin-off del NKRJa. Si tratta di un database ad accesso libero che consiste in un dizionario di costruzioni lessicali russe ed un corpus, dotato di annotazione, dei loro usi. Gli esempi sono presi in maniera casuale dal NKRJa. *FrameBank,* che contiene circa 50 mila esempi, è incentrato principalmente sulle caratteristiche morfosintattiche delle costruzioni ed è per questo dotato di annotazione morfosintattica.

Grammatica di frequenza del russo. Riassume le categorie morfologiche della lingua russa più frequenti.

Statističeskij slovar' jazyka russkoj gazety. Dizionario della lingua russa giornalistica che include i 104 mila lemmi più frequenti nei giornali russi.

CAPITOLO 3
IL WEB COME CORPUS

Introduzione

In questo terzo capitolo introdurremo il concetto di web come corpus che ha iniziato ad emergere nella linguistica dei corpora a partire dal nuovo millennio, quando il computer era ormai entrato a far parte della vita quotidiana delle persone. Il web è così diventato, oltre che una tra le principali fonti di informazione, anche una preziosa fonte da cui ricavare dati linguistici in alternativa ai corpora tradizionali. Vedremo quindi le diverse accezioni di questo concetto e le varie modalità in cui il web può essere usato, anche grazie all'ausilio di altri strumenti linguistici, come risorsa linguistica in quanto fonte di testi per la creazione di corpora o addirittura in quanto corpus di per sé.

3.1 Il concetto di "Web as corpus"

Il web fu per la prima volta considerato come un corpus linguistico all'inizio del Duemila, quando Adam Kilgarriff (2001) scrisse l'articolo "The Web as Corpus". Il noto linguista computazionale evidenziò qui una connessione tra la linguistica dei corpora e il web e pronunciò una frase che fu all'epoca considerata alquanto controversa: "il corpus del nuovo millennio è il web" (Kilgarriff 2001: 345). Ma quali sono le ragioni per cui una frase simile è stata, e viene ancora oggi molte volte pronunciata? La risposta sta nel fatto che il web può essere visto come un'enorme collezione spontanea di testi autentici in formato elettronico, a cui chiunque può avere accesso gratuitamente e con un semplice click. Secondo molti autori infatti, il web rappresenta per il linguista una fonte potenzialmente illimitata di dati linguistici. Nonostante ciò, dal momento in cui questo paragone è stato azzardato, un'eterna diatriba si è accesa tra coloro che ritengono che il web possa effettivamente essere considerato un corpus e quelli che invece non sono d'accordo. Chi reputa corretta questa affermazione lo fa avendo in mente il significato originario del termine latino *corpus*, ovvero una raccolta di testi. Secondo questa definizione, teoricamente ogni collezione di testi può essere chiamata corpus. Chi la reputa errata fa invece riferimento alla moderna accezione del termine corpus, quella che abbiamo ben delineato nel primo capitolo. Infatti, il significato che il

termine corpus, o meglio corpus linguistico, ha acquisito negli ultimi anni nell'ambito della linguistica dei corpora, comprende delle caratteristiche specifiche, che il web non possiede. Le caratteristiche del web, che si discostano da quelle che dovrebbero essere le caratteristiche di un corpus, le illustreremo a breve. Ma l'elemento principale per cui, secondo molti linguisti, il web non è un vero corpus, è la mancanza della finalità linguistica di questo strumento, fondamentale invece nei corpora linguistici.

Prima di andare ad esaminare nel dettaglio le caratteristiche del web può essere utile vedere in breve quali sono i lati positivi e negativi dell'utilizzo di internet a finalità linguistiche. Questa premessa ci permette di inquadrare in linea generale il fenomeno del web nell'ambito della *corpus linguistics*. Ognuno di questi aspetti verrà poi approfondito nel corso del capitolo.

Partendo da quelli positivi abbiamo le dimensioni maggiori rispetto ai corpora tradizionali, il che significa risultati maggiori di una ricerca, in particolare per quanto riguarda i fenomeni linguistici più rari; il web copre più generi e registri linguistici, alcuni dei quali non sono presenti nei corpora tradizionali; i suoi contenuti risultano più aggiornati, permettendo così di trovare fenomeni linguistici non presenti nei corpora tradizionali in quanto troppo nuovi (Lüdeling et al. 2007); è una risorsa immediata, il cui funzionamento è conosciuto da tutti, per testare ipotesi linguistiche o intuizioni personali sull'uso della lingua; può essere usato come "evidenza qualitativa", ovvero provare che una certa forma o costruzione linguistica viene utilizzata in una certa lingua, permettendo di avere un'idea delle proporzioni in cui viene usata, ad esempio se più o meno rispetto ad un'altra, pur trattandosi di dati approssimativi; il web contiene moltissime, se non tutte, le lingue del mondo, alcune delle quali non dispongono ancora di un corpus rappresentativo del loro utilizzo. Passando ai lati negativi abbiamo il fatto che il web, in particolare i motori di ricerca, non siano pensati per trovare forme linguistiche ma contenuti (Ferraresi 2009: 2); le sue dimensioni sono incerte e in continua evoluzione, per cui impossibili da misurare; il suo contenuto non è organizzato ma anarchico e caotico e per lo più sconosciuto; non dispone di alcune caratteristiche fondamentali dei corpora tradizionali, come lemmatizzazione e annotazione morfosintattica, e dei loro principi fondamentali come attendibilità, autorevolezza, rappresentatività; contiene errori dovuti al fatto che si tratta di materiale non controllato prodotto da utenti le cui conoscenze grammaticali potrebbero essere limitate.

Nell'ultimo decennio l'attenzione si è probabilmente concentrata più sui lati positivi di questo strumento che su quelli negativi, dal momento che il web è sempre più utilizzato come fonte di evidenza linguistica. In particolare, il potenziale dei motori di ricerca sta ricevendo sempre più il

supporto e l'attenzione da parte del *Computer Aided Language Learning*[47], nell'insegnamento delle lingue straniere e anche nella ricerca accademica nell'ambito della *corpus linguistics*. Come ci spiega Gatto (2014: 41), questo non vuol dire che i principi cardine tradizionali dei corpora siano stati messi in discussione, si tratta solamente di un incontro tra posizioni teoretiche tradizionali e nuovi strumenti e metodi in risposta ad esigenze pratiche. Ciò che ci si augura per il futuro è una coesistenza e interazione tra i due strumenti in modo da ampliare le risorse a disposizione della comunità linguistica. È doveroso però sottolineare che le differenze tra un corpus linguistico e il web dal punto di vista qualitativo rimangono comunque insormontabili. Tuttavia, non si può affermare in maniera assoluta che il web non sia un valido strumento per la ricerca linguistica, ma va anche considerato il tipo di ricerca che si deve svolgere.

Ora è venuto il momento di analizzare uno per uno alcuni dei principi cardine della linguistica dei corpora (messi in evidenza da Gatto 2014), che abbiamo già visto nel primo capitolo, ma che rivedremo qui dal punto di vista del web, dove sono spesso messi in discussione.

3.1.1 Autenticità

Una delle principali caratteristiche dei corpora linguistici, come dice la stessa definizione, è quella di contenere testi autentici, in una lingua spontanea ed effettivamente prodotta ed utilizzata dai suoi parlanti. I testi contenuti nel web, come abbiamo già sottolineato, sono proprio testi autentici, nati da interazioni umane. L'autenticità è l'elemento che forse maggiormente avvicina il web ai corpora linguistici, rendendolo quindi un buon terreno per il linguista. Ma se per un corpus linguistico l'autenticità è una caratteristica assolutamente positiva, per il web è anche un grande difetto. In questo caso, di fatto, autentico vuole anche dire inaccurato. Questo è dimostrato dal fatto che il web prolifera di errori ortografici, grammaticali, usi impropri della lingua da parte di parlanti non nativi, ma anche di nativi che utilizzano la grammatica in maniera errata sia per scarsa conoscenza, sia per un uso gergale della lingua, che vede spesso infrante alcune regole grammaticali. È proprio qui che la mancanza di autorevolezza dei testi si fa più sentire. Ovviamente, anche il web è ricco di testi autorevoli, come articoli di giornale, testi scientifici, accademici, letterari e molti altri. Il web è però altrettanto ricco di testi non autorevoli come blog, post dei social network e chat. Ed è proprio la lingua di questi testi che spesso si dimostra inattendibile nel caso di una ricerca linguistica. Lo dimostra il fatto che se effettuiamo con Google la ricerca della parola russa

[47] Utilizzo delle nuove tecnologie per l'insegnamento delle lingue straniere

“достопримечательности” scritta in maniera errata, ovvero “достапримечательности” (l’errore è dovuto alla riflessione nella grafia del fenomeno fonologico dell’*akan’e* per cui la vocale /o/ in posizione atona si pronuncia /a/), il motore di ricerca rileva ben 14100 risultati[48]. Si tratta comunque di un risultato poco significativo essendo questa una parola molto frequente, tanto che scritta in maniera corretta i risultati sono 89 milioni. Non lascia quindi dubbi sul fatto che, in caso di una ricerca su quale delle due forme sia corretta, la prima sia quella che fa al caso nostro, ma questo ci dice molto sull’autorevolezza dei testi contenuti nel web.

3.1.2 Rappresentatività

Questo è uno dei concetti più spinosi nella prospettiva del web come collezione di testi simile ad un corpus. Se da una parte le sue grandi dimensioni e la sua vasta inclusività di generi e registri testuali lo rendono uno strumento eterogeneo e potenzialmente rappresentativo, dall’altra è una risorsa non bilanciata e quindi potenzialmente non rappresentativa, soprattutto per il fatto che non si conosce in che proporzione ogni genere e registro è contenuto. Inoltre, nonostante la presenza di un’ampia gamma di generi testuali al suo interno, alcune aree del linguaggio sono poco rappresentate. Mentre con l’affermazione dei social media c’è l’emergere di quel linguaggio a metà tra scritto e parlato di cui abbiamo parlato nel capitolo precedente, ovvero la *Computer-Mediated Communication*, rimane comunque poco rappresentato il linguaggio quotidiano delle conversazioni private, telefoniche e così via, quel linguaggio puramente parlato che oggi si va invece a fondere con quello scritto. Per questo motivo, alcuni linguisti come Leech (2007) ritengono che il web non sia un campione rappresentativo della lingua nell’accezione più generale del temine. Secondo il linguista, quindi, solo un corpus creato con un’attenta selezione dei testi può essere considerato rappresentativo. Dall’altra parte, però, abbiamo anche chi ritiene che il web sia invece da considerarsi rappresentativo proprio per la sua varietà di generi testuali e per il fatto di essere un campione di utilizzo di una lingua in tutto il mondo (o comunque nelle aree in cui una lingua è parlata), come Henzinger e Lawrence (2004). Mettendo d’accordo queste due visioni Gatto sostiene che malgrado il web non possa essere considerato come un campione di lingua come lo è un corpus, la sua varietà e dimensione contribuiscono a controbilanciare i limiti della sua rappresentatività. Comunque, al di là della mancanza di rappresentatività del web, esso rimane una fonte preziosa di dati linguistici dal punto di vista della *corpus linguistics*.

[48] Ricerca effettuata in data marzo 2020

3.1.3 Dimensioni

Le imponenti dimensioni del web giocano sicuramente un punto a favore rispetto ai corpora tradizionali. Una frase che ci fa rendere conto di quanto esso sia grande è stata scritta dal linguista David Crystal (2011: 10): "non è mai esistito un corpus linguistico così grande, che contiene più lingua scritta di tutte le librerie del mondo messe insieme". Questo permette, durante una ricerca linguistica, di acquisire molta più evidenza. La linguistica dei corpora, però, come abbiamo evidenziato nel primo capitolo, è una disciplina scientifica, e, come tutte, si basa su dati quantitativi, la cui fonte, che in questo caso è il corpus, deve avere dimensioni finite e definite. Tuttavia, le dimensioni del web sono tutt'ora sconosciute. Alcuni tentativi di quantificare le pagine web sono stati fatti, ma questi non sono stati in grado di fornire risultati che non fossero pure stime. Si pensi al fatto che di minuto in minuto, se non di secondo in secondo, il web si evolve con l'aggiunta di nuove pagine e l'eliminazione di altre, segno di una sua natura estremamente dinamica. Al momento si può solo parlare di miliardi di pagine, senza sapere esattamente quanti. È quindi improbabile che una fonte di dimensioni sconosciute fornisca dati esatti che siano poi alla base di una ricerca oggettiva e quantitativa, ma ci darebbe solo risposte approssimative. Questo rappresenta quindi un grande limite nell'ambito della *corpus linguistics*, che mette anche in discussione la ripetibilità di un esperimento, concetto cardine della scienza.

Nonostante sia stato tentato da alcuni linguisti il paragone tra il web e un monitor corpora di ultima generazione (Lüdeling et al. 2007), che per definizione è in continua crescita ed evoluzione, questo dovrebbe però crescere in maniera controllata e stabile. Ciò comunque non toglie il vantaggio di essere costantemente aggiornato seguendo in tempo reale l'evolversi della lingua, il ché ci dà la possibilità di analizzare fenomeni linguistici tra i più recenti.
Se quindi ai suoi albori la *corpus linguistics* sosteneva il concetto "più grande è meglio è", come scrisse Sinclair (1991: 18), questo è stato leggermente rivalutato per quanto riguarda il web. Le grandi dimensioni sono qualcosa di positivo purché non siano spropositate e incontrollabili. Rimane pur sempre vero, però, che queste permettono di reperire più informazioni rispetto ai tradizionali corpora.

3.1.4 Composizione

Cosa sia esattamente contenuto in rete è impossibile da definirsi. Basti pensare che le pagine web sono scritte in qualunque lingua, dialetto, stile e registro. Nonostante la ricchezza del suo materiale,

questo rappresenta per internet un limite dal punto di vista della *corpus linguistics*. Più volte, infatti, è stato utilizzato l'aggettivo "anarchico" per definire il web e il contenuto. A questo proposito, sono state individuate da Gatto quattro categorie che generalmente vanno a definire il contenuto di un corpus, ovvero medium, lingua, tema e genere/registro. Vediamo a cosa questi corrispondono per quanto riguarda il web.

Medium. Come abbiamo già più volte sottolineato, il web contiene la lingua scritta, la lingua parlata, ma anche una lingua a metà tra le due. Il limite tra i due mezzi, qui, è sempre più sottile, tanto che persino il linguista Crystal (2011: 20) si è chiesto se la lingua di internet si avvicini più alla lingua scritta, a quella parlata o se sia qualcosa di totalmente differente. Le chat, i blog, e i commenti dei social network ne sono un esempio. Questa caratteristica ovviamente, e lo abbiamo già anticipato nella sezione riguardante l'autenticità del materiale web, non appartiene a tutta la lingua di internet. I testi più "ufficiali" sono infatti una parte molto significativa del web e inoltre compaiono con più rilevanza, tra i risultati di una ricerca, rispetto a post e chat, i quali si trovano spesso in parti più nascoste del web e tendono ad emergere di meno.

Lingua. La lingua franca del web è indiscutibilmente l'inglese. Ma questa, naturalmente, non è l'unica. Nel web sono presenti testi in tutte le lingue del mondo, per molte delle quali non è ancora disponibile un vero e proprio corpus. Ma oltre alle lingue standard, sia le principali che quelle meno diffuse, si trovano anche molte varietà linguistiche e regionalismi. Il multilinguismo è di fatto una delle grandi ricchezze dell'internet. Il sito www.internetworldstats.com mostra il grafico delle 10 lingue più utilizzate nel web in base al numero dei loro utenti (Fig. 3). La lingua russa rientra tra queste al nono posto. Nonostante l'inglese superi ancora di molto le altre lingue, assieme al cinese che lo avvicina molto, queste sono in costante crescita. Questo è possibile grazie ai siti di informazione e notizie, ai siti ufficiali del governo, all'enciclopedia Wikipedia e alle pagine personali come blog e social network. Tutte queste fonti contribuiscono ad ampliare le lingue rappresentate dal web, rendendolo uno strumento prezioso a salvaguardia delle lingue minori o addirittura a rischio. È proprio grazie al materiale presente nel web che sono stati recentemente creati corpora linguistici per lingue come Swahili, Lettone, Basco e molte altre, reperibili su piattaforme corpus-manager come Sketch Engine.

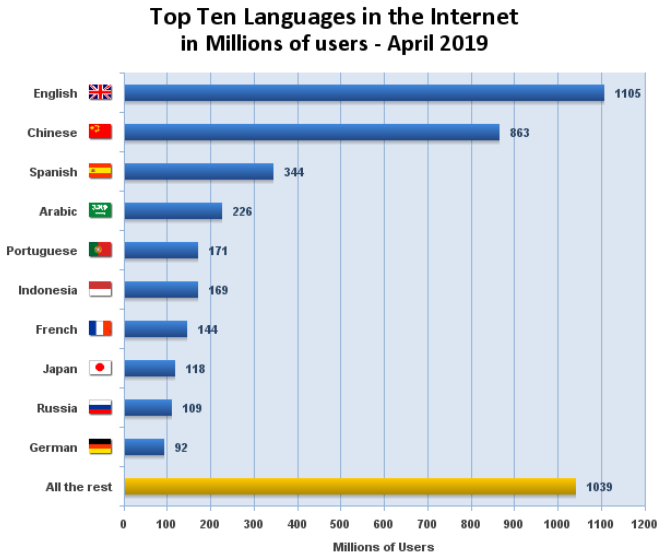


Figura 3. Top-ten lingue del web aggiornata ad aprile 2019

Tema. Nonostante sia stato più volte effettuato il tentativo di dividere il web in categorie tematiche, questo risulta molto difficile da realizzarsi proprio per la natura caotica e anarchica del web e per il suo contenuto sempre in evoluzione. Includere nelle varie categorie tutti i testi presenti nel web sarebbe praticamente impossibile. I motori di ricerca che l'hanno tentato, come ad esempio *Yahoo!*, hanno realizzato una categorizzazione in via puramente rappresentativa, utile sia a indirizzare l'utente verso pagine di suo interesse, restringendo di molto la ricerca, sia a creare corpora ad hoc in base ad un argomento specifico, come vedremo in seguito.

Genere/registro. La distinzione di questi elementi è fondamentale per un corpus dal momento che, perché esso risulti ben bilanciato, deve contenere in maniera equilibrata testi dei diversi generi e registri testuali. È proprio l'impossibilità di effettuare una tale quantificazione dei suoi testi a rendere il web uno strumento meno valido in termini di rappresentatività.

3.1.5 Attendibilità dei risultati

La precisione (*precision*) e il recupero (*recall*) sono due principi statistici alla base dell'*information retrieval*[49], nel nostro caso per quanto riguarda il recupero del materiale dal web. Nel momento in cui si effettua una ricerca, che sia in un corpus tradizionale o tramite un motore di ricerca, i suoi risultati hanno sempre un certo grado di precisione e di recupero. La precisione è massima se la

[49] insieme delle tecniche utilizzate per gestire la rappresentazione, la memorizzazione, l'organizzazione e l'accesso ad oggetti contenenti informazioni quali documenti, pagine web, cataloghi online e oggetti multimediali (Wikipedia)

ricerca ci restituisce solo risultati che corrispondono esattamente alla nostra domanda, e massimo è il recupero se tutti i risultati corrispondono alla nostra domanda. Precisione e correttezza dei dati statistici sono fondamentali nelle indagini quantitative, mentre nelle ricerche di tipo qualitativo non è fondamentale che il livello di precisione sia alto (Lüdeling et al. 2007). Per quanto riguarda i risultati delle ricerche svolte con l'utilizzo del web, un alto livello di recupero è compromesso dalla sua natura instabile e dinamica dato che si hanno pochi risultati corretti (falsi negativi), mentre un buon grado di precisione è compromesso dai limiti dei motori di ricerca ordinari, che non sono nati con finalità di ricerca linguistica, poiché restituiscono troppi risultati errati (falsi positivi). Questo può derivare da diversi fattori legati alle caratteristiche proprie dei motori di ricerca, che vedremo a breve, come ad esempio i numerosi duplicati delle pagine web o la normalizzazione della *query*[50] che avviene allo scopo di fornire più risultati possibili all'utente. Queste funzionalità possono essere utili nel momento in cui si effettua una ricerca nel web a scopo informativo, ma non quando lo scopo è quello linguistico, dal momento che vanno a compromettere l'attendibilità dei risultati.

3.1.6 Metadata e annotazione

I testi presenti nel web, al contrario di quelli che costituiscono i corpora linguistici, non sono ovviamente dotati né di annotazione meta-testuale, né tanto meno di quella morfosintattica, che, stando alla definizione moderna di corpus, sappiamo essere elementi ormai imprescindibili. Per lo meno i dati meta-testuali, però, da alcune tipologie di testi possono essere ricavati manualmente. Gli articoli di giornale, ad esempio, contengono quasi sempre il nome dell'autore, che ne fornisce anche il sesso; allo stesso modo i blog, oltre al nome, spesso contengono anche la loro età e provenienza, così come i post dei social network; per i testi letterari è possibile ricavare i dati dell'autore da siti enciclopedici. Ovviamente, non essendo dati appositamente implementati non saranno reperibili in maniera così immediata come avviene invece in un corpus, in cui è possibile svolgere automaticamente ricerche in base al sesso, all'età e alla provenienza dell'autore per tutti i suoi testi.

[50] Termine tecnico utilizzato in informatica per indicare l'interrogazione di un database da parte di un utente

3.2 Quattro approcci all'utilizzo del web come corpus

Ci sono diverse modalità con cui il web può essere utilizzato come un corpus linguistico. A questo proposito, il vantaggio del web è quello di essere una fonte di testi già in formato elettronico che si prestano a studi di tipo linguistico, sia sotto forma di pagine web scaricabili come fonte per la creazione di corpora offline, sia tramite motori di ricerca come fonte di dati linguistici online. Questo ci darà la possibilità di dimostrare come l'utilizzo di alcuni strumenti e software di supporto, più o meno indirizzati alla ricerca linguistica, permettano di superare alcuni dei limiti propri del web che abbiamo finora introdotto e di sfruttarlo al meglio dal punto di vista linguistico. Gli approcci che andremo ora ad illustrare sono stati proposti per la prima volta da Baroni e Bernardini (2006: 10-14) e sono stati poi spesso ripresi negli anni a seguire nella letteratura della *corpus linguistics.* Introduciamo qui le quattro modalità di utilizzo del web come corpus con una breve definizione, dopodiché le vedremo nel dettaglio una ad una con l'illustrazione del funzionamento dei vari strumenti e software di supporto e con delle dimostrazioni pratiche.

1. Web come *corpus surrogate:* si tratta dell'utilizzo del web a scopi linguistici tramite i motori di ricerca commerciali, come Google, Yahoo, Yandex, oppure attraverso software finalizzati all'analisi linguistica che permettono di accedere al web come se fosse un corpus, come ad esempio WebCorp.
2. Web come *corpus shop:* secondo questo approccio l'utente seleziona ed effettua il download di testi dal web per la creazione di corpora specializzati "monouso", a scopi traduttivi o terminologici, in maniera semi-automatica tramite software come BootCat.
3. Web come *corpus proper:* con questo approccio è possibile guardare al web come ad un corpus rappresentativo della lingua del web, ovvero la *Computer Mediated Communication*. Il web diventa quindi una sorta di corpus specializzato per indagare le caratteristiche tipiche di questa lingua.
4. *Mega-corpus mini-web:* questo approccio rappresenta il tentativo di creare un nuovo oggetto (un mega-corpus o un mini-web), adattato alla ricerca linguistica, che combina caratteristiche del web (grande, aggiornato, con interfaccia simile al web ecc.) con quelle dei corpora (annotazione, stabilità, ricerche complesse ecc.)

In base a questi quattro approcci è possibile fare una distinzione tra due concetti: *web as corpus* e *web for corpus* (De Schryver 2002). Il primo vede il web stesso come corpus (approccio 1 e 3) mentre il secondo vede il web come fonte per ricavare il materiale con cui costruire altri corpora (approccio 2 e 4).

Nell'analisi pratica di questi approcci che seguirà non ci soffermeremo sul terzo approccio, ovvero quello di web come *corpus proper*, in quanto lo scopo di questo lavoro non è tanto analizzare la lingua del web ma piuttosto mostrare come poter sfruttare il web a scopi di ricerca linguistica.

3.2.1 Il web come *corpus surrogate*

I due strumenti che permettono di utilizzare il web secondo questo approccio sono: i motori di ricerca commerciali, ovvero gli strumenti che gli utenti di qualsiasi categoria utilizzano quotidianamente per andare a reperire informazioni contenutistiche nel web, e i software volti alla ricerca linguistica che sono utilizzati da una fascia specifica di utenti, ovvero generalmente linguisti o traduttori. Vedremo ora in che modo questi due strumenti, che si basano esclusivamente su materiale reperito dalla rete, permettono di svolgere analisi di tipo linguistico.

3.2.1.1 Motori di ricerca commerciali

Per capire il funzionamento dei motori di ricerca commerciali è opportuno partire dalla definizione di motore di ricerca in generale. Il motore di ricerca è un programma pensato per aiutare le persone ad individuare informazioni nel web formulando una semplice ricerca con parole chiave (Oja, Pearson 2012). Esso connette le parole chiave dell'utente alle parole chiave delle pagine web, ovvero gli indici che ne sintetizzano il contenuto in una forma simile ad un titolo. Tali indici vengono infine mostrati all'utente seguiti da un'anteprima del testo. La ricerca tramite motore di ricerca avviene in tre fasi operative: 1) l'analisi (*crawling*), nella quale viene scansionato il web e vengono individuati i siti nuovi o recentemente modificati; 2) l'indicizzazione (*indexing*), in cui per ogni pagina individuata ne viene analizzato il contenuto, che viene poi indicizzato estraendo le parole contenute nel testo, nel titolo e nelle varie voci. In questo indice sono presenti per lo più parole chiave, mentre le parole più frequenti e le *stop-words*, come articoli o preposizioni, vengono escluse. Tutti i dati recuperati ed elaborati in questa fase vengono poi archiviati in un database degli indici per le future ricerche; 3) la risposta (*searching*), che avviene nel momento in cui l'utente inserisce la sua *query* e, dopo aver analizzato gli indici presenti nell'archivio, il sistema restituisce quelli che corrispondono alla sua ricerca.

Dopo averne analizzato il meccanismo alla base, andiamo a vedere prima di tutto cosa significa l'aggettivo "commerciale", e a seguire quali sono le funzionalità che i motori di ricerca offrono

all'utente. È bene sottolineare, però, che le funzionalità che vedremo sono state pensate per rendere la ricerca migliore dal punto di vista del contenuto e non dal punto di vista linguistico. Come abbiamo già anticipato, infatti, tali strumenti non nascono con finalità linguistiche. Nonostante molte di queste funzionalità ostacolino una qualsivoglia ricerca linguistica con l'utilizzo dei motori di ricerca, alcune opzioni, seppur implementate con altre finalità, permettono di effettuare una ricerca simile a quella che si svolgerebbe tramite un corpus.

Alcuni motori di ricerca vengono definiti "commerciali" per un motivo ben preciso. Questi "hanno finalità di lucro e attingono profitto dalle informazioni che l'utente stesso rilascia all'interno del sito quando fa ricerche e visita determinate pagine. Queste informazioni saranno poi trasmesse alle aziende che, a pagamento, orienteranno un certo tipo di pubblicità verso il cliente"[51]. Quindi, nel momento in cui si effettua una ricerca tramite un motore di ricerca commerciale, i risultati non sono oggettivi ma personalizzati, verranno cioè mostrate, in base ad algoritmi di rilevanza, le pagine che corrispondono all'uso, alle abitudini e alle ricerche pregresse dell'utente. Questo meccanismo è sicuramente utile nell'utilizzo quotidiano del web ma ha due conseguenze fondamentali dal punto linguistico: i risultati di una ricerca saranno diversi per ogni utente, anche in base al momento e al luogo geografico in cui questa viene effettuata, andando a violare il principio scientifico della ripetibilità di un esperimento; inoltre, dal momento che i risultati di una ricerca vengono mostrati in numero limitato, questi non saranno dati oggettivi, quali devono essere dei dati scientifici, ma dati rilevanti in base ad un principio esterno rispetto alla finalità linguistica della ricerca.

Dal punto di vista della *query*, il motore di ricerca effettua un processo di normalizzazione dello *spelling* per cui non verranno distinte caratteristiche dell'ortografia quali lettere maiuscole, parole scritte separatamente o unite da un trattino, punteggiatura ecc. (Rosenbach 2007) in modo da fornire più risultati possibili. Questo va ovviamente a scapito della ricerca del linguista. Anche la mancanza di lemmatizzazione ed annotazione si fa senza dubbio sentire, rendendo quasi impossibile la ricerca di specifici elementi grammaticali permessa invece dai corpora linguistici.

Oltre a tutti questi svantaggi, i motori di ricerca hanno però anche alcuni lati positivi, ovvero alcune caratteristiche od opzioni che li avvicinano per qualche aspetto ai corpora. In primo luogo possiamo dire che essi siano uno tra gli strumenti più immediati e popolari per testare ipotesi linguistiche o per comprovare dubbi di natura linguistica, come ad esempio la correttezza o la frequenza d'utilizzo di una forma rispetto ad un'altra. Un secondo elemento, che rende il web più simile ad un corpus, riguarda il modo in cui i risultati vengono presentati all'utente, ovvero sotto forma di stringhe di

[51] Fonte: https://www.unidlab.com/motori-di-ricerca/

testo simili a delle concordanze in cui vediamo un certo numero di occorrenze di una parola o un gruppo di parole nel loro contesto, con una vaga indicazione della loro frequenza (Gatto 2014: 79). Infine, i motori di ricerca prevedono una serie di opzioni che permettono di effettuare anche ricerche complesse, sia dal punto di vista linguistico che meta-testuale. La qualità dei risultati che si ottengono da una ricerca linguistica tramite internet nella maggior parte dei casi dipende anche dalla capacità dell'utente di formulare una *query* appropriata, che possa avvicinarsi ad una ricerca effettuata tramite un corpus. Per questo motivo andiamo ad analizzare le opzioni di ricerca più avanzata di due motori di ricerca, Google, che è il più popolare al mondo, tra cui anche nel nostro paese, e Yandex, che è invece il più popolare in Russia. Capire bene come questi funzionano è il primo passo per una ricerca più efficace dal punto di vista linguistico.

Opzioni di ricerca di Google

Nonostante, come già sottolineato, le pagine web non sono dotate annotazione meta-testuale (né tanto meno morfosintattica), come invece accade nei corpora, alcune opzioni di ricerca avanzata disponibili nel motore di ricerca Google permettono una sorta di filtraggio dei risultati in base ad alcuni elementi. Vediamo quali sono in questa immagine (Fig. 4).

Figura 4. Ricerca avanzata di Google

- Lingua: è possibile ricercare i risultati in una o in tutte le lingue tra le 40 disponibili.
- Area geografica: per ricercare pagine web appartenenti ad una determinata area geografica è possibile selezionare un dominio specifico in cui effettuare la ricerca, che può essere .it per quanto riguarda l'Italia, .ru per la Russia, .uk per l'Inghilterra e così via. Questo permette anche di esplorare le diverse sfumature di una lingua, ad esempio l'inglese britannico e americano oppure il russo parlato in Russia o in Ucraina.

- Ultimo aggiornamento: permette di selezionare la data di ultima modifica di un file, che va dalle ultime 24 ore all'ultimo anno, oppure una data qualsiasi. Questo può rendere il web una sorta di corpus diacronico, anche se con molte limitazioni.
- Sito o dominio: è possibile svolgere una ricerca all'interno di siti specifici recandosi nell'apposita sezione della ricerca avanzata (come in foto) e incollando il link nello spazio corrispondente alla voce sito o dominio, oppure inserendo nella *query* la dicitura "site:" seguito dall'URL del sito desiderato. Inoltre, inserendo ad esempio domini come .edu o .gov è possibile ricercare solamente pagine in ambito rispettivamente accademico o governativo, il che garantisce tendenzialmente una lingua ufficiale, autorevole e priva di errori. Se invece si vuole che la ricerca avvenga, ad esempio, solamente nell'Enciclopedia Wikipedia, il dominio da inserire sarà wikipedia.org.
- Posizione dei termini: Google permette di selezionare la posizione in cui si vuole che i termini ricercati compaiano all'interno della pagina, come ad esempio nel titolo, nel testo o in un punto qualsiasi, per evitare che titoli o link indesiderati compromettano la ricerca.
- Tipo di file: ci viene data la possibilità di selezionare il tipo di file in cui è contenuta la nostra *query* tra pdf, doc, ppt e altri. Anche in questo caso tale opzione è una sorta di garanzia di correttezza del linguaggio in quanto, solitamente, documenti di questo tipo sono più autorevoli rispetto a pagine come blog, e si presuppone che sia più corretto dal punto di vista ortografico e grammaticale.

Per quanto riguarda la ricerca dal punto di vista linguistico, vediamo ora quali opzioni avanzate sono offerte dal motore di ricerca Google, sempre nella sezione "ricerca avanzata" (Fig. 5).

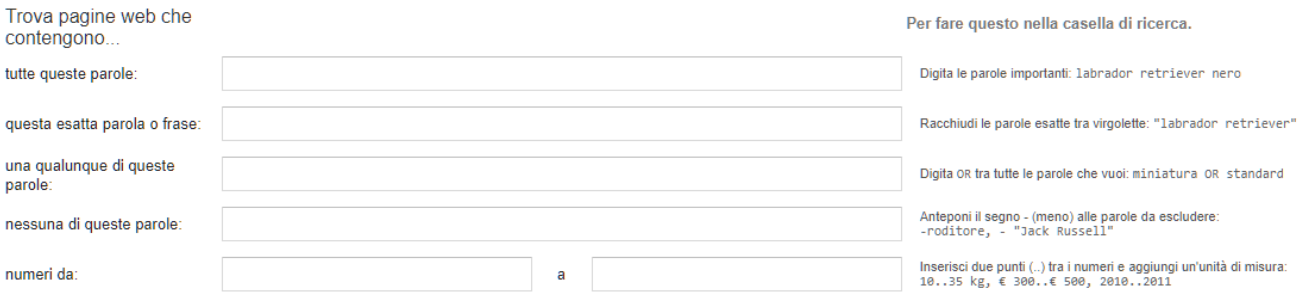


Figura 5. Ricerca avanzata di Google

Il testo della nostra ricerca, ovvero la *query*, può essere composto da una o più parole. È possibile effettuare una ricerca che vada a recuperare tutte le parole presenti nella *query*, in qualsiasi ordine e

normalizzate, oppure ricercare una parola o una frase esatta, con un preciso ordine e una precisa ortografia. Inoltre possono anche essere recuperate pagine che contengano una qualunque tra una serie di parole. Ancora, da una *query* si possono escludere determinate parole che non vogliamo compaiano nelle pagine recuperate, e infine, si può ricercare una serie di numeri che dovrà essere presente nei risultati ottenuti.

Funzioni molto simili a queste, oltre che dalla sezione "ricerca avanzata" di Google, possono essere sfruttate anche direttamente dalla barra di ricerca grazie ai cosiddetti "operatori booleani", da integrare al testo della nostra ricerca. Questi operatori, infatti, permettono di identificare pagine web che contengono una particolare combinazione di parole. Essi sono:

- + per cercare un gruppo di parole che devono essere tutte presenti (in qualsiasi ordine), ad esempio "чёрный + красный + зелёный + белый"
- OR per cercare un gruppo di parole che possono essere presenti o meno. Ad esempio "*кот OR собака"*
- - per escludere dai risultati determinate parole. Ad esempio "машина -автомобиль" mi permette di ottenere pagine in cui la parola "машина" abbia principalmente il significato di macchina di altro tipo, e non automobile.
- "" per cercare una combinazione esatta di parole. Ad esempio "отплатить за услугу"
- * per cercare una frase in cui manca un elemento, il quale viene sostituito dall'asterisco (funziona meglio se l'intera frase viene messa tra virgolette). Ad esempio, avendo in mente la frase idiomatica "быть в своей тарелке", ovvero "sentirsi a proprio agio", abbiamo ricercato la frase ""быть в своей *"", per vederne le possibili varianti. Dai risultati abbiamo visto che, oltre a "тарелке" si trova spesso accompagnato anche dalla parola "стихии", con lo stesso significato. Altro esempio, per sapere quale verbo russo si utilizza per esprimere la frase "il paziente ha subito un'operazione" oppure "il paziente è stato sottoposto ad un'operazione" basterà esprimere una *query* di questo tipo: "пациент * операция" e il motore di ricerca proporrà frasi in cui tra le parole "пациент e операция" (che si troveranno anche in casi diversi dal nominativo inserito nella ricerca) compaiono altri elementi. Alcuni dei risultati trovati sono: "пациенту, перенесшему операцию"; "пациентов, перенёсших операцию"; "пациенты, перенесшие операцию"; "пациенту, проведя операцию"; "пациенту провели уникальную операцию"; "пациенту сделали операцию; пациентов провели операции" ecc. Da questi risultati possiamo quindi dedurre che i verbi più utilizzati in questo caso sono: "перенести" se è il paziente a subire l'operazione, "провести" o "сделать" se sono i medici ad aver eseguito l'operazione sul paziente.

Opzioni di ricerca Yandex

Yandex è il motore di ricerca russo più popolare. I suoi operatori, che sono diversi da quelli di Google, sono creati proprio per adattarsi alla lingua russa che è ricca di grammatica e sintassi. Vediamo quali sono[52]:

- ! davanti ad una parola permette di ricercare la parola in quella forma precisa, mantenendo caso, numero e tempo verbale. Ad esempio: "купить !собаку" per trovare la parola "собака" al caso accusativo
- + davanti alle *stop-words* permette di trovare una parola accompagnata da un determinato articolo o preposizione o altri di questi elementi. Ad esempio, possiamo ricercare "работа +на дому" per trovare la combinazione di parole *работа* e *дом* unite dalla preposizione "на".
- "" per trovare una corrispondenza esatta, come in Google
- [] per mantenere l'ordine delle parole di una frase. Ad esempio билеты "[из москвы в париж]"
- | tra due parole se si vuole ricercare una delle due
- () per effettuare ricerche più complesse come "купить машину (недорого|ВАЗ)" per trovare la collocazione "купить машину" accompagnata da una delle due parole tra parentesi.

Con la conoscenza di queste opzioni si riusciranno ad effettuare ricerche complesse tramite i motori di ricerca anche dal punto di vista linguistico. Secondo Gatto, il settaggio dei parametri meta-testuali può addirittura essere visto come la creazione di un sub-corpus temporaneo dal web. L'esperta ci spiega inoltre che, nonostante gli innumerevoli limiti del web, queste funzionalità possono renderlo meno anarchico e meno inospitale per la ricerca linguistica e permettono di dimostrare che può essere una valida fonte di evidenza, sia qualitativa che quantitativa, di uso della lingua. Ovviamente, è necessario prestare attenzione ad elaborare una *query* corretta ma soprattutto interpretare in modo adeguato i risultati, senza dare niente per scontato. Un banale esempio di presa di coscienza dei risultati ottenuti dal web è il fatto che, essendo una fonte di dimensioni così imponenti, se una ricerca fornisce poche centinaia e a volte anche poche migliaia di risultati, significa che potrebbe trattarsi di errori.

[52] Fonte: https://yandex.com/support/direct/keywords/symbols-and-operators.html

Le tipologie di ricerca che più il web, in particolare i motori di ricerca, si prestano a fare sono tre: ricerca di frasi e collocazioni, ricerca di strutture fraseologiche e test dei candidati di una traduzione (Gatto 2014: 88).

Ricerca di frasi e collocazioni. La ricerca di frasi è sicuramente più valida rispetto a quella di singole parole, che, in una fonte così ampia, può essere fuorviante. Grazie alle sue dimensioni il web fornisce una grande quantità di elementi come collocazioni e frasi idiomatiche, cosa che solitamente i corpora tradizionali non fanno, o fanno in maniera molto limitata, per le sue dimensioni più contenute. Grazie all'operatore booleano delle virgolette che abbiamo visto sopra, ad esempio, è possibile verificare se una certa sequenza di parole esiste in una lingua. In tutto ciò bisogna comunque tenere conto del rumore che una fonte di dimensioni tali contiene, tanto che, più che a validare una teoria, il web si presta più ad invalidarla.

Ricerca di strutture fraseologiche. Questo viene reso possibile sia dall'operatore delle virgolette, per verificare la correttezza o l'esistenza di una struttura fraseologica, che da quello dell'asterisco, per verificare, all'interno di una di queste strutture, quali sono gli elementi plausibili nelle diverse posizioni della frase.

Test dei candidati di una traduzione. Questa operazione è possibile grazie all'evidenza d'utilizzo della lingua che il web fornisce all'utente, in questo caso il traduttore. Ciò richiede ovviamente, da parte del traduttore, un'analisi dei risultati per verificare ad esempio se il contesto sia quello da lui ricercato, se l'indicazione della frequenza indichi un'effettiva evidenza o se si tratti di errori, e tutte quelle accortezze necessarie a verificare l'attendibilità di una ricerca.

3.2.1.2 Software linguistici

Esistono alcuni software pensati per rendere il web più utile dal punto di vista della ricerca linguistica. Più precisamente, questi rendono più facile formulare in un motore di ricerca una *query* che sia appropriata dal punto di vista linguistico e manipolano i risultati così da renderli più adeguati all'analisi linguistica (Gatto 2014:105). Strumenti di questo tipo si sono dimostrati efficaci nel restituire al linguista dati utili, specialmente nell'ambito dell'insegnamento e in quello della ricerca di frasi, neologismi, e termini rari o obsoleti (Fletcher 2007). Nonostante tutti i vantaggi che i motori di ricerca possono offrire per l'analisi linguistica, molti, e li abbiamo visti, rimangono quelli negativi, che non è possibile ignorare. Sono stati proprio questi limiti a far nascere la volontà di creare strumenti che potessero sfruttare il potenziale del web, le sue dimensioni e la sua ricchezza

per la ricerca linguistica. Strumenti come questi, che vedremo nel corso di questo capitolo, sono di diverso tipo e rientrano in vari dei quattro approcci all'utilizzo del web come corpus. Alcuni degli aspetti in cui si differenziano riguardano il modo in cui recuperano i dati dal web, il loro grado di dipendenza dai motori di ricerca, la stabilità e la verificabilità dei risultati (Gatto 2014: 106). Ci sono ad esempio quelli che fungono da intermediario tra i bisogni del linguista e i motori di ricerca, ma anche quelli totalmente indipendenti da essi, che vanno autonomamente a recuperare il materiale nel web al fine di costruire corpora offline (Lüdeling et al. 2007: 16).

In questa sezione andremo ad analizzare il software che rientra nell'approccio di cui stiamo trattando, ovvero quello di web come *corpus surrogate*. Si tratta di uno di quegli strumenti che abbiamo detto essere strettamente legati ai motori di ricerca. Ciò significa che il materiale sarà recuperato dal motore di ricerca, secondo le sue funzionalità tipiche che abbiamo visto, ma verrà poi rielaborato dallo strumento in questione, che offre molte opzioni e funzioni utili dal punto di vista della ricerca linguistica.

Il software di cui siamo parlando è WebCorp[53], nato negli anni Novanta proprio per testare l'ipotesi che il web potesse essere utilizzato come ampio corpus di testi per studi linguistici (Morley 2006: 283). Si tratta di una sorta di *concordancer*, che è una delle funzioni principali dei corpora tradizionali. Lo abbiamo già visto nel primo capitolo ma, per riassumere brevemente, la funzione di *concordancer* permette di visualizzare le concordanze, cioè una lista di uno o più termini nel loro contesto (il cosiddetto formato KWIC). Se però il *concordancer* di un corpus tradizionale si basa sul materiale di cui esso è composto, strumenti come questi si basano sulle pagine web, facendo del web un vero e proprio corpus di testi. WebCorp è il più rappresentativo dei vari *concordancer* online (a titolo informativo citiamo anche KWic finder[54]) e il suo funzionamento ci viene illustrato nel dettaglio da Renouf et al. 2007, che andremo qui a riassumere. La ricerca tramite questo software funziona esattamente come effettuare una qualsiasi ricerca tramite un motore di ricerca. L'interfaccia infatti è molto simile. La differenza sta nel fatto che WebCorp ci dà la possibilità di settare alcuni parametri come quello del *case-sensitive*, la lunghezza del testo che viene visualizzato prima e dopo la parola ricercata, la lingua, il motore di ricerca da cui recuperare il materiale. Quest'ultimo è possibile sceglierlo tra FAROO, Bing, FAROO news, Bing news, the Guardian Open Platform. Vediamo quindi la volontà dei creatori del software di proporre motori di ricerca in grado di fornire anche materiale più attendibile e controllato, proveniente da siti ufficiali e con un linguaggio controllato, come può essere quello di articoli di giornali e notizie.

[53] accessibile alla pagina http://www.webcorp.org.uk/live/
[54] https://www.kwicfinder.com/KWiCFinder.html

Il materiale recuperato da una ricerca tramite questo software non viene in alcun modo modificato, ciò che cambia è solamente il formato in cui viene mostrato. Oltre che prima, è possibile settare alcuni parametri anche dopo la ricerca, ovvero: scegliere se mostrare o meno il link della pagina da cui proviene il testo, scegliere un determinato sito in cui effettuare la ricerca tra alcuni proposti o inserendo il link, filtrare i risultati per data o selezionare un intervallo temporale entro cui effettuare la ricerca, selezionare la posizione nella frase in cui si deve trovare il termine, mettere un filtro alle parole utilizzando gli operatori supportati dal motore di ricerca scelto. Renouf et al. avevano previsto per il futuro l'implementazione dell'annotazione morfosintattica, cosa che oggi è effettivamente presente e che viene specificato nella pagina di introduzione al programma, sottolineando che WebCorp è "Our large-scale search engine with more search options, part-of-speech tags and quantitative analyses."

L'utilizzo di un software di questo tipo, rispetto al solo motore di ricerca, permette di affinare la ricerca e creare un ambiente più adatto alla ricerca linguistica, con la possibilità di elaborare il materiale sia prima che dopo. Fornisce anche alcuni dati statistici per quanto riguarda le collocazioni e permette di creare liste di frequenza da siti specifici. Il vantaggio è quello di essere uno strumento immediato, rispetto ai corpora, per verificare un'ipotesi linguistica (Lew 2009: 2).

È utile però tenere in considerazione anche i lati negativi di uno strumento di questo tipo, molti dei quali sono gli stessi dei motori di ricerca, data appunto la loro stretta dipendenza. In primo luogo, da alcune ricerche svolte con il suo utilizzo, molto più rispetto al solo motore di ricerca, è risultata la presenza di rumore, ovvero casi in cui le parole non facessero parte di un normale testo ma di link, URL, indici, liste o titoli (il cosiddetto *boilerplate*, ovvero il linguaggio utilizzato in parti strutturali di una pagina come la testata, il piè di pagina, le informazioni di navigazione ecc. (Gatto 2014: 124)). Bisogna quindi fare molta attenzione a controllare il materiale dopo la ricerca. Inoltre, i risultati vengono mostrati da un numero limitato di pagine, che va dalle 50 alle 100 pagine a seconda del motore di ricerca selezionato, il che vuol dire, considerando anche l'elevato rumore, che il numero di pagine potenzialmente rilevante potrebbe essere basso ai fini della ricerca. Infine, trattandosi di materiale web, che come abbiamo detto è in continua evoluzione, persiste la non riproducibilità della ricerca tipica della *corpus linguistics*.

Per capirne meglio il funzionamento e vederne nel concreto vantaggi e svantaggi, proponiamo un esempio di ricerca svolta (in data marzo 2020) con l'utilizzo di questo software. È stata ricercata la collocazione "задать вопрос" e questo è un fermo immagine rappresentativo dei risultati (Fig. 6).

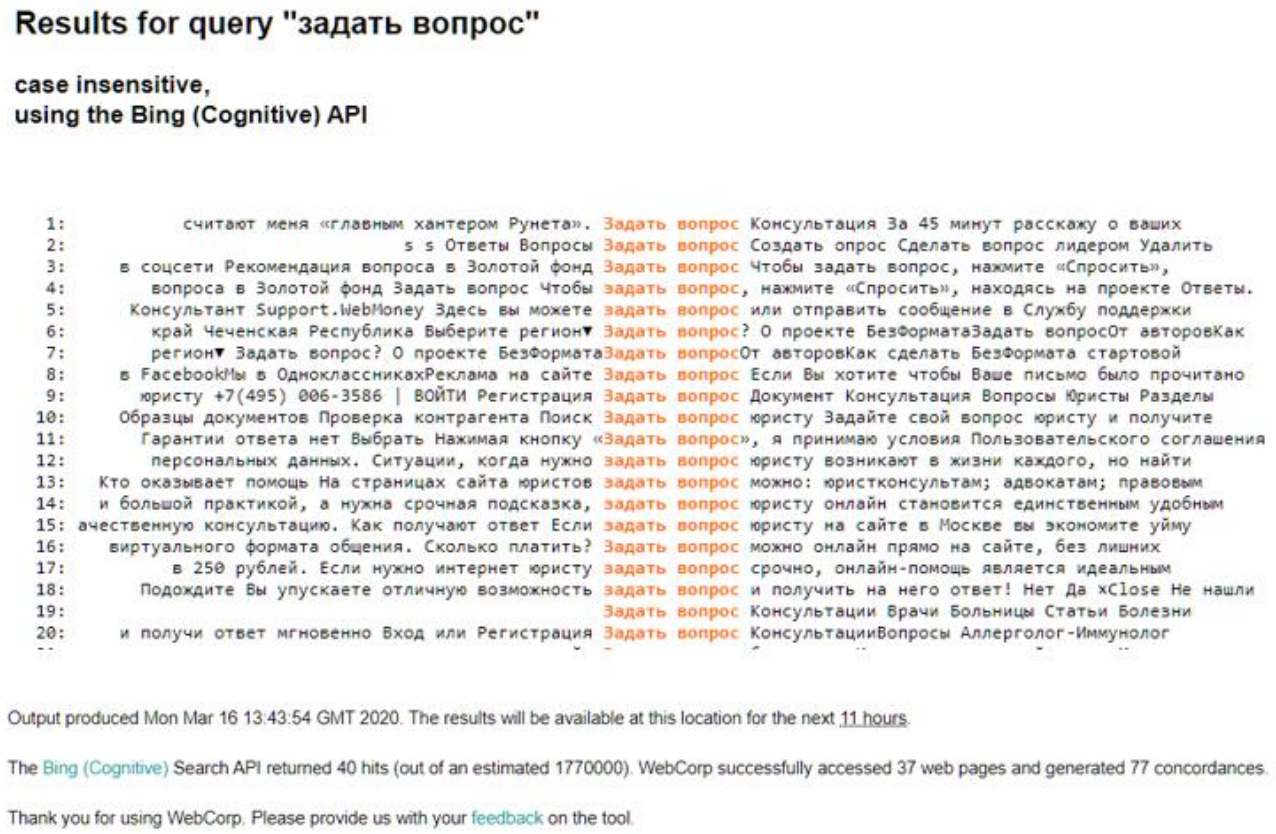


Figura 6. Risultati della ricerca con WebCorp

Vediamo prima di tutto la visualizzazione tipica delle collocazioni, molto utile in quanto permette di osservare in che modo una parola può essere utilizzata, in che contesti, gli elementi che la possono precedere o seguire e così via. Ci sono però vari elementi da mettere in luce. In primo luogo, come ci viene detto nella parte inferiore della pagina, le concordanze rilevate sono solamente 77 (qui ne vediamo solo 20 essendo solamente un'immagine campione) estratte da 37 pagine web, e più di così non ci è possibile visualizzarne. Il software, quindi, fornisce sì dati statistici, ma riguardo a una porzione limitata di pagine. In secondo luogo, pur avendo selezionato l'opzione *case insensitive*, probabilmente a causa del numero limitato dei risultati mostrati, non vediamo forme diversa da quella inserita nella ricerca, né per quanto riguarda il verbo né per il sostantivo. Terzo, è frequente la presenza di *boilerplate*. In questo caso infatti, la formula "задать вопрос" è presente molto spesso nei siti russi nelle sezioni in cui è possibile rivolgere domande ad un contatto di assistenza o rivolgere un qualche tipo di domanda in generale, ad esempio in un forum, cosa che in un corpora tradizionale non si troverebbe. Questi sono alcuni dei casi in questione (Fig. 7):

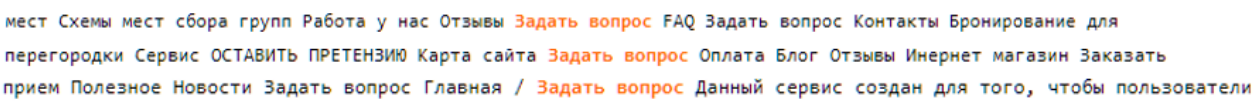


Figura 7. Risultati di una ricerca tramite WebCorp

Infine ci viene spiegato che i risultati di questa ricerca saranno salvati in remoto per 11 ore, dopodiché non saranno più accessibili. La stessa ricerca svolta successivamente potrebbe avere risultati differenti a causa della dinamicità delle pagine web.

3.2.2 Il web come *corpus shop*

Questo approccio prevede l'utilizzo di strumenti linguistici volti alla creazione di corpora specializzati comparabili e offline utilizzando materiale tratto dal web. Corpora di questo tipo sono tipicamente utilizzati dai traduttori specializzati che si affidano ad essi per la creazione di schede terminologiche, ovvero liste di termini legati ad un determinato ambito nella lingua di partenza, con il corrispondente nella lingua di arrivo. Per prima cosa andiamo ad analizzare i passaggi che questo processo prevede a livello teorico. Dopodiché vedremo un software in grado di svolgere l'intero processo in maniera quasi totalmente automatica.

Il primo step è quello di effettuare una *query*, tramite un qualsiasi motore di ricerca, che contenga una combinazione di parole chiave inerenti all'argomento del testo specializzato da tradurre. Possono essere utili in questa fase le varie opzioni offerte dai motori di ricerca, come i parametri sulla lingua, sul dominio (come domini di siti scientifici o accademici), la selezione di URL specifici da cui recuperare il materiale, gli operatori booleani ecc. Una volta effettuata la *query* si passa all'analisi e valutazione delle pagine web restituite dal motore di ricerca, ovvero la loro rilevanza rispetto all'argomento in questione in base al titolo della pagina, al contenuto, allo stile e al registro. I testi ritenuti adatti a far parte del corpus saranno poi scaricati ed infine ripuliti di tutto il *boilerplate* e altro materiale irrilevante. Se tutto ciò dovesse essere svolto manualmente, il traduttore dovrebbe dotarsi degli strumenti adatti a svolgere tali operazioni, come un motore di ricerca, programmi per processare e ripulire i testi una volta scaricati, *concordancer* per la visualizzazione dei testi e così via. Vediamo già da qui come un processo di questo tipo richiederebbe innanzitutto competenze informatiche piuttosto avanzate ma anche un tempo piuttosto lungo. A semplificare il tutto entra in gioco BootCat[55], un software in grado di svolgere tutte queste

[55] Scaricabile gratuitamente dal sito https://bootcat.dipintra.it/ oppure accessibile online dal sito si Sketch Engine

operazioni in una sola volta e in un unico luogo, ma soprattutto in maniera quasi totalmente automatica. In questo modo il traduttore potrà creare il proprio corpus specializzato monouso ogni volta che vorrà. Allo scopo di esemplificarne il funzionamento, che ci viene illustrato da Baroni e Bernardini (2004), vediamo i passaggi con cui è possibile realizzare un corpus specializzato nell'ambito della linguistica dei corpora.

La prima e, praticamente, unica cosa di cui si avrà bisogno è una combinazione di parole chiave inerenti al tema del testo da tradurre affinché il software possa individuare i testi adeguati. Queste parole chiave sono chiamate *seeds* e, una volta inserite, il programma le andrà automaticamente a ricercare in un motore di ricerca specifico a gruppi di due o tre combinate in maniera casuale (i cosiddetti *tuples*). La scelta dei *seeds* dovrà essere fatta attentamente, evitando termini che possano risultare ambigui o polivalenti che potrebbero recuperare pagine fuori tema. Questo causerebbe infatti molto rumore e i dati statistici non sarebbero più attendibili. Gran parte del successo del corpus dipenderà proprio da questa fase e dall'abilità del traduttore di scegliere i termini chiave corretti, per lo meno per quanto riguarda la prima ricerca. Vediamo nelle seguenti immagini la scelta dei *seeds* (Fig. 8) e successivamente i *tuples* (Fig. 9) che il programma ha restituito in automatico.

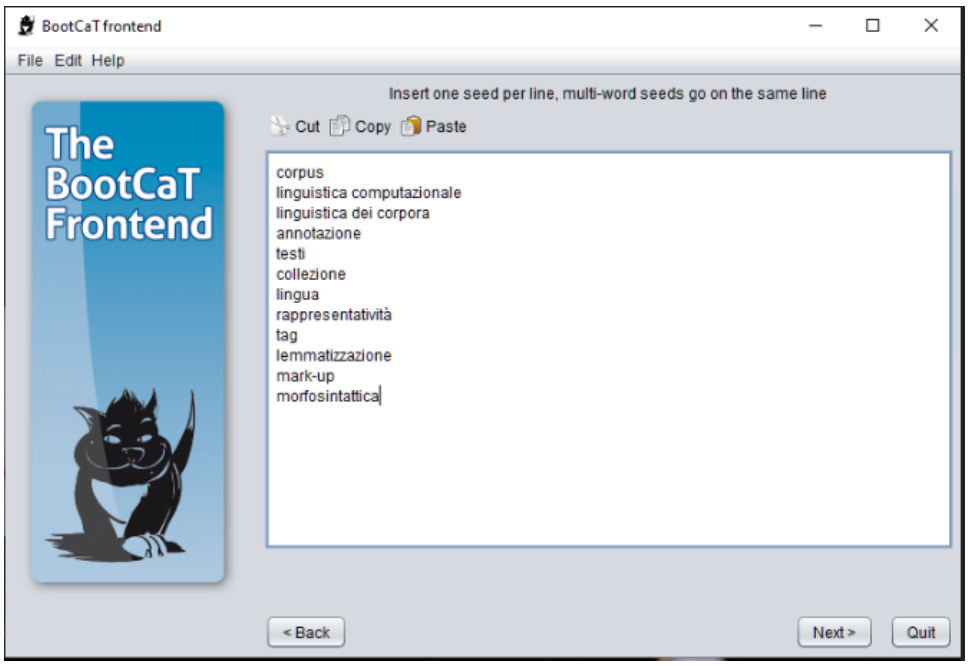


Figura 8. *Seeds* utilizzati per la creazione del corpus specializzato

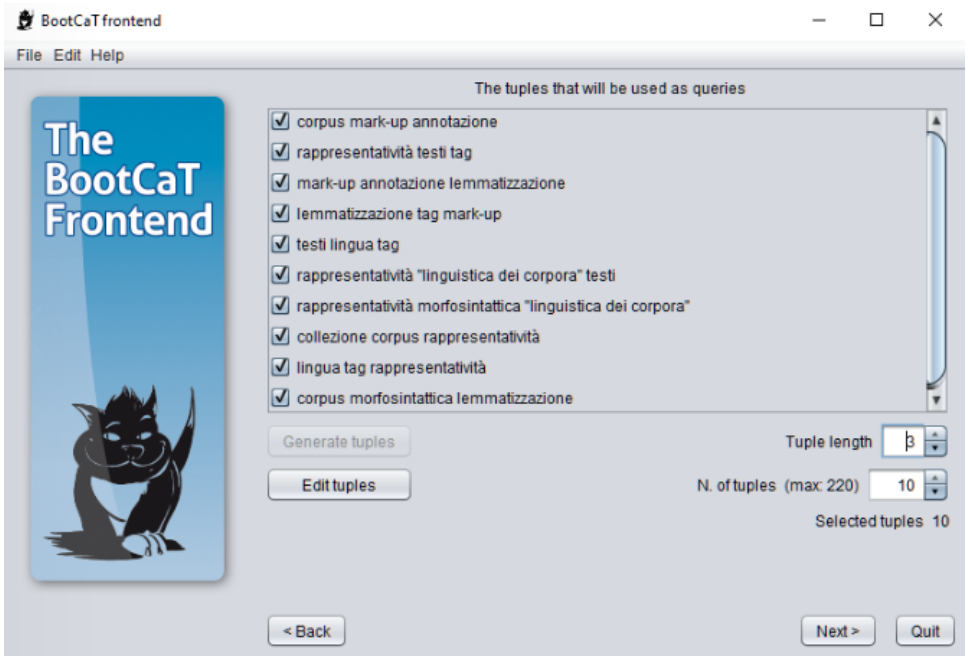


Figura 9. *Tuples* proposti in automatico dal software

A partire da queste parole il software andrà a recuperare le principali pagine web (che di default sono 10) e che dovranno poi essere manualmente scaricate. BootCat, in questa fase, permette di limitare la ricerca a specifici siti web o domini o di escluderne alcuni. Consente inoltre di limitare la scelta a testi di uno o più formati, come ad esempio il PDF, che è stato qui scelto in quanto abbiamo visto essere più autorevole.

Prima di procedere a scaricare le pagine, a processare il materiale convertendolo da HTML a testo semplice e a ripulirlo di tutti gli elementi indesiderati tra cui *boilerplate* e duplicati, il software permette all'utente di visualizzare gli URL delle pagine in modo da poterle visitare e valutare se sono rilevanti o meno alla sua ricerca, ed eventualmente eliminarli. Sempre allo scopo di aumentare la rilevanza, verranno automaticamente escluse pagine web troppo lunghe o troppo corte, dal momento che spesso non contengono materiale utile. Tutte queste operazioni fanno sì che il processo di creazione del corpus somigli a quello di corpora creati con metodi più tradizionali (Gatto 2014: 149).

Arrivati a questo punto abbiamo già un piccolo corpus a nostra disposizione. Da questo, però, è possibile andare a ripetere la procedura così da ampliarlo fino alle dimensioni desiderate. Per fare questo si dovrà ricavare da questo corpus una lista di parole chiave o una lista di frequenza da cui estrarre una nuova combinazione di *seeds* e ripetere la ricerca. Si avranno così altre dieci pagine web in aggiunta alle dieci iniziali.

Questo, ovviamente, è molto più di quanto il traduttore potrebbe fare manualmente e affidandosi a diversi programmi, e può essere svolto in tempi decisamente più brevi. Il materiale finale, che a questo punto risulta piuttosto pulito, sarà infine salvato localmente sul pc dell'utente. Il risultato è un vero e proprio corpus mirato alla traduzione specializzata che potrà essere utilizzato tramite un qualsiasi *concordancer* o piattaforma che permette di inserire il proprio corpora, come Sketch Engine (Fig. 10).

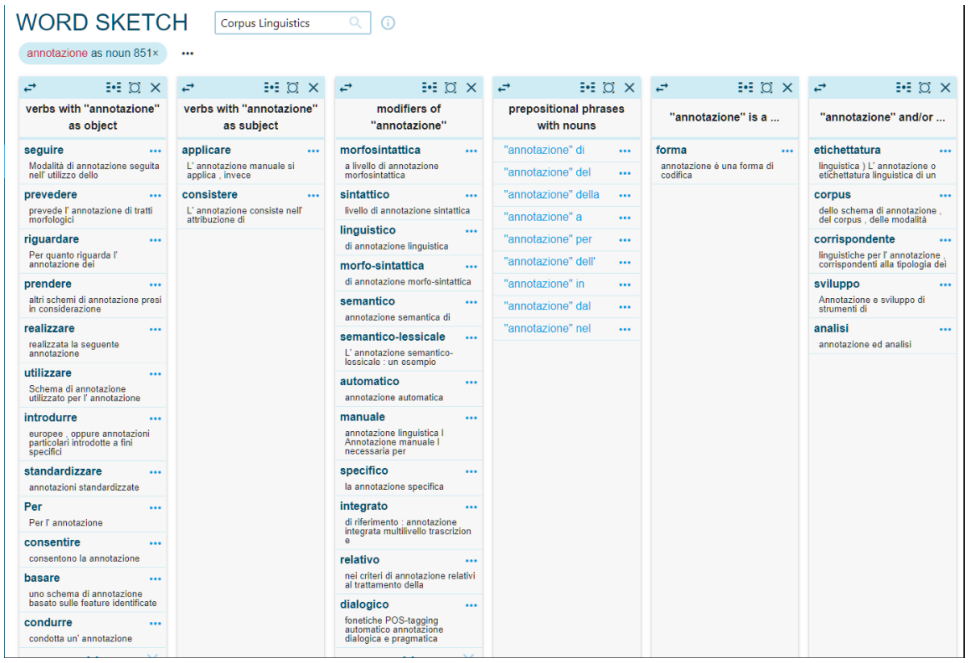


Figura 10. Esempio di utilizzo tramite Sketch Engine del corpus che abbiamo realizzato

Sketch Engine, infatti, in quanto corpus manager, oltre a contenere corpora realizzati dai suoi creatori, permette anche di lavorare su corpora già esistenti e pronti all'uso (ovvero già ripuliti, lemmatizzati, annotati e precaricati nel server) creati con la procedura appena vista. L'utente, quindi, potrà caricare il corpus da lui creato, che sia un mini corpus specializzato o un mega corpus generico come quelli che vedremo tra poco, e utilizzarlo sfruttando tutte le funzioni che Sketch Engine offre. Inoltre, c'è anche la possibilità di creare un corpus personalizzato, monolingue o multilingue, con lo stesso Sketch Engine, inserendo i propri testi o scegliendoli da una lista proposta dal programma. In questo caso possiamo dire che questo, seppur offra molte meno opzioni, sia un valido sostituto di WebCorp.

Affinché il corpus sia utile a scopi traduttivi, però, la procedura non può fermarsi qua. Il traduttore, a questo punto, necessita di un corpus comparabile a quello che ha appena creato, che potrà essere realizzato ripetendo la stessa procedura, con gli stessi termini chiave, ma in un'altra lingua, ovvero quella di arrivo. Nell'immagine vediamo i *seeds* che sono stati utilizzati per creare un corpus russo comparabile a quello precedente (Fig. 11). In quanto comparabile, questo corpus dovrà infatti contenere testi simili per genere e argomento all'altro.

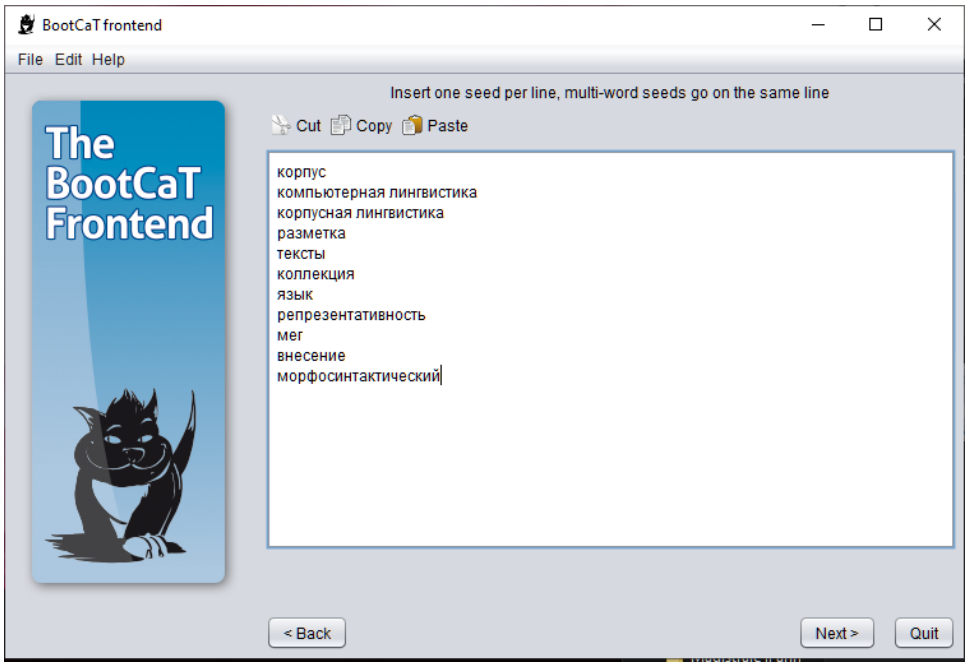


Figura 11. *Seeds* utilizzati per creare il corpus comparabile russo

Questa tecnica, in primo luogo per l'estrazione terminologica, è sempre più diffusa negli studi traduttivi. Negli ultimi anni, infatti, la creazione automatica di corpora comparabili è stata vista come la soluzione principale alla scarsità di corpora comparabili preesistenti, specialmente per la realizzazione di database terminologici (Gatto 2014:154), fondamentali nella traduzione specializzata. La creazione di schede terminologiche e liste di termini paralleli è sempre stata un passaggio fondamentale per i traduttori specializzati, ma il web ha reso questo processo più facile e veloce (Bernardini 2006). Ciò che sarà poi concretamente utile alla traduzione è la comparazione della terminologia per verificare come determinati termini sono stati tradotti nell'altra lingua.

3.2.2 Mega corpus/mini web

Oggetto di questo approccio sono i grandi corpora generici creati dal materiale presente nel web. Come ci spiegano Baroni e Bernardini (2006: 13-14), essi consistono in un nuovo oggetto, una sorta di mini-web (o mega corpus) adattato alla ricerca linguistica, con caratteristiche sia del web che dei corpora tradizionali: come il web è grande, aggiornato, contiene materiale testuale da pagine web e ha un'interfaccia simile ai motori di ricerca per l'accesso al materiale; come un corpus è annotato, permette ricerche complesse, è relativamente equilibrato. Possiamo quindi dire che questi corpora da una parte offrono tutte le potenzialità del web, come le dimensioni e la varietà di generi testuali, mentre dall'altra assicurano alcuni degli standard dei corpora tradizionali, come attendibilità, riproducibilità e stabilità (Gatto 2014: 164). Inoltre, le tipologie testuali in essi contenuti vanno dai testi più "classici" dei corpora, come di narrativa, scientifici, accademici, ai testi tipici del web come blog e post dei social media (Baroni et al. 2009: 212). Secondo alcuni studiosi, questo approccio mette finalmente d'accordo le potenzialità del web con i suoi limiti, inevitabili se si vuole sfruttare appieno ciò che di buono sa offrire. Si può dire che ci sia da parte della comunità dei linguisti l'accettazione di questi limini (poco controllo di contenuti, rumore, mancanza di bilanciamento) derivati non tanto dal fatto di essere basati sul web ma dall'essere corpora di grandi dimensioni costruiti in un breve arco di tempo e con poche risorse (Baroni, Ueyama 2006).

Proprio da questo approccio nascono i mega corpora multilingue del web, alcuni dei quali li abbiamo visti nel capitolo precedente relativamente alla lingua russa come quelli di piattaforme come Leeds Collection of Internet Corpora, Sketch Engine, Aranea, WaCky. Questi, oltre ad essere disponibili per numerosissime lingue, anche tra le meno diffuse, sono dotati delle tipiche funzionalità dei corpora linguistici, prima tra queste l'annotazione.

Andiamo quindi a vedere la procedura con cui si possono realizzare i mega corpora del web, che è stata illustrata da diverse figure tra cui Šarov (2006) e Baroni e Bernardini (2006, 2009). Questa volta, però, rimarremo a livello teorico evidenziando solamente in linea generali quelle che sono le diverse fasi della procedura. Ci sono poi tutta una serie di elementi da tenere in considerazione che non indagheremo nel dettaglio. Si tratta infatti di un lavoro difficilmente realizzabile da un singolo utente, come è stato possibile invece per la realizzazione di piccoli corpora specializzati, e che solitamente viene svolto da un team di ricercatori. La difficoltà in questo caso, oltre che a esistere dal punto di vista tecnico, esiste anche e soprattutto dal punto di vista del bilanciamento e della rappresentatività del corpus.

Le fasi di lavoro sono sostanzialmente le stesse necessarie alla creazione dei corpora specializzati. La prima è quella di recupero delle pagine web, per cui vengono sempre sfruttati i motori di ricerca. Una volta recuperato il materiale, si passa poi allo step in cui il materiale dovrà essere processato e ripulito degli elementi indesiderati. Il tutto può avvenire, anche in questo caso, con l'utilizzo di software come BootCat. Con altri software appositi il materiale recuperato verrà come ultima cosa lemmatizzato e annotato.

La fase cruciale tra quelle appena viste è proprio quella di recupero del materiale, a cui dedicheremo maggiore attenzione. Mentre per i corpora specializzati era sufficiente inserire come *seeds* alcune parole chiave inerenti all'argomento del testo da tradurre, per un corpora generico le cose sono diverse. Le pagine web che andranno a comporre il corpus dovranno infatti variare in terminologia, contenuto e genere (Baroni et al. 2009) per raggiungere un certo livello di rappresentatività della lingua in questione. Fermo restando che un metodo ben preciso e definito per raggiungere questo obiettivo non è ancora stato trovato, Baroni e Bernardini ci propongono il metodo da loro utilizzato nella realizzazione del progetto WaCky.

La lista di *seeds*, ovvero la combinazione di parole che formano la *query*, è stata creata selezionando una serie casuale di bi-grammi di parole cosiddette *content words*, ovvero parole piene e non funzionali o grammaticali. Queste, generalmente, devono aggirarsi intorno alle 400 e 500 (Šarov 2006: 438) e devono essere parole generiche non appartenenti a nessun dominio specifico. Gatto (2014: 168) ci spiega che sono preferibili coppie di parole in quanto parole singole potrebbero causare il recupero di documenti indesiderati, come definizioni da dizionario, titoli o nomi di compagnie che contengono quella parola e molto altro. Sono invece da evitare combinazioni di più di due parole dal momento che potrebbero portare al recupero di pagine in cui sono contenute liste sconnesse di parole. Baroni e Bernardini hanno notato che scegliendo le parole da fonti scritte come notiziari o dal materiale di corpora già esistenti tendono ad essere recuperati documenti appartenenti alla sfera pubblica, ovvero di ambito accademico, giornalistico, socio-politico ecc. Scegliendo invece le parole chiave da liste di parole tratte dai dizionari risultano pagine appartenenti più alla sfera di interesse personale, come ad esempio i blog. Siccome corpora di questo tipo dovrebbero tendenzialmente contenere entrambe le tipologie di testi, sono state utilizzate *seeds* da entrambe le fonti. Inoltre, come è stato fatto per la creazione di ukWac, per creare la lista di *seeds* è possibile anche utilizzare liste di frequenza tratte dal corpus nazionale di tale lingua, in questo caso il BNC, ma anche da dizionari di apprendenti, che solitamente contengono termini più formali e accademici. Le parole recuperate vengono poi appaiate casualmente in bi-grammi e con queste il recupero del materiale potrà finalmente avere luogo.

Grazie a corpora di questo tipo vengono superati i limiti tipici del web, come le dimensioni sconosciute e in continua evoluzione, che qui, seppur ancora sconosciute, sono per lo meno stabili. Questo garantisce la ripetibilità di un qualsiasi esperimento linguistico e la stabilità dei dati statistici ottenuti, che hanno in questo caso validità scientifica. La composizione di questi corpora, al contrario del web preso nel suo insieme, è in un certo modo strutturata, con il tentativo di raggiungere un certo bilanciamento e una certa rappresentatività. Parlando invece di lati negativi, nonostante sia presente l'annotazione morfosintattica, manca quella meta-testuale. Inoltre, in mancanza di una revisione manuale dovuta all'enorme numero di testi, l'annotazione può presentare degli errori, richiedendo in ogni caso una certa cautela nel valutare i dati ottenuti da questi corpora.

3.3 Altri strumenti web utili dal punto di vista linguistico

3.3.1 Google Books

Google Books è uno strumento sviluppato da Google che raccoglie libri integrali in forma digitalizzata. Si presenta con un'interfaccia simile al motore di ricerca Google ma in questo caso qualsiasi parola o frase sarà ricercata all'interno di questi libri. I testi, che sono sia antichi che più recenti e attualmente in commercio, sono presenti in moltissime lingue. Ogni lingua ha la propria interfaccia, ovvero Google Libri per l'italiano o Google Книги per il russo, anche se da ognuna sono accessibili i libri in qualsiasi lingua. Ad esempio, se dalla pagina italiana Google Libri ricerchiamo una parola o una frase in russo, questa verrà recuperata ugualmente. Nonostante non sia noto il numero esatto di testi o parole contenute in questo motore di ricerca, si stima che il numero di parole si aggiri attorno ai 500 miliardi, comprese tutte le lingue. Le opzioni offerte da questo motore di ricerca che possono risultare utili dal punto di vista della ricerca linguistica sono varie. In primo luogo è possibile selezionare il periodo di appartenenza del testo, ossia XIX, XX o XXI secolo, oppure inserire un intervallo temporale specifico a cui devono appartenere i libri in cui avverrà la ricerca. In questo modo fungerà da corpus diacronico con il quale analizzare l'evoluzione dell'utilizzo di determinate parole o forme nel tempo. In secondo luogo, possono essere svolte anche qui ricerche linguistiche più avanzate recandosi alla pagina della ricerca avanzata di Google Libri (Fig. 12).

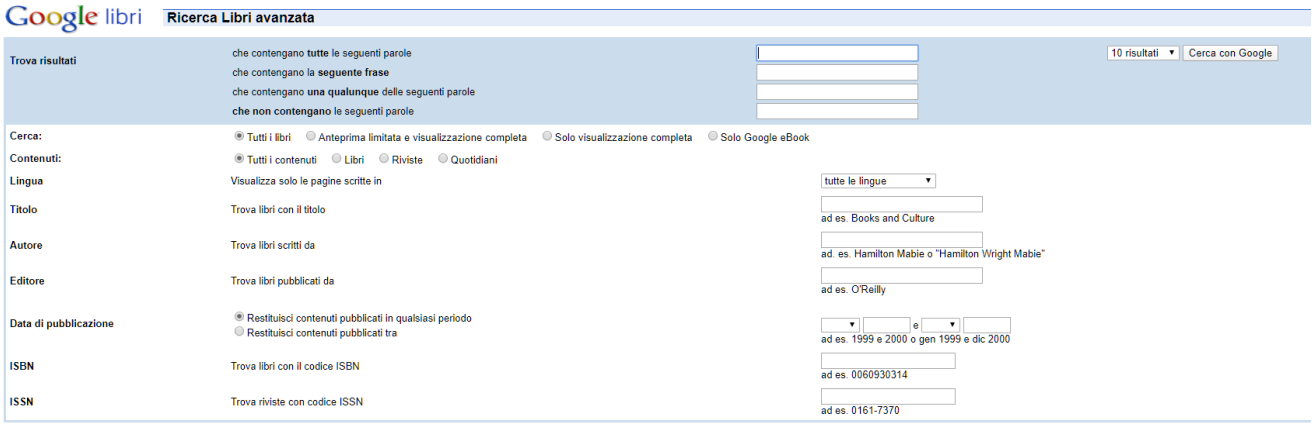


Figura 12. Ricerca avanzata di Google Libri

In questo modo, oltre a poter svolgere una ricerca avanzata dal punto di vista linguistico, con opzioni che sono le stesse offerte dal motore di ricerca Google (in azzurro nell'immagine), si potranno anche settare vari parametri per quanto riguarda elementi meta-testuali, come la tipologia testuale (libri, riviste, quotidiani), la lingua del testo (a scelta tra 46 lingue), l'autore, l'anno di pubblicazione ecc. Per questo aspetto possiamo dire che *Google Books* sia molto simile ad un corpus linguistico tradizionale. Ciò che manca, però, è l'annotazione morfosintattica.

3.3.2 Google Ngram Viewer

Google Ngram Viewer è un motore di ricerca legato a Google Books, ovvero basato sul suo materiale, il quale mostra la frequenza d'uso di parole o frasi in un periodo di tempo specifico. Come esempio è stata ricercata la parola "социальная сеть" (Fig. 13) e il grafico in questa immagine mostra come a partire dagli anni Duemila ci sia stato un incremento esponenziale del suo utilizzo, dovuto al fatto che i social network sono nati proprio in questi anni.

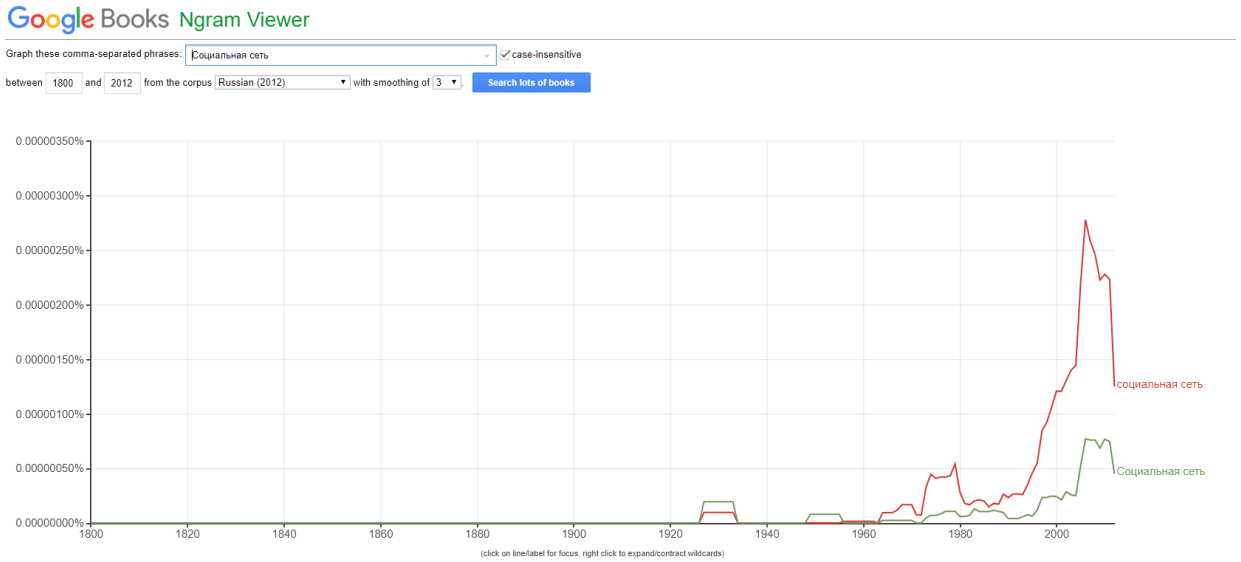


Figura 13. Esempio di ricerca tramite Google Ngram Viewer

Un altro esempio è dato dall'aggettivo "советский" (Fig. 14). Vediamo dall'immagine come dagli anni Venti l'utilizzo di questo aggettivo abbia cominciato a crescere fino a rimanere più o meno costante tra gli anni Quaranta e gli anni Ottanta, fino a diminuire notevolmente dagli anni Novanta. È chiaro come la frequenza di utilizzo di questo termine vada di pari passo con l'evoluzione storica, che in questo caso ha visto l'emergere, l'affermazione e la dissoluzione dell'Unione Sovietica in Russia.

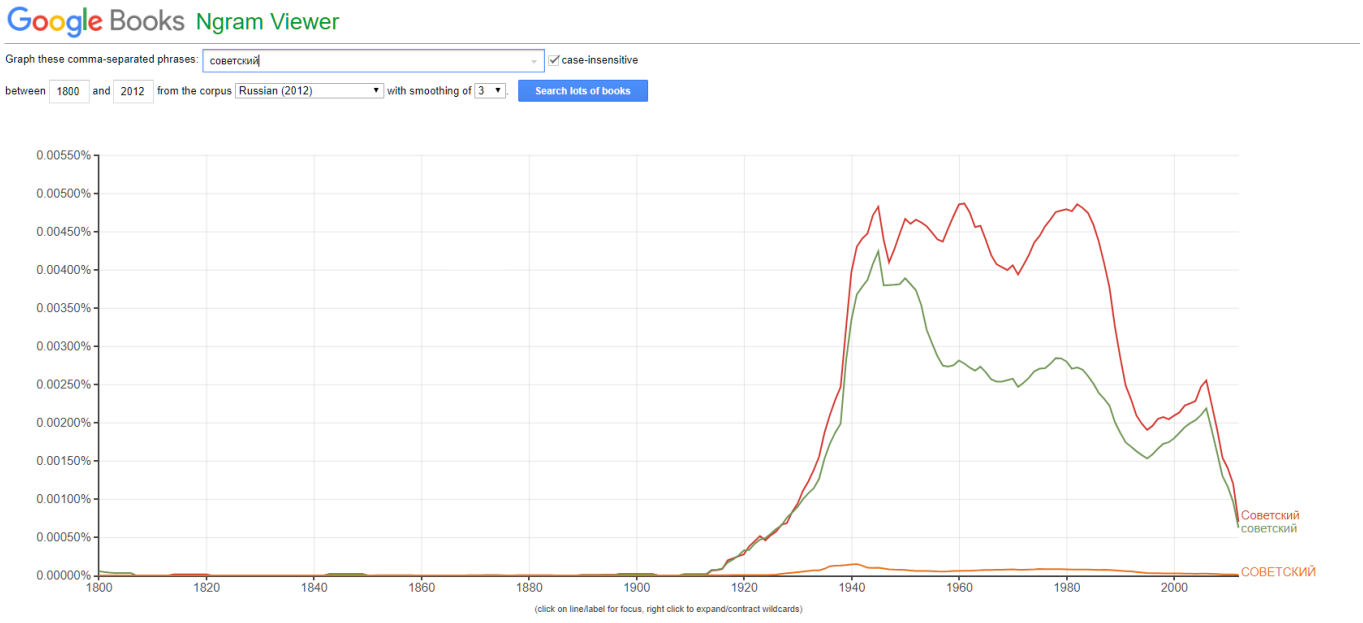


Figura 14. Esempio di ricerca tramite Google Ngram Viewer

Anche Google Ngram Viewer permette di settare alcuni parametri di ricerca. In primo luogo la lingua, che è selezionabile tra le principali lingue del mondo ovvero inglese americano o britannico, cinese, tedesco, spagnolo, francese, russo, italiano, ebraico. Per ognuna di queste è disponibile un corpus di testi aggiornato a data 2009 o 2012. In secondo luogo si può scegliere l'arco temporale entro cui effettuare la ricerca. Infine, questo motore di ricerca include anche la funzione *case-sensitive*, cioè permette di scegliere se ricercare una parola in una forma specifica (caso o declinazione) o in tutte le forme. Dopo aver effettuato la ricerca è possibile anche vedere tutti i testi in cui la parola o frase compare. I testi, ovviamente, sono protetti da copyright quindi verranno mostrate solamente le parti in cui sono contenuti gli elementi ricercati.

3.3.3 Google Scholar

Google Scholar è un motore di ricerca simile a Google Books ma che contiene testi appartenenti alla letteratura accademica come articoli scientifici, tesi di laurea, libri ed estratti. Anche in questo caso non sono note le dimensioni esatte di questa raccolta di testi ma si stima che contenga miliardi di parole. Nonostante sia incentrato per lo più su testi attuali, sono presenti anche testi scritti a partire dal XVIII secolo. Anche in questo caso c'è la possibilità di selezionare un arco temporale entro cui svolgere la ricerca, così come un autore o un ambito specifico. Dal punto di vista linguistico, uno strumento come questo può essere usato per esplorare collocazioni tipiche della lingua accademica grazie all'implementazione degli operatori booleani.

Come ci spiega anche Gatto (2014: 95), possiamo affermare che strumenti come Google Books, Google Ngram Viewer e Google Scholar sono la dimostrazione dello sforzo di Google di far interagire i servizi Google e la linguistica dei corpora più tradizionale. Ovviamente, ognuno di questi strumenti possiede tutti i limiti propri dei motori di ricerca che, nonostante questo sforzo, non sono comunque nati con lo scopo primario di effettuare ricerche linguistiche.

3.3.4 Wikipedia

Wikipedia è la più famosa enciclopedia online libera e multilingue. Proprio per il fatto di essere multilingue (copre più di 300 lingue), oltre che un'enciclopedia viene anche spesso considerata un corpus parallelo o comprabile (in quanto non sempre i testi sono l'esatta traduzione da una lingua all'altra.) di testi enciclopedici. Autori come Gamallo Otero e Gonzalez Lopez (2010: 21) la descrivono infatti come una fonte multilingue affidabile. Grazie al fatto di essere il più grande

archivio di testi simili in diverse lingue, Wikipedia è spesso fonte di estrazione terminologica per le parole chiave usate nella creazione di corpora specializzati monouso. Il potenziale di questa enciclopedia dal punto di vista linguistico è tale che sono in atto tentativi di creare corpora paralleli dal materiale in essa contenuto (Mohammadi, Ghasem Aghaee 2010).

I testi presenti in Google Books, Google Scholar e Wikipedia, così come quelli di molte altre enciclopedie o raccolte di testi scientifici online (citiamo ad esempio Academia.edu) possono essere utilizzati come fonte per la creazione di corpora specializzati con i metodi e gli strumenti che abbiamo analizzato nel corso di questo capitolo.

Possiamo concludere dicendo che la definizione di web come corpus è relativa, in quanto ci sono diversi modi in cui si può essere sfruttato, così come è relativo affermare che il web è uno strumento valido o meno di analisi linguistica, dal momento che questo dipende dal tipo di ricerca e dall'elemento linguistico da analizzare. L'obiettivo di questo capitolo è stato quello di illustrare i lati positivi e negativi del web come strumento di ricerca linguistica, le diverse modalità in cui può essere sfruttato e in quali casi è più o meno valido rispetto ad un corpus linguistico. I successi di una ricerca tramite questo strumento dipendono dalla consapevolezza che l'utente, il linguista o il traduttore hanno sia del potenziale che dei limiti della rete, sia dei modi in cui più essere utilizzato in maniera più efficace a scopi linguistici.

CAPITOLO 4
CORPORA TRADIZIONALI VS WEB CORPORA IN PRATICA

Introduzione

In questo capitolo conclusivo faremo un confronto tra i corpora tradizionali e i web corpora della lingua russa che abbiamo visto nel corso di questo lavoro mettendo in pratica il loro utilizzo. In particolare, confronteremo le prestazioni dell'uno e dell'altro tipo di corpus nell'analisi di alcuni aspetti linguistici. Questi saranno grammatica, collocazioni, anglicismi, termini gergali, tecnicismi e l'uso di due termini simili. Lo scopo di questa analisi vuole essere quello di testare quale tipologia di corpus risulta più utile a seconda del tipo di ricerca da svolgere. In funzione rappresentativa per la categoria dei corpora tradizionali verrà utilizzato il corpus russo per eccellenza, ovvero il NKRJa. Per la categoria dei web corpora, invece, saranno utilizzati principalmente ruTenTen (Sketch Engine), Araneum Russicum Maximum (Aranea Family) e il GIKRJa. Oltre al confronto tra il NKRJa e i corpora del web, questi ultimi verranno a loro volta messi a confronto tra di loro in alcune ricerche specifiche allo scopo di valutare quale di essi sia il più funzionale.

4.1 La ricerca

4.1.1 Grammatica

Nome collettivo

"Картошка", che significa "patata" o "patate" è un termine il cui utilizzo solleva spesso dei dubbi negli apprendenti italiani di lingua russa. Mentre in italiano è un nome individuale, e può quindi essere reso sia al singolare che al plurale, in russo è un nome collettivo e si usa solamente al singolare. Per questo motivo, a causa dell'influenza della lingua madre, sono frequenti i casi in cui viene usato erroneamente al plurale da persone di lingua madre italiana. A causare ulteriori dubbi entra in gioco il fatto che nella maggior parte dei contesti in cui viene utilizzato in funzione di complemento oggetto, il termine "картошка" ha terminazione in "-и". Non si tratta però di un accusativo plurale ma di un genitivo singolare partitivo. Il genitivo russo, infatti, oltre ad avere

funzione di specificazione, è anche usato come corrispondente del partitivo italiano (“vorrei dell’acqua”, “ho comprato del pane”). In casi come questi il complemento oggetto si troverà al caso genitivo in quanto l’azione si riferisce solo ad una parte indefinita dell’oggetto (in russo avremo: “мне бы хотелось воды” e “я купила хлеба”).

Un altro caso di genitivo partitivo lo ritroviamo nei complementi oggetto di verbi con prefisso “на-”. Questo prefisso è l’indicatore formale del modo d’azione cumulativo (*kumuljativnyj sposob dejstvija*), il quale indica appunto l’idea di accumulo come risultato dell’azione. Tali verbi, che sono ad esempio “наварить” (cucinare molto cibo) o “накупить” (comprare molte cose), reggono il caso genitivo oppure il loro complemento oggetto è retto da sostantivi che indicano quantità (“много”, “масса”, “куча” ecc.).

Affinché ogni dubbio venga chiarito, è possibile ricorrere all’utilizzo di un corpus linguistico per dimostrare come questo termine venga effettivamente utilizzato nella lingua russa e verificare se ci siano dei casi in cui esso ricorra al plurale. Vediamo quale corpus, per una ricerca puramente grammaticale come questa, ci può essere più di aiuto.

La nostra ricerca tramite il corpus prevede due fasi: la prima in cui ricerchiamo il termine “картошка” per osservare come viene utilizzato; la seconda in cui verifichiamo se ci sono dei casi in cui il sostantivo con terminazione in “-и” si trovi al caso nominativo o accusativo.

ruTenTen. Per la prima fase della ricerca digitiamo all’interno di Sketch Engine il termine “картошка”.

Queste sono immagini rappresentative dei risultati in formato *word sketch* (Fig. 15)

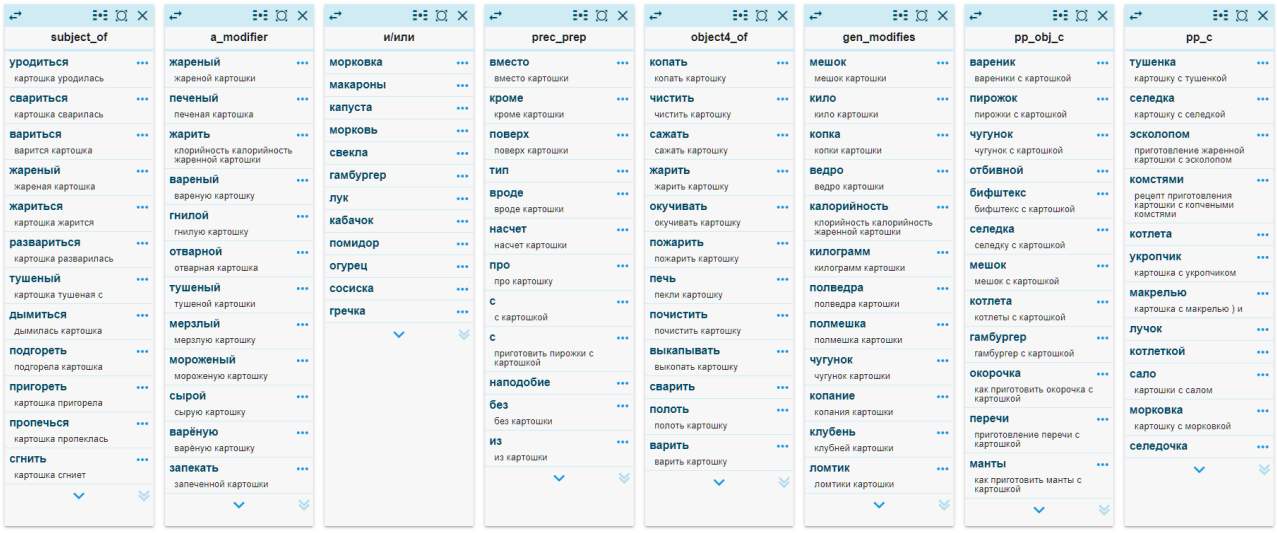


Figura 15. Ricerca della ricerca del termine "картошка" in formato *word sketch* tramite Sketch Engine

e in formato *concordancer* (Fig. 16)

Figura 16. Risultati della ricerca del termine "картошка" in formato *concordancer* tramite Sketch Engine

Per trarre le proprie conclusioni l'utente potrà scegliere il formato che preferisce. In entrambi i casi viene confermato ciò che abbiamo detto precedentemente, ovvero che il sostantivo viene utilizzato al singolare e i casi in cui ha terminazione in "-и" corrispondono al genitivo semplice o partitivo.

Per la seconda fase della ricerca iniziamo osservando l'uso del sostantivo con terminazione in "-и" (Fig. 17) (in questo caso solo in formato *concordancer* perché il formato *word sketch* non supporta il *case-sensitive*)

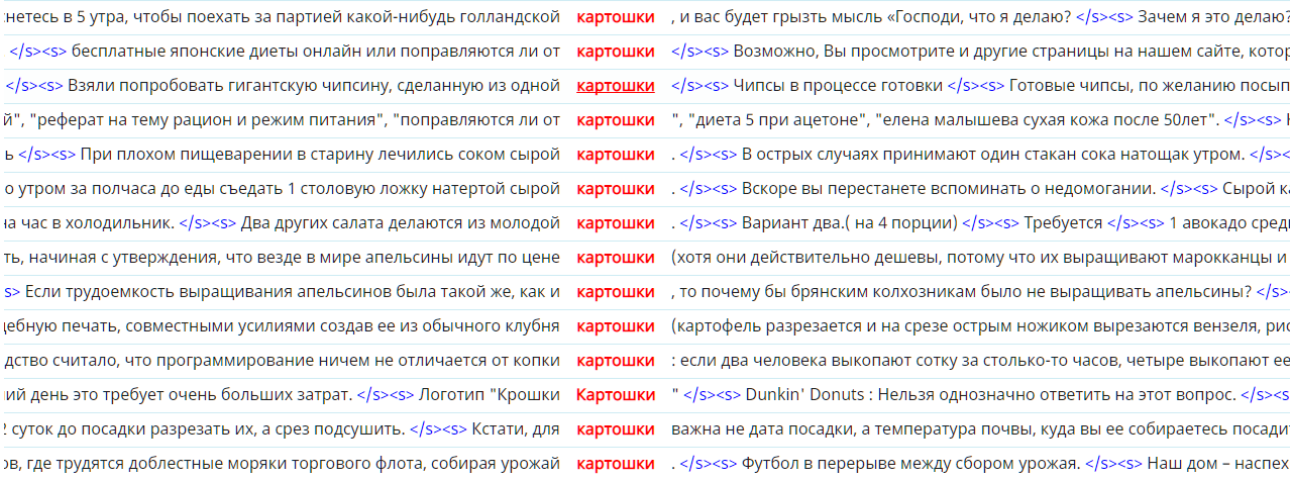


Figura 17. risultati della ricerca del termine "картошка" con terminazione in -и

Anche in questo caso notiamo che si tratta sempre di genitivi semplici o partitivi. Ma per avere la conferma effettiva del fatto che la terminazione in "-и" corrisponda solo al genitivo, dal momento che è impossibile scorrere manualmente tra tutti gli oltre 46 mila risultati, è necessario ricercare il

termine “картошки” come sostantivo al caso nominativo o accusativo. Ed è proprio qui che riscontriamo il problema principale di questo web corpus. Sketch Engine, infatti, permette di ricercare un termine selezionando la parte del discorso da questo riquadro (Fig. 18)

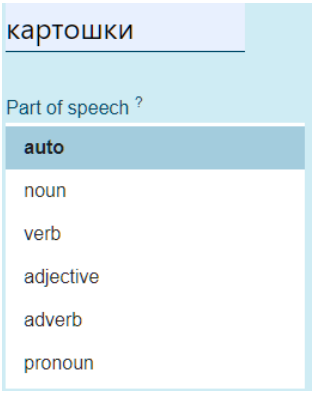
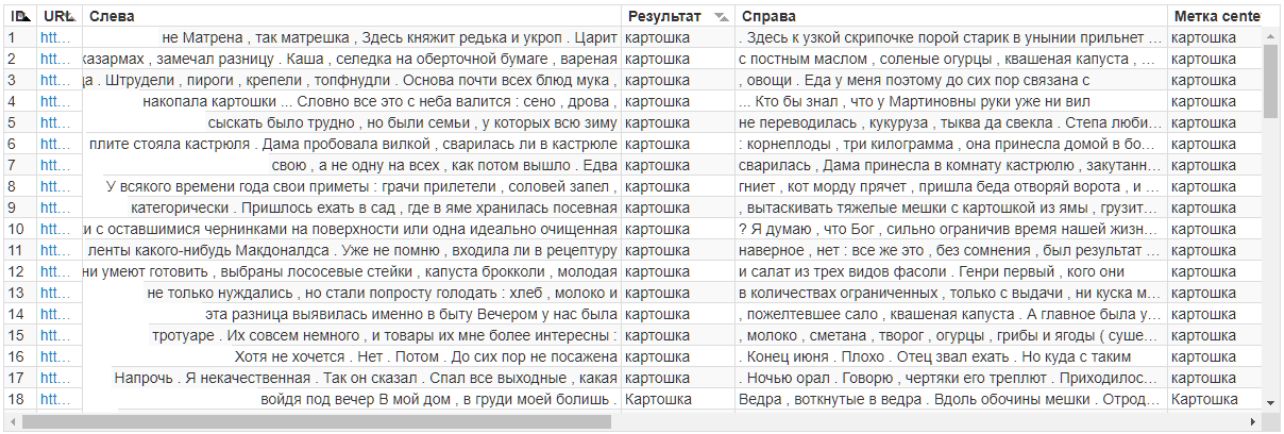


Figura 18. Opzioni di ricerca in Sketch Engine

ma non è possibile settare ulteriori parametri come ad esempio il caso, fondamentale per la nostra ricerca. Con la semplice ricerca del sostantivo “картошки” non potremo avere la conferma del fatto che si tratti esclusivamente di genitivo o se ci sia invece anche solo qualche raro caso in cui questa forma indichi un caso nominativo o accusativo.

La conclusione che possiamo trarre dopo questa ricerca è che Sketch Engine non si presta a ricerche di tipo grammaticale. Lo stesso possiamo dire, anche senza la dimostrazione pratica, del corpus Aranea in quanto, come Sketch Engine, non permette ricerche grammaticali complesse.

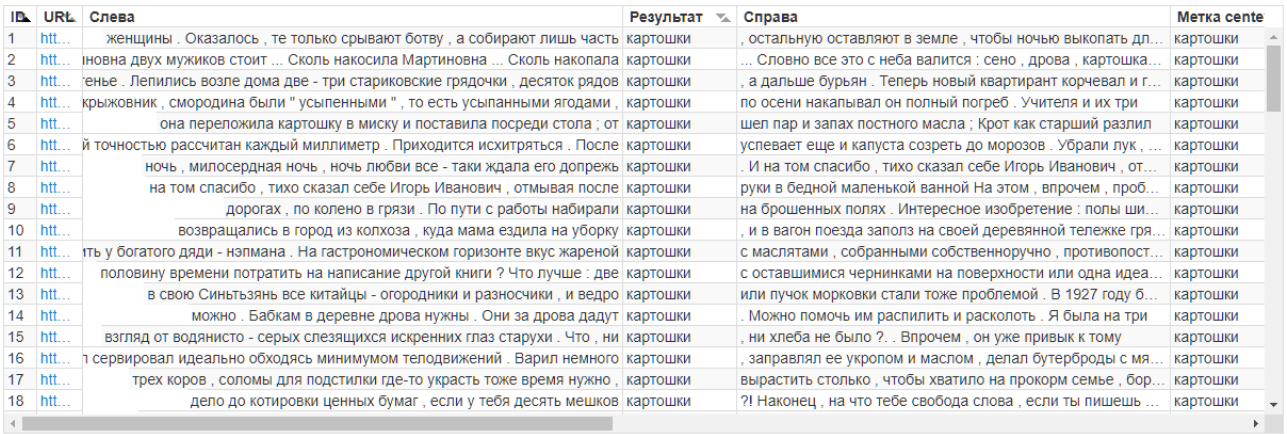
GIKRJa. Per quanto riguarda la prima fase, la differenza sostanziale tra questo corpus e il precedente sta nell’opzione *case-sensitive*. Mentre in Sketch Engine non è presente per formato *word sketch* ma lo è per il formato *concordancer*, qui risulta sempre attiva, senza la possibilità di disattivarla. Se infatti digitiamo la parola “картошка”, essa apparirà solo nella sua forma base al nominativo (Fig. 19).



ID	URL	Слева	Результат	Справа	Метка сente
1	htt…	не Матрена , так матрешка , Здесь княжит редька и укроп . Царит	картошка	. Здесь к узкой скрипочке порой старик в унынии прильнет …	картошка
2	htt…	казармах , замечал разницу . Каша , селедка на оберточной бумаге , вареная	картошка	с постным маслом , соленые огурцы , квашеная капуста , …	картошка
3	htt…	да . Штрудели , пироги , крепели , топфнудли . Основа почти всех блюд мука ,	картошка	, овощи . Еда у меня поэтому до сих пор связана с	картошка
4	htt…	накопала картошки ... Словно все это с неба валится : сено , дрова ,	картошка	... Кто бы знал , что у Мартиновны руки уже ни вил	картошка
5	htt…	сыскать было трудно , но были семьи , у которых всю зиму	картошка	не переводилась , кукуруза , тыква да свекла . Степа люби…	картошка
6	htt…	плите стояла кастрюля . Дама пробовала вилкой , сварилась ли в кастрюле	картошка	: корнеплоды , три килограмма , она принесла домой в бо…	картошка
7	htt…	свою , а не одну на всех , как потом вышло . Едва	картошка	сварилась , Дама принесла в комнату кастрюлю , закутанн…	картошка
8	htt…	У всякого времени года свои приметы : грачи прилетели , соловей запел ,	картошка	гниет , кот морду прячет , пришла беда отворяй ворота , и …	картошка
9	htt…	категорически . Пришлось ехать в сад , где в яме хранилась посевная	картошка	, вытаскивать тяжелые мешки с картошкой из ямы , грузит…	картошка
10	htt…	и с оставшимися чернинками на поверхности или одна идеально очищенная	картошка	? Я думаю , что Бог , сильно ограничив время нашей жизн…	картошка
11	htt…	ленты какого-нибудь Макдоналдса . Уже не помню , входила ли в рецептуру	картошка	наверное , нет : все же это , без сомнения , был результат …	картошка
12	htt…	ни умеют готовить , выбраны лососевые стейки , капуста брокколи , молодая	картошка	и салат из трех видов фасоли . Генри первый , кого они	картошка
13	htt…	не только нуждались , но стали попросту голодать : хлеб , молоко и	картошка	в количествах ограниченных , только с выдачи , ни куска м…	картошка
14	htt…	эта разница выявилась именно в быту Вечером у нас была	картошка	, пожелтевшее сало , квашеная капуста . А главное была у…	картошка
15	htt…	тротуаре . Их совсем немного , и товары их мне более интересны :	картошка	, молоко , сметана , творог , огурцы , грибы и ягоды (суше…	картошка
16	htt…	Хотя не хочется . Нет . Потом . До сих пор не посажена	картошка	. Конец июня . Плохо . Отец звал ехать . Но куда с таким	картошка
17	htt…	Напрочь . Я некачественная . Так он сказал . Спал все выходные , какая	картошка	. Ночью орал . Говорю , чертяки его треплют . Приходилос…	картошка
18	htt…	войдя под вечер В мой дом , в груди моей болишь .	Картошка	Ведра , воткнутые в ведра . Вдоль обочины мешки . Отрод…	Картошка

Figura 19. Risultati della ricerca del termine "картошка" nel GIKRJa

Per osservare l'utilizzo generale di una parola questo risulta essere un limite. Se l'opzione *case-sensitive* è utile, per non dire fondamentale, per alcune tipologie di ricerche, può essere anche controproducente per altre. La cosa migliore per un corpus è quella di permettere la sua attivazione e disattivazione a seconda delle necessità. Ciò che possiamo fare in questo caso per venire incontro all'esigenza della nostra ricerca è cercare il termine direttamente con terminazione in "-и" per osservare come viene utilizzato (Fig. 20).



ID	URL	Слева	Результат	Справа	Метка сente
1	htt…	женщины . Оказалось , те только срывают ботву , а собирают лишь часть	картошки	, остальную оставляют в земле , чтобы ночью выкопать дл…	картошки
2	htt…	иновна двух мужиков стоит ... Сколь накосила Мартиновна ... Сколь накопала	картошки	... Словно все это с неба валится : сено , дрова , картошка…	картошки
3	htt…	тенье . Лепились возле дома две - три стариковские грядочки , десяток рядов	картошки	, а дальше бурьян . Теперь новый квартирант корчевал и г…	картошки
4	htt…	крыжовник , смородина были " усыпенными " , то есть усыпанными ягодами ,	картошки	по осени накапывал он полный погреб . Учителя и их три	картошки
5	htt…	она переложила картошку в миску и поставила посреди стола ; от	картошки	шел пар и запах постного масла ; Крот как старший разлил	картошки
6	htt…	й точностью рассчитан каждый миллиметр . Приходится исхитряться . После	картошки	успевает еще и капуста созреть до морозов . Убрали лук , …	картошки
7	htt…	ночь , милосердная ночь , ночь любви все - таки ждала его допрежь	картошки	. И на том спасибо , тихо сказал себе Игорь Иванович , от…	картошки
8	htt…	на том спасибо , тихо сказал себе Игорь Иванович , отмывая после	картошки	руки в бедной маленькой ванной На этом , впрочем , проб…	картошки
9	htt…	дорогах , по колено в грязи . По пути с работы набирали	картошки	на брошенных полях . Интересное изобретение : полы ши…	картошки
10	htt…	возвращались в город из колхоза , куда мама ездила на уборку	картошки	, и в вагон поезда заполз на своей деревянной тележке гря…	картошки
11	htt…	ить у богатого дяди - нэпмана . На гастрономическом горизонте вкус жареной	картошки	с маслятами , собранными собственноручно , противопост…	картошки
12	htt…	половину времени потратить на написание другой книги ? Что лучше : две	картошки	с оставшимися чернинками на поверхности или одна идеа…	картошки
13	htt…	в свою Синьтьзянь все китайцы - огородники и разносчики , и ведро	картошки	или пучок морковки стали тоже проблемой . В 1927 году б…	картошки
14	htt…	можно . Бабкам в деревне дрова нужны . Они за дрова дадут	картошки	. Можно помочь им распилить и расколоть . Я была на три	картошки
15	htt…	взгляд от водянисто - серых слезящихся искренних глаз старухи . Что , ни	картошки	, ни хлеба не было ?. . Впрочем , он уже привык к тому	картошки
16	htt…	л сервировал идеально обходясь минимумом телодвижений . Варил немного	картошки	, заправлял ее укропом и маслом , делал бутерброды с мя…	картошки
17	htt…	трех коров , соломы для подстилки где-то украсть тоже время нужно ,	картошки	вырастить столько , чтобы хватило на прокорм семье , бор…	картошки
18	htt…	дело до котировки ценных бумаг , если у тебя десять мешков	картошки	?! Наконец , на что тебе свобода слова , если ты пишешь …	картошки

Figura 20. Risultati della ricerca del termine "картошки" nel GIKRJa

In questo caso sta all'utente osservare i risultati, il quale potrà trarre le stesse conclusioni che abbiamo visto sopra.

Per la seconda fase, il GIKRJa si presenta come un buon corpus per le ricerche grammaticali dal momento che permette il settaggio di tutta una serie di opzioni per quanto riguarda le singole parti del discorso. Lo vediamo in questa immagine (Fig. 21):

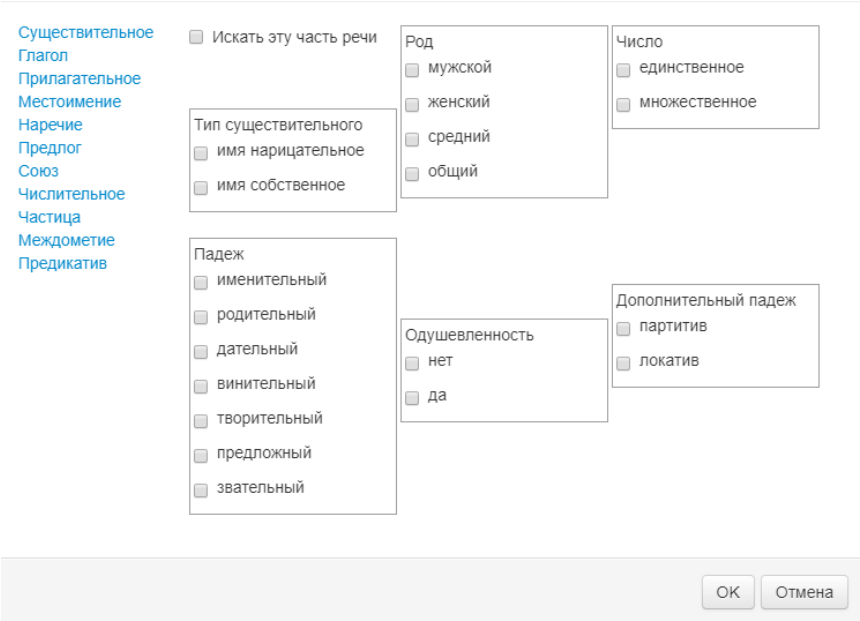


Figura 21. Opzioni di ricerca del corpus GIKRJa

Per un sostantivo, come nel nostro caso, si possono selezionare tipologia, caso, genere, numero, animato o inanimato e addirittura casi come partitivo e locativo. Se a prima vista può sembrare molto semplice, nel momento in cui proviamo a formulare una ricerca un po' più elaborata dal punto di vista grammaticale sfruttando queste opzioni, ci accorgiamo che in realtà non è un'operazione così immediata. La selezione di queste opzioni non permette di ricercare direttamente un termine che abbia tali caratteristiche, ma bensì di formulare graficamente la query necessaria poi alla ricerca.

Se ad esempio volessimo cercare la parola "картошка" al caso accusativo, una volta selezionati i parametri "sostantivo" e "caso accusativo", nella casella in cui inserire la query appare questa dicitura (Fig. 22):

Figura 22. Formulazione query complessa nel GIKRJa

A questa, però, dovrà esserne aggiunta manualmente un'altra che includa il termine che vogliamo ricercare (картошка) e che abbia questa forma: [lemma=картошка]

La query finale sarà quindi questa: [lemma=картошка] [pos="N...a.."]

L’inserimento di tale formulazione non è qualcosa che l’utente potrà intuire facilmente, è infatti necessaria l’attenta lettura delle istruzioni per la formulazione di ricerche complesse. Una volta assodato questo, la ricerca può proseguire. Continuando si accorre però ad altri problemi. In primo luogo la ricerca richiede tempi piuttosto lunghi, che in certi casi superano addirittura i 20 minuti, ma soprattutto la ricerca con questa query risulta non funzionare (Fig. 23).

ID	URL	Слева	Результат	Справа	Метка cente
1	htt…	поворчали да и успокоились - мол , пускай уж . В конце концов ,	картошкой Е…	кормят Франция и Испания , а нефте- и газопроводы и в	картошкой
2	htt…	более всего поразило объявление начальника поезда о бесплатных обедах (	картошка Ро…	с мясом) для людей старше 55 лет (выдача только по	картошка
3	htt…	свою страну со словами : " Если мы будем есть гамбургеры и	картошку ты…	лет , мы станем выше , наша кожа побелеет , и из брюнетов	картошку
4	htt…	ести " , владельцы передвижных закусочных-"фритри " даже приготовили для	картошки соус	, носящий имя визитной карточки бельгийской столицы - А…	картошки
5	htt…	и дать настояться 10 минут . Украсить укропом . " Сугроб " 3 вареных	картошки гри…	(любые : жареные , соленые , маринованные) 2 луковиц…	картошки
6	htt…	жарим на высоком огне в течение 5 минут . Выкладываем на	картошку по…	количества риса . Утрамбовываем . На рис выкладываем …	картошку
7	htt…	ешиваем , оставляем на некоторое время . Фарш+чеснок+лук+ссчп . Вносим	картошку ду…	на противне , минут на 30-40 . На вареные яйца напялива…	картошку
8	htt…	sCounter['903474733 '] = 'Subject ' ; } }) ; Свинина с виноградом и запеченной	картошкой В…	: активное – 20 минут пассивное – 50 минут Ингредиенты (…	картошкой
9	htt…	ки остатки соуса ! Приятного аппетита ! Свинина с виноградом и запеченной	картошкой В…	: активное - 20 минут пассивное - 50 минут Ингредиенты (…	картошкой
10	htt…	Сан-Джорджо (San Giorgio) под Неаполем купила на рынке вместе с	картошкой гр…	, сообщили журналистам в среду представители правоохр…	картошкой
11	htt…	свою страну со словами : " Если мы будем есть гамбургеры и	картошку ты…	лет , мы станем выше , наша кожа побелеет , и из брюнетов	картошку
12	htt…	кстати про инвестиции : " Доволен своим урожаем Петро : Весной посадил я	картошки ве…	. И столько же вырастил , много ли - мало А все как-никак	картошки
13	htt…	Мне нравятся с	картошкой б…	. Они , кстати , продаются и в обычном супермаркете	картошкой
14	htt…	очень помогает жить хозяйство - печь протопи , кур накорми , снег разгреби ,	картошку пр…	и так далее . То есть ты если в деревне не	картошку
15	htt…	редиску посадить надо , кустарник постричь . Траву сгрести и вообще . К	картошке гря…	подготовить . А соседка моя безногая в церкву не попала , в	картошке
16	htt…	на	картошке ма…	, да . погода ещё хорошая . на море сегодня собираемся .	картошке
17	htt…	уживании в данном заведении . Приятно , конечно , когда предлагают взять к	картошке тост	или десерт к чаю , но когда начинают переспрашивать по …	картошке
18	htt…	картошка да картошка , А когда же молоко ?! С этой ебанной	картошки Хуй	не лезет глубоко ! Поломалася машина , Не работает мото…	картошки

Figura 23. Risultati di una ricerca complessa tramite il GIKRJa

Vediamo infatti che tra i risultati sono presenti varie forme del sostantivo tra cui casi diversi da quello accusativo.

Il GIKRJa, quindi, nonostante teoricamente abbia tutte le carte in regole per esserlo, non risulta un corpus valido per ricerche di tipo grammaticale. Questo può essere dovuto al fatto che sia ancora in fase di sviluppo o che l’annotazione, che in corpora di queste dimensioni avviene in maniera automatica, presenti molti errori.

NKRJa. Per la prima parte della ricerca, tramite il NKRJa funziona allo stesso modo dei corpora del web ed è possibile compiere le stesse osservazioni e trarre le stesse conclusioni. La differenza sostanziale di questo corpus è evidente nella seconda parte della ricerca. Il suo punto di forza è proprio quello della ricerca complessa dal punto di vista grammaticale. Una volta inserita la parola “картошки” il NKRJa permette di settare moltissimi parametri (Fig. 24).

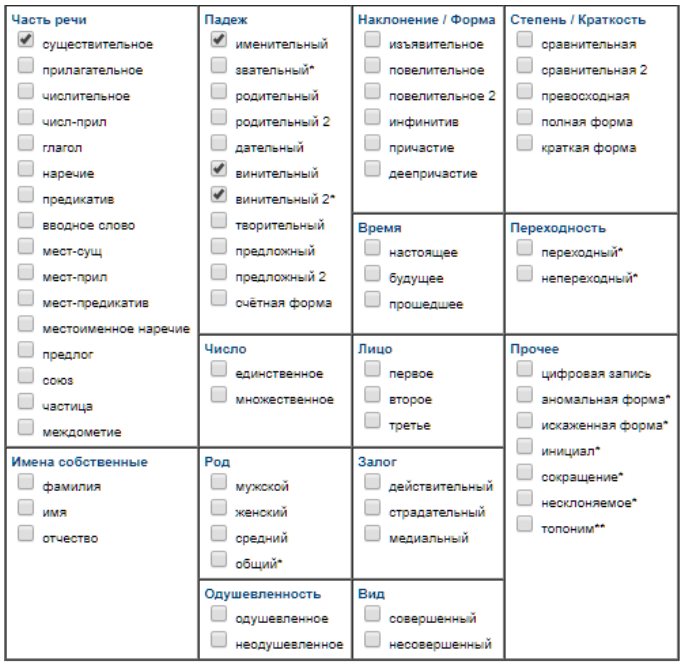


Figura 24. Settaggio dei parametri di ricerca per il sostantivo "картошки"

Per la nostra ricerca abbiamo selezionato “sostantivo al caso nominativo e accusativo”.

Dalla ricerca non emerge alcun risultato, il che risolverà ogni dubbio: non è possibile utilizzare il sostantivo “картошки” al nominativo o all’accusativo plurale (Fig. 25).

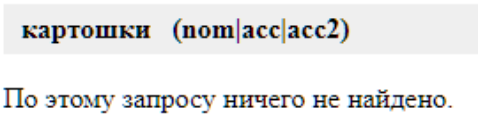


Figura 25. Risultati di ricerca per il termine "картошки" al caso nominativo e accusativo

Genitivo partitivo

Oltre ad esprimere indefinitezza o a denotare il modo d’azione cumulativo, come accennato poco fa, il genitivo partitivo denota anche parzialità. Questo in russo, per i sostantivi maschili singolari, può essere espresso con la terminazione “-y”, differenziandosi dal genitivo semplice in “-a”. Ne sono un esempio le frasi “я пил чаю” che significa “ho bevuto un po’ di te” oppure “я добавил немного сахару” che sta per “ho aggiunto un po’ di zucchero”. Questo fenomeno sta però via via riducendosi lasciando sempre più spazio al classico suffisso “-a” del genitivo maschile. In questo

caso ci serviremo dello strumento del corpus per analizzare meglio l'utilizzo e l'evolversi del genitivo partitivo singolare maschile in "-у".

Prendendo come esempio la locuzione "немного сахару", possiamo confrontare la frequenza del suo utilizzo rispetto alla stessa con la terminazione del genitivo maschile singolare semplice, ovvero "немного сахара".

NKRJa. Tramite il NKRJa vediamo che "немного сахару" (32 occorrenze) (Fig. 26) è molto più frequente rispetto a "немного сахара" (13 occorrenze) (Fig. 27), emergendo in un numero superiore al doppio.

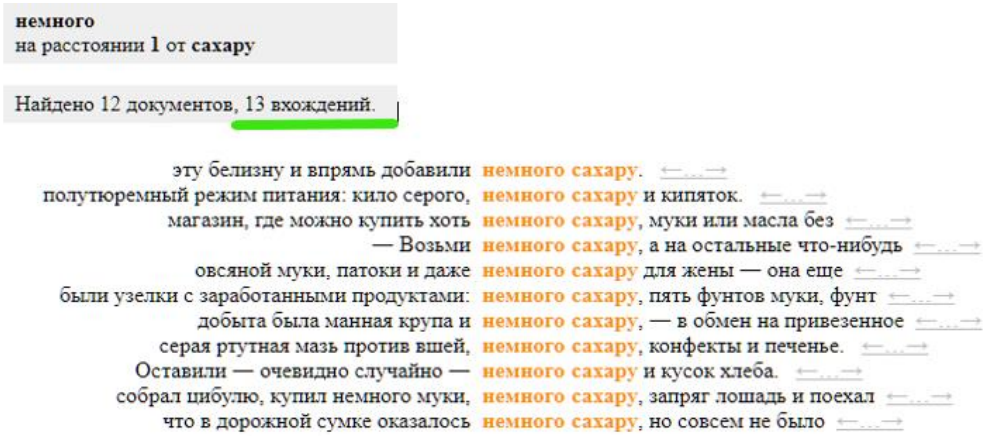


Figura 26. Risultati della ricerca della locuzione "немного сахару" tramite il NKRJa

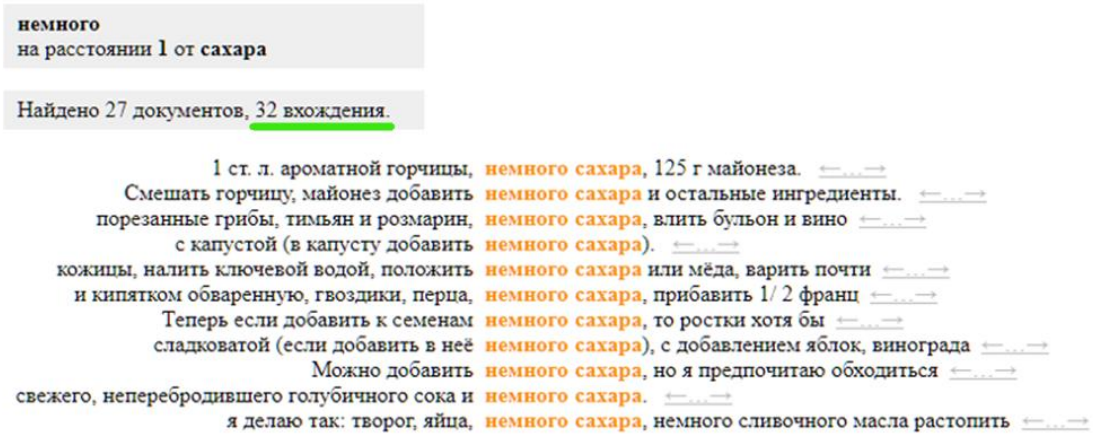


Figura 27. Risultati della ricerca della locuzione "немного сахара" tramite il NKRJa

Ora ripetiamo la stessa ricerca utilizzando alcuni corpora del web.

ruTenTen. Tramite questo web corpus emerge immediatamente il fatto che, in proporzione alle sue dimensioni, la forma al genitivo partitivo è presente in misura estremamente ridotta rispetto al

NKRJa (Fig. 28). Se nel corpus tradizionale la forma al genitivo semplice era presente oltre il doppio delle volte rispetto al genitivo partitivo, qui si parla di 20 volte di più.

Figura 26. Numero di occorrenze per le locuzioni "немного сахару" e "немного сахара" tramite Sketch Engine

Aranea. Nel corpus Araneum Russicum Maximum l'infrequenza del genitivo partitivo rispetto al genitivo semplice è ancora più evidente (Fig. 29).

Figura 27. Numero di occorrenze per le locuzioni "немного сахару" e "немного сахара" tramite Aranea

GIKRJa. Lo stesso si può notare anche con questo corpus, il quale presenta solamente 80 occorrenze per il genitivo partitivo e 1.985 per il genitivo semplice (Fig. 30).

Figura 28. Numero di occorrenze per le locuzioni "немного сахару" e "немного сахара" tramite il GIKRJa

La scarsa presenza del genitivo partitivo nel web, in cui la lingua si aggiorna di giorno in giorno, conferma il fatto che questa è una forma sempre meno utilizzata. Dall'altra parte, il NKRJa, in cui c'è una presenza importante anche della lingua del passato, mantiene in maniera più consistente la presenza di alcune forme di cui oggigiorno si va perdendo l'utilizzo.

Dal momento che nel NKRJa le occorrenze della locuzione "немного сахару" sono un numero piuttosto ridotto, e grazie anche al fatto che il corpus permette di ordinare i risultati in base al criterio cronologico, ci è stato possibile ricavare l'arco di tempo a cui appartengono i testi in cui sono presenti le occorrenze di questa locuzione al genitivo partitivo. È emerso che tali testi sono stati pubblicati in un periodo che va dal 1747 al 1994. Ciò confermerebbe che negli anni Duemila questa forma sia quasi scomparsa. Essendo i web corpora più aggiornati, sarebbe interessante compiere la stessa operazione per verificare in che misura siano presenti casi recenti del suo utilizzo. Purtroppo, però, la maggior parte dei corpora del web manca di annotazione meta-testuale.

L’unico a contenere informazioni sull’anno di pubblicazione online dei testi è il GIKRJa, dal quale vediamo che le 80 occorrenze appartengono a testi (di cui tutti blog tranne otto) scritti tra il 1999 al 2015. 80 occorrenze in 16 anni, in una fonte così ampia come è il web, sono un numero piuttosto limitato e ci confermano lo scarso utilizzo di questa forma.

Oltre che tramite un corpus, tradizionale o del web che sia, un calcolo di frequenza come questo può essere effettuato anche utilizzando i motori di ricerca.

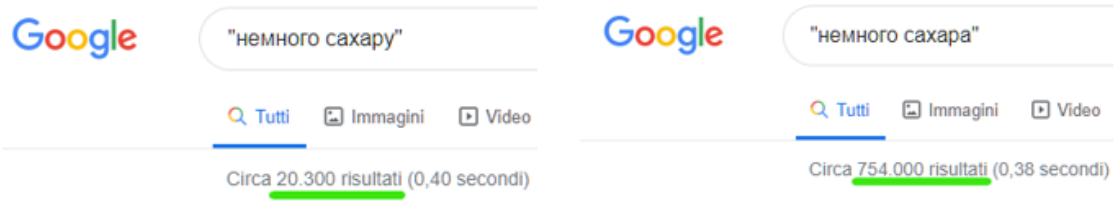


Figura 29. Numero di occorrenze per le locuzioni "немного сахару" e "немного сахара" tramite Google

Figura 30. Numero di occorrenze per le locuzioni "немного сахару" e "немного сахара" tramite Yandex

Anche Google (Fig. 31) e Yandex (Fig. 32) ci confermano l’utilizzo ridotto di questa forma. Ovviamente Yandex, che è un motore di ricerca russo, è in grado di reperire molti meno risultati rispetto ad un motore di ricerca globale come Google. In questo caso però non sono i numeri che dobbiamo guardare ma le proporzioni dei risultati dell’una e dell’altra forma.

A conclusione di questa ricerca abbiamo visto come il NKRJa sia il corpus in cui è meno marcata, seppur comunque evidente, la sproporzione di utilizzo che esiste tra il genitivo partitivo maschile con terminazione in “-y” e quello che mantiene la terminazione classica del genitivo maschile in “-a”. I web corpora, quindi, si sono dimostrati lo strumento migliore per evidenziare il fatto che questo fenomeno grammaticale stia entrando in disuso. Dall’altra parte, il NKRJa risulterebbe invece molto valido, rispetto ai corpora del web, per lo studio di fenomeni linguistici meno recenti o della lingua letteraria russa.

Verbi bi-aspettuali

Sono definiti bi-aspettuali quei verbi che, mancando di coppia aspettuale, fungono sia da verbo di aspetto imperfettivo che perfettivo. A determinare l'aspetto non sarà quindi la forma del verbo ma il contesto in cui viene utilizzato.

Come ci spiegano Zaliznjak e Šmelev (2000: 71), alcuni di questi sono verbi slavi di origine antica, i quali sono in continua diminuzione, mentre per la maggior parte si tratta di verbi di origine straniera che sono per tanto in costante espansione. Un esempio possono esserlo i verbi "исследовать" per la prima categoria e "диагностировать" per la seconda.

Nonostante questi verbi siano "ufficialmente" considerati bi-aspettuali, e vengano indicati nei dizionari come verbi di aspetto sia imperfettivo che perfettivo senza una coppia aspettuale, alcuni di essi, col passare del tempo, vengono integrati dal sistema aspettuale con la prefissazione o la suffissazione. Ne sono un esempio i verbi "поисследовать" e "продиагностировать".

Ci serviremo dei corpora per analizzare l'utilizzo di questi verbi prefissati nel tempo e avere un'idea della proporzione in cui si trovano rispetto alla loro versione senza prefisso.

Abbiamo riassunto i dati ricavati in queste tabelle (Tab. 1 e 2), in cui sono stati riportati il numero di occorrenze dei verbi nei vari corpora e il periodo a cui esse appartengono (quest'ultimo dato è stato ricavato esclusivamente dal NKRJa in quanto Sketch Engine non include i dati meta-testuali e il GIKRJa non comprende testi anteriori al XX secolo). Inoltre, è stata ricavata la percentuale dei casi in cui ai nostri verbi viene applicato il prefisso perfettivo.

	Исследовать (1700-2017)	Поисследовать (1884-1953)	Percentuale della prefissazione
NKRJa	9050	8	0,08 %
Sketch Engine	550.920	427	0,07 %
GIKRJa	45.743	798	1,74 %

Tabella 1. Risultati della ricerca dei verbi "исследовать" e "поисследовать" tramite diversi corpora

	Диагностировать (1877-2017)	Продиагностировать (1997-2013)	Percentuale della prefissazione
NKRJa	266	10	3,75 %
Sketch Engine	69.793	2.345	3,35 %
GIKRJa	5780	558	9,65 %

Tabella 2. Risultati della ricerca dei verbi "диагностировать" e "продиагностировать" tramite diversi corpora

Le conclusioni che possiamo trarre da questi dati sono due: in primo luogo, facendo una media delle percentuali ricavate dai tre corpora, risulta che al verbo “диагностировать” viene applicato il prefisso perfettivo nel 5,58 % dei casi mentre al verbo “исследовать” solamente nello 0,6 %. Quest’ultimo, dunque, tende a mantenere la sua bi-aspettualità, al contrario del primo la cui integrazione al sistema aspettuale è molto più marcata. Notiamo però come la prefissazione avviene in percentuale nettamente più alta nel corpus GIKRJa per entrambi i verbi (1,74 % per “поисследовать” e 9,65 % per “продиагностировать”). Questo ci indica che il fenomeno è più frequente nel linguaggio del web appartenente a blog e social network, tipico di questo corpus, rispetto alla lingua scritta sia letteraria che del web in senso più ampio; in secondo luogo, nonostante il verbo “исследовать” sia presente nella lingua russa da oltre un secolo e mezzo prima rispetto a “диагностировать”, e anche la sua integrazione al sistema aspettuale sia avvenuta prima rispetto al secondo verbo (per lo meno secondo i dati del corpus), pare che l’integrazione abbia attecchito maggiormente con il verbo “диагностировать”. Vediamo infatti come il verbo prefissato “продиагностировать” sia rimasto in uso con costanza fino ai nostri anni mentre il verbo “поисследовать” sia rimasto fermo agli anni Cinquanta del secolo scorso. Ovviamente questi non sono dati assoluti, il verbo “поисследовать” è continuato ad esistere anche oltre questi anni, e lo provano i dati dei web-corpora il cui materiale è piuttosto recente, seppur con utilizzo molto meno frequente. Ciò che però emerge è che il tentativo di prefissazione sia in un certo qual modo fallito entrando in disuso.

Una ragione per cui il verbo prefissato “продиагностировать” ha avuto maggiore successo potrebbe risiedere nel valore semantico del suo prefisso. Mentre il prefisso “по-” del verbo “поисследовать” è il classico prefisso perfettivo e non denota alcun particolare significato, il prefisso “про-”, nei verbi di moto ma non solo, racchiude generalmente in sé l’idea di

attraversamento, in qualche modo anche di raggiungimento di un risultato dopo un processo (vedasi ad esempio il verbo “читать-прочитать” per cui l’aver terminato di leggere un testo ha implicato un processo di lettura pagina dopo pagina).

Il processo diagnostico, in particolare, non è un qualcosa di realizzabile in maniera immediata o con un una semplice azione ma si tratta di una serie di azioni, esami, valutazioni prima di arrivare ad una conclusione. Vediamone alcuni esempi (Fig. 33):

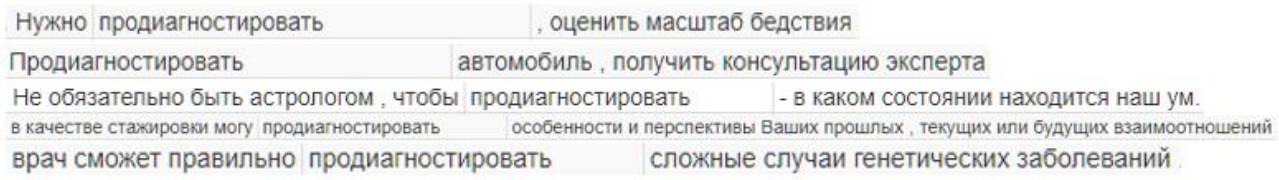


Figura 31. Alcuni esempi di utilizzo del verbo "продиагностировать" tramite il GIKRJa

Nel caso dell’automobile si parla di una diagnosi che richiede il consulto di un esperto, che dovrà andare più a fondo nel capire quali problemi il veicolo possa avere; nel caso del dottore che ha individuato una malattia genetica, si intuisce dall’utilizzo dell’espressione “сложные случаи” che non è stato facile riscontrarla e che sono stati necessari diversi esami per arrivare alla diagnosi finale. Quindi, il verbo "продиагностировать" potrebbe aver continuato il suo corso dal momento che un perfettivo con questo prefisso rende molto bene il concetto di “effettuare ad una diagnosi”.

Per quanto riguarda l’analisi diacronica dei verbi, l’annotazione meta-testuale del NKRJa è stata fondamentale. I corpora del web, invece, si sono rivelati uno strumento molto utile per lo studio dei verbi bi-aspettuali in quanto permettono di stabilire quanto essi rimangano tali e quanto invece si integrino nel sistema aspettuale con l’acquisizione di un prefisso perfettivo.

4.1.2 Collocazioni

Prendiamo l’esempio di una collocazione come ulteriore elemento di comparazione tra corpora tradizionali e del web. Per questa analisi abbiamo scelto la collocazione “носить имя”, che significa “portare il nome” di qualcuno o qualcosa. Ci serviremo dei corpora linguistici in primo luogo per osservare alcuni esempi del suo utilizzo e capire come e in che contesti la ritroviamo, e in secondo luogo per la ricerca di altre collocazioni che vedono la presenza del sostantivo “имя”, così da avere una conoscenza più ampia dell’uso di questo sostantivo.

NKRJa. Cercando la nostra collocazione nel NKRJa possiamo osservare, scorrendo tra i risultati, che essa può essere utilizzata con:

- nomi propri (Fig. 34)

Коровьев, — хозяйка бала должна непременно **носить имя** Маргариты, во-первых, а во-вторых
Такая женщина просто обязана **носить имя** Маргарита, а вовсе не
Борющейся за право **носить имя** Василия Ивановича Чапаева!

Figura 32. Risultati della ricerca della collocazione "носить имя" tramite il NKRJa

- titoli (Fig. 35)

до того, что негоже столице **носить имя** чужого короля, когда есть
никто в Афинах не достоин **носить имя** «царь» — отныне глава государства
не хотите сделать им честь **носить имя** их короля»), а что

Figura 33. Risultati della ricerca della collocazione "носить имя" tramite il NKRJa

- titoli seguiti da un nome proprio (Fig. 36)

бывшего Советского Союза удостоилась чести **носить имя** командарма Лизюкова и руками
СССР с 1929 г. стал **носить имя** начдива Василия Чапаева.
торжественно объявил: «Открытый остров будет **носить имя** создателя русского флота Петра

Figura 34. Risultati della ricerca della collocazione "носить имя" tramite il NKRJa

- nomi comuni di persona (Fig. 37)

рука об руку все достойные **носить имя** человека.
судьба отняла у вас право **носить имя** вашего отца.
о том, достойна ли она **носить имя** его любимой тетки.

Figura 35. Risultati della ricerca della collocazione "носить имя" tramite il NKRJa

Queste sono alcune tra le principali osservazioni che si possono compiere con i risultati ottenuti dal NKRJa. Il totale delle occorrenze per questa collocazione è di 56, un numero non elevato ma sufficiente per avere un'idea dei contesti in cui viene utilizzata e allo stesso tempo accettabile per poterle analizzare tutte.

Lo stesso tipo di ricerca può essere svolto anche con un qualsiasi web corpus, con la differenza che l'utente potrà osservare solo una parte dei risultati, dal momento che saranno molto numerosi per

via delle dimensioni di questi corpora. Ad esempio, i risultati per questa collocazione tramite Sketch Engine e Aranea sono 42 mila, e tramite il GIKRJa 1.676.

Per una ricerca di questo tipo sono uno strumento piuttosto valido, seppur con qualche limite, anche i motori di ricerca, con i quali si avranno risultati ancora più numerosi. Con Google abbiamo ad esempio 1,3 milioni di riscontri. Questi sono alcuni esempi (Fig. 38).

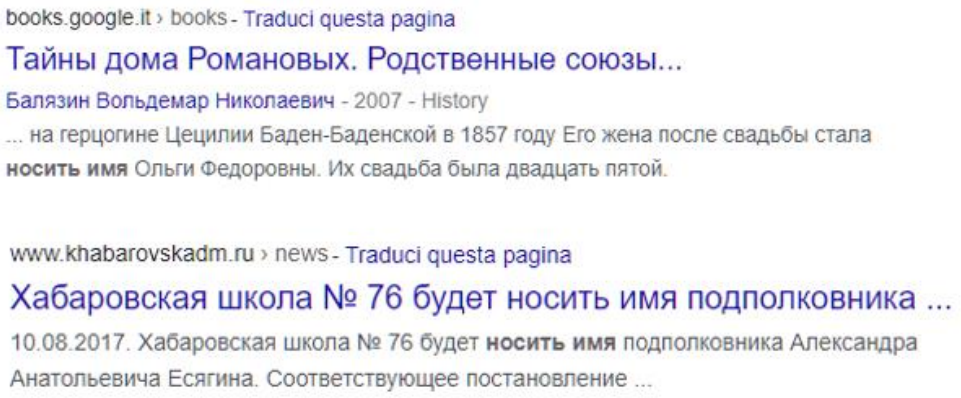


Figura 36. Risultati della ricerca della collocazione "носить имя" tramite Google

Per via del formato con cui i risultati vengono mostrati, meno allineato rispetto al formato KWIC, l'osservazione sarà meno immediata. Per evitare questo problema si può ricorrere all'utilizzo del software WebCorp, il quale recupera il materiale dai motori di ricerca ma li mostra in formato *concordancer*, che risulta più adatto all'osservazione dei risultati (Fig. 39).

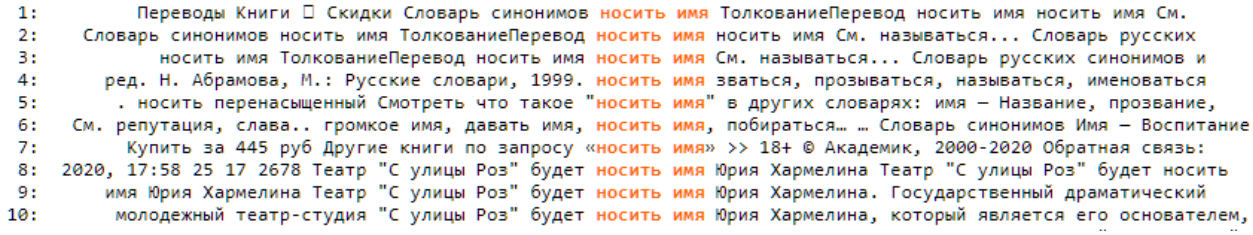


Figura 37. Risultati della ricerca della collocazione "носить имя" tramite WebCorp

Inoltre, sono molti i casi in cui la collocazione non si troverà all'interno di un vero e proprio testo ma in siti di dizionari o traduttori automatici, come vediamo di seguito (Fig. 40).

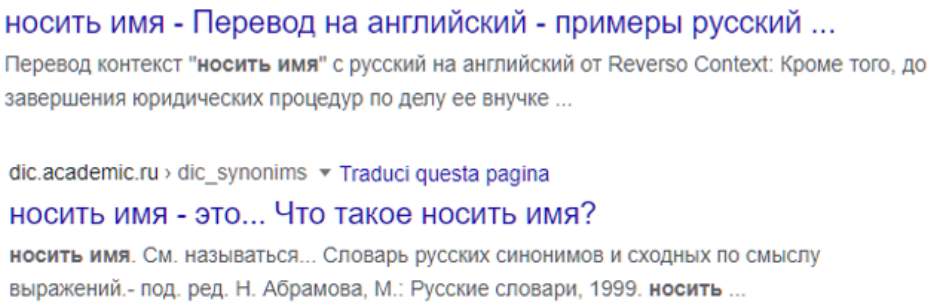


Figura 38. Risultati della ricerca della collocazione "носить имя" tramite Google

Come abbiamo spiegato nel capitolo precedente, infatti, il web contiene molto rumore, ovvero elementi che non sono utili ai fini di una ricerca di tipo linguistico. Per questo motivo, tali problemi non potranno essere risolti nemmeno con l'utilizzo di WebCorp, nel quale il materiale irrilevante sarà comunque presente. Starà quindi all'utente ignorare tali risultati. Perciò, se in molti casi il più grande vantaggio del web è la quantità elevata di contenuti, e di conseguenza di risultati, in questo caso non sono i numeri a rendere una ricerca migliore.

Un'ultima differenza riscontrata tra i vari corpora sta nella funzione *case-sensitive*. Questa risulta attiva nel NKRJa e nel GIKRJa per quanto riguarda la ricerca di una collocazione (a differenza della ricerca di un singolo verbo o sostantivo, per cui vengono restituiti in tutte le loro forme) ma anche in Google per via dell'operatore del *perfect match*, per cui il verbo è presente solamente all'infinito. Gli altri web corpora, invece, restituiscono la collocazione in tutti i tempi verbali. Questi ultimi permettono quindi un'osservazione leggermente più ampia per quanto riguarda l'utilizzo di una collocazione. Questa differenza, però, non dipende dalla tipologia ma dalle funzionalità dei singoli corpora.

Sostanzialmente, per concludere, nell'analisi di una collocazione si potranno trarre le stesse conclusioni da corpora tradizionali, corpora del web e motori di ricerca. Ciò che però vogliamo sottolineare è che: un corpus tradizionale è uno strumento valido e sufficiente ai fini di questa tipologia di ricerca; i corpora del web sono ugualmente validi ma non necessari in presenza dei primi, dato che il numero elevato di risultati risulta inutile; i motori di ricerca sono più confusionari nella visione complessiva rispetto ai primi due, per cui è necessario ricorrere ad un *concordancer* come WebCorp. In entrambi i casi, però, sarà presente dal materiale irrilevante.

Nella seconda parte della nostra ricerca, che riguarda l'individuazione di altre collocazioni per il sostantivo "имя", il processo da svolgere sarà lo stesso con l'ausilio di qualsiasi corpus, tradizionale o web che sia, ad eccezione di uno. L'unica strada al raggiungimento di questo obiettivo sarà quella

di ricercare tramite un corpus il solo sostantivo “имя” e, scorrendo tra i risultati, cercare possibili collocazioni che lo includano. Questo è sicuramente meno immediato e richiede tempi più lunghi rispetto ad altre ricerche. Il sostantivo in questione è oltretutto molto diffuso e i dati da analizzare saranno parecchi. Proviamo, a scopo esemplificativo, ad analizzare le prime dieci pagine di risultati per questo sostantivo e a vedere quante collocazioni riusciremo a ricavare.

NKRJa. I risultati per la ricerca del sostantivo “имя” tramite il NKRJa sono 49.241. Ciò che ci è stato possibile osservare dalle prime dieci pagine è che esso può essere utilizzato assieme a verbi quali: давать-дать (dare, assegnare un nome), делать-сделать (себе) (fare/farsi un nome), вычёркивать-вычеркнуть (cancellare un nome), писать-написать (scrivere un nome), увековечивать-увековечить (perpetuare un nome), скандировать (scandire un nome), узнавать-узнать (conoscere un nome), получать-получить (ricevere un nome), называться (dire/fare un nome), произносить- произнести e выговаривать-выговорить (pronunciare un nome) (Fig. 41).

тот же. Ему издавна дали **имя** Бык. Ураган Бык врывался внезапно
был успех, он сделал Кеплеру **имя** , но успех неполный и отчасти
как-то слишком подчеркнуто), вычеркнуть его **имя** из памяти, учебников и современного
начертила мужской контур, написала моё **имя** и дату рождения.
Хорст Дреслер-Андерс сделал себе **имя** многочисленными статьями в искусствоведческих изданиях
находил $ 100, чтобы увековечить своё **имя** на кирпичике, вмонтированном у входа
поддерживали стража ворот, скандируя его **имя** .
кто пробился в плей-офф, узнает **имя** своего соперника.
для детей и юношества, получивший **имя** этого известного русского живописца.
Причём в документе называлось **имя** конкретного страховщика ответственности— Международная страховая
Считалось, что, произнеся её **имя** , накликаешь беду.
Ирина даже не захотела выговорить **имя** "Олег".

Figura 39. Risultati della ricerca del sostantivo "имя" tramite il NKRJa

Il procedimento sarà lo stesso anche con l’utilizzo di un web corpus. Uno di questi che però si presta particolarmente ad una ricerca di questo tipo è Sketch Engine.

Sketch Engine. Grazie alla sua funzione *word sketch*, che abbiamo introdotto nel dettaglio nel secondo capitolo, basterà inserire il sostantivo “имя” per avere una panoramica generale e completa di tutte le sue collocazioni (Fig. 42).

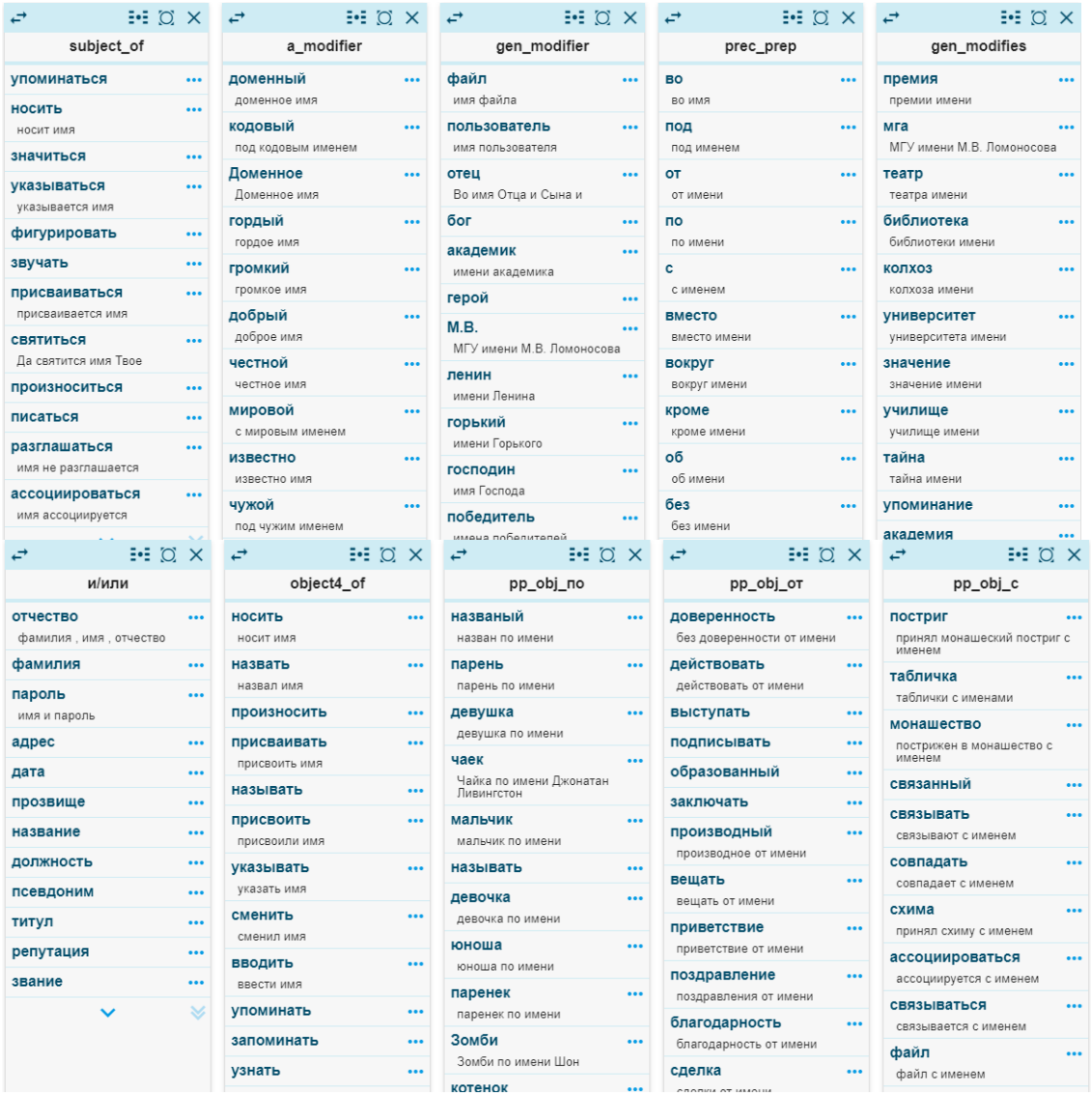


Figura 40. Risultati della ricerca del sostantivo "имя" tramite Sketch Engine

Il numero di collocazioni individuate per il sostantivo “имя” è decisamente superiore rispetto a quelle individuate manualmente. Inoltre abbiamo in unico luogo, oltre ai verbi, anche gli oggetti a cui può riferirsi e tutta una serie di altri elementi di cui riportiamo qui solo un campione. Considerando che i risultati per tale sostantivo sono 5 milioni, eseguire manualmente un’operazione del genere sarebbe impossibile. Il vantaggio assoluto di questa funzione è l’immediatezza con cui ci

viene mostrato il comportamento di una parola e tutte le sue collocazioni. Per questo tipo di ricerca il software Sketch Engine non ha eguali.

4.1.3 Anglicismi

Per testare la ricerca di anglicismi tramite i corpora utilizzeremo il verbo “лайкать-лайкнуть”. Questo deriva dall’inglese *like* ed ha il significato di mostrare approvazione sui social network cliccando sul tipico pulsante a forma di cuore o di pollice alzato. Una volta assorbito dalla lingua russa, il verbo ha acquisito la forma tipica del verbo russo con tanto di aspetto perfettivo e imperfettivo, discostandosi da quei verbi di origine straniera che, come abbiamo detto prima, mantengono un'unica forma bi-aspettuale. Dal momento che nei principali dizionari russi questo verbo non è ancora presente, un corpus può essere utile per capire meglio il suo funzionamento, la sua reggenza e i contesti in cui può essere utilizzato.

NKRJa. Ricercando all’interno del corpus il verbo “лайкать” nella sua forma imperfettiva, abbiamo cinque risultati (escludendo il penultimo esempio in quanto si tratta di un caso di omonimia con il verbo “abbaiare”) (Fig. 43).

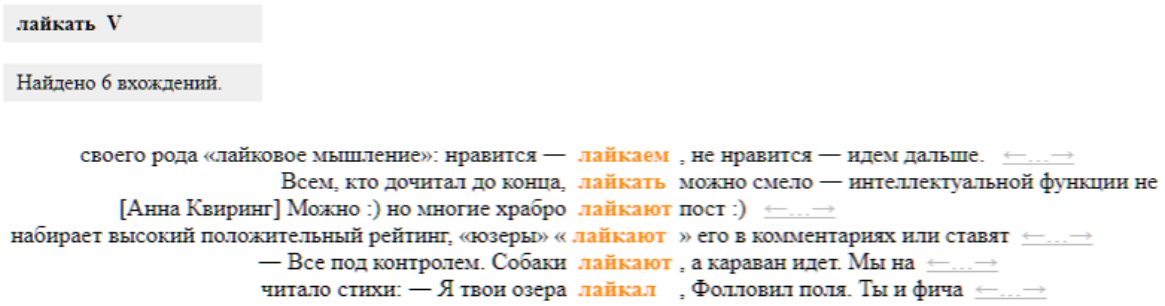


Figura 41. Risultati della ricerca del verbo "лайкать" tramite il NKRJa

Nella sua forma perfettiva abbiamo invece sei risultati (anche in questo caso escludendo l’ultimo esempio, che riflette una forma omonima) (Fig. 44).

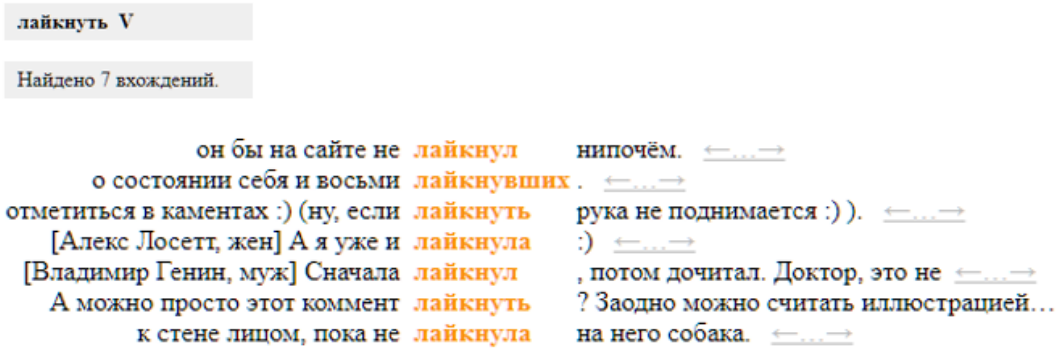


Figura 42. Risultati della ricerca del verbo "лайкать" tramite il NKRJa

Con un totale di 11 occorrenze le considerazioni che si possono trarre sul comportamento di questo verbo non sono molte, se non che regge il caso accusativo e che come complemento oggetto può avere il sostantivo “пост”.

ruTenTen. Se cerchiamo il verbo “лайкать” nel corpus russo di Sketch Engine la prima cosa che si nota è che il numero dei risultati è decisamente maggiore rispetto al NKRJa. La ricerca tramite questo corpus è *case-sensitive*, e, di conseguenza, verranno ritrovati i verbi che corrispondono alla forma che abbiamo inserito nella query. Ricercando il verbo “лайкать”, quindi, troveremo solo la sua forma all’infinito. I risultati sono 216. All’aspetto perfettivo abbiamo invece 147 risultati (per quanto riguarda i web-corpora non abbiamo tenuto conto dei casi di omonimia con il verbо “abbaiare” dal momento che per numeri così alti si tratta di una percentuale minima). Nonostante questa mancanza, i risultati sono decisamente un numero più alto e le osservazioni che si possono fare di questo verbo sono comunque molte.

Scopriamo così che è possibile:

- scambiarsi a vicenda questo gesto di approvazione sulle varie piattaforme (Fig. 45);

Figura 43. Risultati della ricerca del verbo "лайкать/лайкнуть" tramite Sketch Engine

- chiedere o invitare qualcuno a farlo per un nostro contenuto (Fig. 46);

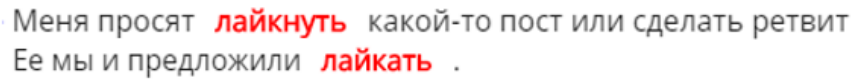


Figura 44. Risultati della ricerca del verbo "лайкать/лайкнуть" tramite Sketch Engine

- farlo attivamente (Fig. 47)

Figura 45. Risultati della ricerca del verbo "лайкать/лайкнуть" tramite Sketch Engine

- o dandogli pieno valore (Fig. 48);

Figura 46. Risultati della ricerca del verbo "лайкать/лайкнуть" tramite Sketch Engine

- “lanciarli”, ovvero lasciarli senza troppa attenzione a diversi contenuti (Fig. 49);

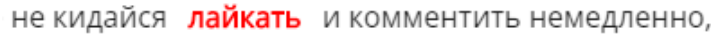


Figura 47. Risultati della ricerca del verbo "лайкать/лайкнуть" tramite Sketch Engine

- lasciarli a contenuti quali video, foto, post, pagine o articoli (Fig. 50);

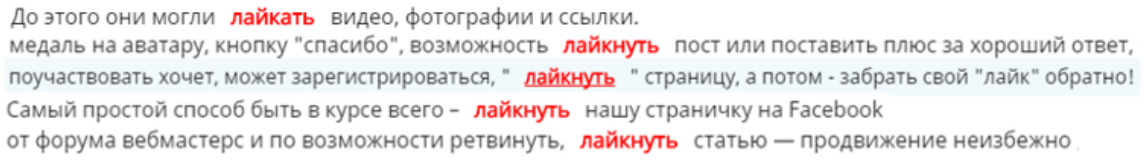


Figura 48. Risultati della ricerca del verbo "лайкать/лайкнуть" tramite Sketch Engine

Rispetto a Sketch Engine, però, con i corpora Aranea e GIKRJa abbiamo un numero decisamente maggiore di risultati. Rispettivamente per la forma imperfettiva e perfettiva abbiamo 2.666-1.930 occorrenze con Aranea e 7.034-9.257 occorrenze con il GIKRJa. Dal GIKRJa abbiamo inoltre la possibilità, grazie alle informazioni meta-testuali, di individuare gli anni in cui questo anglicismo ha iniziato ad emergere nella lingua russa. Lo vediamo per la prima volta nel 2009 con tre occorrenze mentre a partire dal 2010 ricorre con frequenza sempre maggiore fino ad oggi.

Per una questione puramente numerica proviamo ora a svolgere la stessa ricerca con il motore di ricerca Google. I numeri qui sono davvero alti. Solamente all'infinito il verbo supera i due milioni di occorrenze in entrambi gli aspetti (Fig. 51).

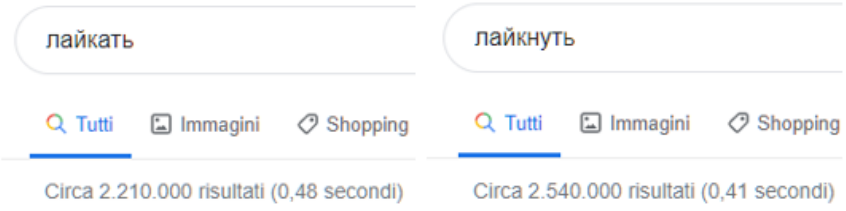


Figura 49. Risultati della ricerca del verbo "лайкать/лайкнуть" tramite Google

Se con Sketch Engine il verbo "лайкать-лайкнуть" poteva sembrare ancora presente in maniera marginale nella lingua russa, con gli altri due corpora del web ma soprattutto con Google capiamo quanto questo sia un anglicismo ormai diffuso.

Grazie all'osservazione di questi risultati abbiamo inoltre scoperto, da un'occorrenza riscontrata in Sketch Engine, che in russo esiste anche, se pur ancora molto raro, l'anglicismo per l'antonimo del verbo *like*, ovvero *unlike*. Questo è "анлайкать-анлайкнуть" (Fig. 52).

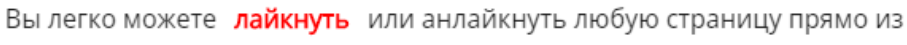


Figura 50. Verbo "анлайкнуть" rilevato tramite Sketch Engine

Abbiamo quindi approfondito la ricerca anche per questo verbo. Nel NKRJa, com'era prevedibile, non è presente. Nei web corpora lo troviamo invece solamente una volta all'aspetto imperfettivo e una all'aspetto perfettivo in ruTenTen (Fig. 53),

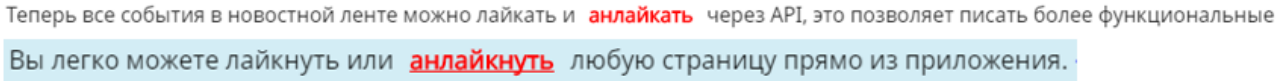


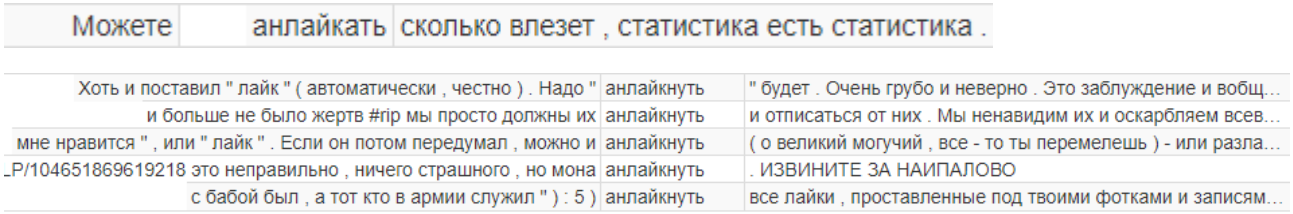
Figura 51. Risultati della ricerca del verbo "анлайкать-анлайкнуть" tramite Sketch Engine

ed una volta all'imperfettivo in Aranea (Fig. 54).

Figura 52. Risultati della ricerca del verbo "анлайкать-анлайкнуть" tramite Aranea

Una speranza di vedere numeri anche di poco maggiori la riponiamo nel GIKRJa. Il suo materiale appartiene infatti al web più “profondo” ovvero social network e blog, e il nostro verbo proviene proprio da questi ambiti.

Di questo, all’infinito, sono risultate un’occorrenza alla forma imperfettiva e cinque alla forma perfettiva (Fig. 55)



Можете	анлайкать	сколько влезет , статистика есть статистика .	
Хоть и поставил " лайк " (автоматически , честно) . Надо "	анлайкнуть	" будет . Очень грубо и неверно . Это заблуждение и вобщ…	
и больше не было жертв #rip мы просто должны их	анлайкнуть	и отписаться от них . Мы ненавидим их и оскарбляем всев…	
мне нравится " , или " лайк " . Если он потом передумал , можно и	анлайкнуть	(о великий могучий , все - то ты перемелешь) - или разла…	
_P/104651869619218 это неправильно , ничего страшного , но мона	анлайкнуть	. ИЗВИНИТЕ ЗА НАИПАЛОВО	
с бабой был , а тот кто в армии служил ") : 5)	анлайкнуть	все лайки , проставленные под твоими фотками и записям…	

Figura 53. Risultati della ricerca del verbo "анлайкать-анлайкнуть" tramite il GIKRJa

Abbiamo effettivamente un numero leggermente maggiore di risultati. Si tratta di un anglicismo più recente rispetto al precedente, che è apparso per la prima volta in questo corpus nel 2013. Per i fenomeni più rari, quindi, il GIKRJa è in grado di restituire più risultati rispetto agli altri corpora del web.

Come ultima fonte per la ricerca di questo verbo raro utilizziamo il motore di ricerca Google (Fig. 56).

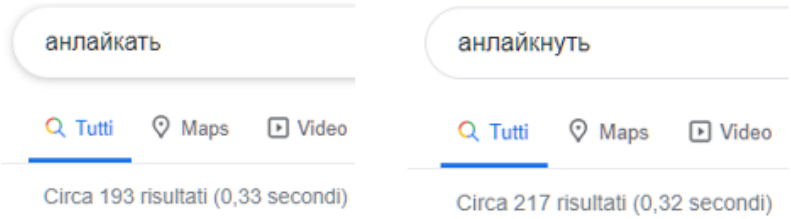


Figura 54. Risultati della ricerca del verbo "анлайкать-анлайкнуть" tramite Google

I numeri sono decisamente superiori rispetto ai web corpora ma, in una fonte dalle dimensioni come quelle del web, si tratta di numeri infinitesimali. Ciò conferma ulteriormente la rarità di questo verbo e ci permette di concludere che esso non sia entrato in uso tanto quanto лайкать-лайкнуть. La sua introduzione nella lingua russa è avvenuta ma si trova ancora ai primi stadi. È però probabile

che tra qualche anno la sua presenza aumenterà. Per quanto riguarda la ricerca di questi due verbi, l'integrazione dei motori di ricerca ai corpora del web ci ha da una parte dimostrato la diffusione del primo, mentre dall'altra ci ha confermato la rarità del secondo. Limitare una ricerca ad un solo strumento può quindi essere fuorviante, specialmente per fenomeni nuovi o rari.

Rimanendo nell'ambito degli anglicismi, facciamo un altro esempio con il termine "оффтоп", dall'inglese *off-topic*, che indica un qualcosa che risulta essere fuori tema in un determinato contesto. Questo anglicismo è ormai penetrato piuttosto solidamente nella lingua russa, così come in quella italiana. In particolar modo, ma non solo, è molto diffuso nel web in ambienti quali forum e pagine dei social network, in cui gli utenti si ritrovano ad intavolare discussioni e a scambiarsi domande su un determinato argomento. È proprio in questo contesto che spesso gli utenti avvisano gli altri del fatto che stanno per parlare o fare domande su un argomento che si discosta da quello su cui verte la conversazione, affermando appunto che si tratta di qualcosa *off-topic*. Questa volta vediamo, in termini prettamente numerici, come si comportano i vari corpora di lingua russa.

NKRJa. Dalla ricerca per questo anglicismo emergono 40 occorrenze. Capiamo subito che si tratta di un termine che ha fondato radici piuttosto solide nella lingua russa. Analizzando i dati meta-testuali di questo anglicismo vediamo infatti che il primo testo in cui compare risale all'anno 2004. A differenza di "лайкать-лайкнуть" l'uso di questo anglicismo è dunque molto meno recente.

Aranea. Con il web corpus Aranum Russicum Maximum abbiamo ovviamente un numero nettamente superiore di occorrenze, ovvero 8,614.

Sketch Engine. ruTenTen ci restituisce un numero di risultati ancora più alto, ovvero 11.824.

GIKRJa. Abbiamo qui 19.728 occorrenze, un numero che supera i precedenti in maniera abbastanza rilevante. Il GIKRa è un corpus i cui contenuti provengono proprio dai luoghi in cui questo anglicismo è più diffuso. Si è infatti rivelato, tra gli altri web corpora, quello in grado di restituire più occorrenze. Possiamo quindi confermare che il GIKRJa è lo strumento che offre maggiori risultati sia per quanto riguarda gli anglicismi.

Anche in questo caso integriamo l'utilizzo dei motori di ricerca. I risultati per questo termine tramite Google sono oltre 5 milioni, il che ci dà ulteriore prova del fatto che si tratta di un anglicismo piuttosto diffuso.

Ad esprimere lo stesso concetto di “оффтоп” esiste in russo anche l’anglicismo “оффтопик”. Abbiamo svolto una ricerca tramite il NKRJa e i corpora del web per verificare quale dei due, a parità di significato, sia il più frequente nella lingua russa. Riassumiamo i dati in questa tabella (Tab. 3).

	Оффтоп	Оффторик
NKRJa	40	14
Aranea	8.614	3.205
Sketch Engine	11.824	2.412
GIKRJa	19.728	4.201
Google	5,1 milioni	2,6 milioni

Tabella 3. Numeri delle occorrenze degli anglicismi "оффтоп" e "оффтопик" tramite diversi corpora

Nonostante l’anglicismo “оффтопик” si avvicini di più al termine inglese da cui deriva, notiamo che nella lingua russa il più utilizzato è invece “оффтоп”.

4.1.4 Termini gergali

Per la ricerca di termini che fanno parte dello slang russo abbiamo scelto due termini gergali della parola *den’gi*, che sono “бабло” e “бабки”, i quali solitamente indicano denaro sporco, ottenuto illegalmente, ciò che in italiano chiameremmo “la grana”. I due sono stati ricercati con il NKRJa, con i principali web corpora, ovvero ruTenTen, Aranea e il GIKRJa e con i motori di ricerca Google e Yandex. Vediamo in questa tabella il numero di occorrenze dei due termini ricavati dai diversi strumenti (Tab. 4 e 5).

	Деньги	Бабло	Percentuale di utilizzo dello slang rispetto al termine neutro
NKRJa	74.892	83	0,11 %
Sketch Engine	6 milioni	41.223	0,68 %
Aranea	7,5 milioni	43.002	0,57 %
GIKRJa	4,9 milioni	82.715	1,68 %
Google	463 milioni	10,8 milioni	2,33 %
Yandex	29 milioni	9 milioni	31 %

Tabella 4. Numero delle occorrenze del termine gergale "бабло" tramite diversi corpora

	Деньги	Бабки	Percentuale di utilizzo dello slang rispetto al termine neutro
NKRJa	74.892	2.881	3,8 %
Sketch Engine	6 milioni	73.327	1,22 %
Aranea	7,5 milioni	65.792	0,87 %
GIKRJa	4,9 milioni	98.806	2,01 %
Google	463 milioni	12,3 milioni	2,65 %
Yandex	29 milioni	5 milioni	17,2 %

Tabella 5. Numero delle occorrenze del termine gergale "бабки" tramite diversi corpora

La prima cosa che si può notare è la differenza sulla frequenza di utilizzo dei due termini. “Бабки” risulta infatti utilizzato nella lingua russa più frequentemente rispetto a “бабло”. L’unica eccezione sono i risultati del termine “бабло” ottenuti con il motore di ricerca russo Yandex. Ciò può essere dovuto al fatto che “Бабло” è il titolo di un noto film russo, per cui una buona parte del materiale contenente questo termine potrebbe fare riferimento ad esso.

In secondo luogo, vediamo la percentuale di utilizzo dei termini gergali “бабло” e “бабки” rispetto al termine neutro “деньги”. Mentre “бабло” risulta essere più utilizzato nei web corpora e nei motori di ricerca, in cui il linguaggio è più colloquiale, “бабки” è presente in percentuale maggiore nel NKRJa. Deduciamo quindi che quest’ultimo sia un termine gergale molto adoperato nella lingua letteraria russa. Facendo una media dei dati sull’utilizzo dei due termini tratti dalle nostre fonti,

risulta che “бабло” viene sostituito come termine gergale in cambio di “деньги” nell’1,07 % dei casi mentre “бабки” nel 2,11 %[56].

Da ultimo, possiamo invece trarre delle conclusioni riguardo gli ambiti di utilizzo di questi due termini. Rispetto ai web corpora Sketch Engine e Aranea, i risultati ottenuti tramite il GIKRJa sono decisamente più alti. Così come per gli anglicismi, il GIKRJa è in grado di recuperare un gran numero di risultati anche per quanto riguarda i termini gergali. Questi sono infatti particolarmente utilizzati in contesti in cui un linguaggio più colloquiale è preferito, come lo possono essere i blog o i social network, da cui il GIKRJa recupera il suo materiale.

4.1.5 Confronto di due termini con significato simile

Passiamo ora ad una ricerca che vede il confronto tra due termini che condividono parzialmente il significato, ovvero “день” e “сутки”. Ricercando il loro significato in un dizionario di lingua russa, ad esempio nel Kovalev, troveremo che tra i principali significati del primo ci sono “giorno”, “giornata”, “giorno di 24 ore”, mentre l’unico significato per il secondo è “giorno di 24 ore”. Per quest’ultimo, al contrario di “день”, non vengono inclusi esempi del suo utilizzo, il che potrebbe lasciare ad un apprendente di lingua russa dei dubbi riguardo a quando sia meglio utilizzare l’uno o l’altro per esprimere il significato che condividono, ovvero “giorno di 24 ore”. Dal momento che con il solo utilizzo di un dizionario emerge questo limite, ci affideremo ad un corpus per capire meglio gli utilizzi del termine “сутки” (Tab. 5).

	День	Сутки
NKRJa	420.088	20.446
Sketch Engine	20 milioni	1,4 milioni
Aranea	23,8 milioni	1,9 milioni
GIKRJa	10 milioni	445.047

Tabella 6. Numero delle occorrenze dei termini "день" e "сутки" tramite diversi corpora

In primo luogo osserviamo la frequenza di utilizzo dei due termini. Com’era ovvio dato il suo significato decisamente più limitato, “сутки” è molto meno frequente di “день”, che ha invece

[56] Questo dato è stato ottenuto escludendo il motore di ricerca Yandex in quanto, come abbiamo sottolineato, molti dei casi in cui il termine “бабло” viene utilizzato, fa riferimento al famoso film. Questo avrebbe portato a risultati imprecisi.

ambiti di utilizzo molto più ampi. Ciò che però emerge in particolare dal numero di risultati ottenuti con il corpus GIKRJa è che il termine “сутки” ha qui molte meno occorrenze rispetto ad altri web corpora di dimensioni simili, e, proporzionalmente alle sue dimensioni, anche al NKRJa. Ancor prima di osservare i casi in cui viene utilizzato, il termine “сутки” sembrerebbe quindi appartenere ad un linguaggio più formale. Ciò spiegherebbe il fatto che rispetto a corpora generici, che siano web o tradizionali, questo termine è meno presente in un corpus il cui materiale proviene principalmente da blog e social network e che di conseguenza contiene un linguaggio più colloquiale e informale.

Per quanto riguarda l’analisi degli utilizzi di questo termine, vediamone alcuni esempi tratti da diversi corpora. Dai risultati si noterà che il termine “сутки” viene utilizzato per indicare una certa quantità di giorni alla settimana o al mese, come ad esempio i giorni di lavoro o di apertura di un locale, la durata di un’attività oppure per indicare ciò che in italiano è l’espressione “24 ore su 24” (Fig. 57).

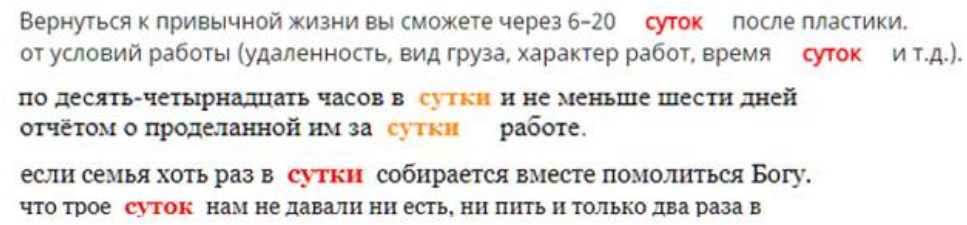


Figura 55. Risultati della ricerca del termine "сутки" tramite diversi corpora

Per questa specifica ricerca non sono da evidenziare particolari differenze rispetto all’uso di corpora tradizionali o del web. Queste considerazioni potranno quindi essere tratte da qualsiasi tipologia di corpus.

4.1.6 Tecnicismi

Concludiamo la nostra analisi con un paragone tra corpora tradizionali e corpora del web per quanto riguarda la ricerca di un termine specialistico. Il termine in questione appartiene all’ambito medico ed è “биопсия” che significa “biopsia”. È infatti utile, per la traduzione o la stesura di un testo specialistico, conoscere ad un livello più approfondito gli utilizzi e le collocazioni di tecnicismi appartenenti ad esso.

Vediamo quindi il numero di occorrenze rilevate dai diversi corpora per questo termine e quale risulti il più consono ai fini della nostra ricerca.

NKRJa. Il NKRJa restituisce 54 occorrenze per questo tecnicismo. Trattandosi di un numero poco elevato è possibile osservarle tutte e trarre le dovute considerazioni sull'uso della parola. Ciò che ci è stato possibile constatare è che:

- I verbi per cui questo termine può fungere da complemento oggetto sono: "брать-взять", "делать-сделать" e "приводить-привести" (Fig. 58)

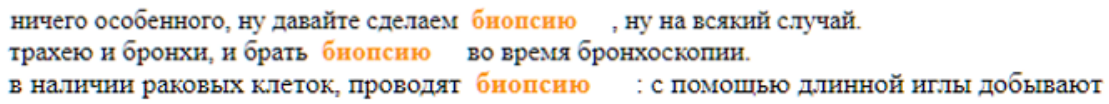


Figura 56. Risultati della ricerca del termine "биопсия" tramite il NKRJa

- Gli attributi a cui può essere associato sono: "пункционная", "прицельная", "аспирационная", "игловая" e "повторная" (Fig. 59)

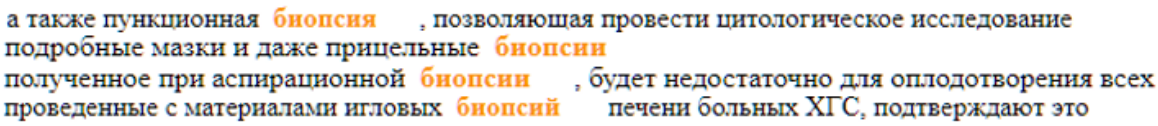


Figura 57. Risultati della ricerca del termine "биопсия" tramite il NKRJa

- I complementi di specificazione possono essere: "хорион" e "печень" (Fig. 60)

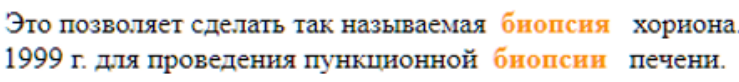


Figura 58. Risultati della ricerca del termine "биопсия" tramite il NKRJa

Questo è tutto ciò che siamo riusciti a ricavare dalla ricerca di questo termine tramite il NKRJa.

Aranea. Con questo web corpus i risultati sono 43.313. La differenza per quanto riguarda il numero delle occorrenze è notevole. Ciò che si può ricavare da una quantità tale di risultati è sicuramente molto di più rispetto a ciò che è emerso dal NKRJa e la conoscenza di questo termine risulterebbe decisamente più approfondita, nonostante il tempo che un'analisi di questo tipo richiede. Ovviamente una sola persona non sarebbe in grado di analizzare tutti i risultati ma sarebbe comunque necessario farlo per una buona parte di essi. Possiamo quindi affermare che un web corpus è in grado di fornire una quantità di informazioni nettamente superiore rispetto ad un corpus tradizionale come il NKRJa.

Sketch Engine. Come abbiamo visto per quanto riguarda l'analisi delle collocazioni, la funzione *word sketch* prevista dal corpus Sketch Engine è preziosa in una ricerca di questo tipo. Il numero di occorrenze rilevate per il termine "биопсия" è simile a quello ottenuto con Aranea, ovvero 39.921.

Ciò che qui fa la differenza è che non è necessaria da parte dell'utente un'analisi voce per voce al fine di ricavare informazioni sull'utilizzo di questo termine, come richiederebbe un qualsiasi altro corpus. Il tutto avviene infatti in maniera automatica e, data la mole di materiale su cui questa analisi automatica avviene, i risultati saranno davvero consistenti.

Se con il NKRJa abbiamo ricavato tre verbi e cinque aggettivi che si possono associare al nostro termine, Sketch Engine ce ne offre un'intera lista, di cui ne riportiamo una parte nell'immagine che segue (Fig. 61).

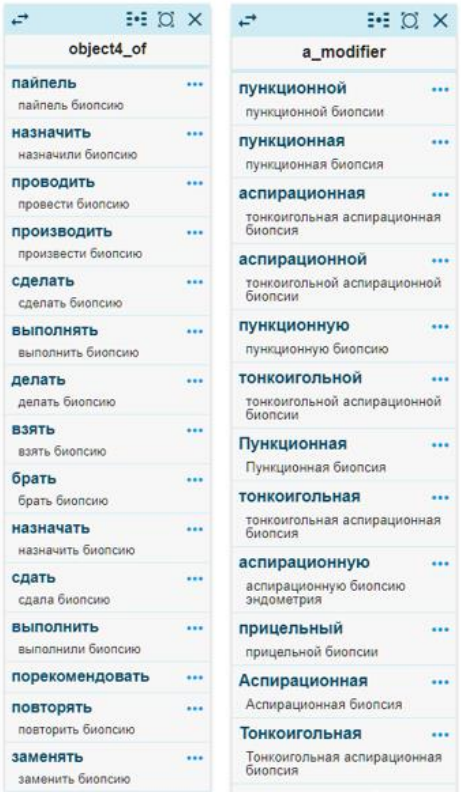


Figura 59. Risultati della ricerca del termine "биопсия" tramite Sketch Engine

Sketch Engine si riconferma dunque lo strumento migliore per un generale approfondimento del comportamento e utilizzo di un termine, non solo per quanto riguarda termini generici ma anche tecnici e specialistici.

GIKRJa. Ciò che abbiamo constatato per il corpus Aranea vale anche per il GIKRJa. Va però sottolineato che i risultati ottenuti da questo web corpus per i termini tecnici sono nettamente inferiori rispetto agli altri. La terminologia scientifica e in generale più tecnica, così come abbiamo avuto modo di vedere con i termini meno colloquiali, è infatti poco ricorrente nelle sezioni del web come quelle da cui il GIKRJa ricava il suo materiale.

CONCLUSIONI

A conclusione di questa nostra ricerca riassumiamo le considerazioni che abbiamo potuto trarre da essa. Quali sono i casi in cui i web corpora sono da preferirsi rispetto ad un corpus tradizionale e vice versa?

I corpora tradizionali sono sicuramente la scelta migliore per ricerche puramente grammaticali. Il NKRJa è dotato di annotazione morfosintattica completa e con margine di errore molto basso, grazie alla quale è in grado di distinguere le parti del discorso come sostantivi, aggettivi e verbi ma anche casi, aspetti e tempi verbali, elementi fondamentali nell'analisi di determinati aspetti grammaticali. I corpora del web, in mancanza di questa funzionalità, sono invece risultati inadatti a ricerche di tipo grammaticale.

Il NKRJa si è rivelato uno strumento superiore ai corpora del web anche nell'analisi di determinati fenomeni linguistici dal punto di vista cronologico. Si differenzia infatti altresì per la sua accurata annotazione meta-testuale, la quale permette di risalire alla data di pubblicazione di ogni testo in esso contenuto. Infine, si distingue dai web corpora per la possibilità di studiare fenomeni linguistici che oggi giorno sono entrati in disuso o che sono utilizzati con frequenza sempre più ridotta, grazie alla consistente quantità di materiale appartenente a secoli passati. Possiamo invece considerarlo a pari livello dei web corpora quando si tratta di osservare l'utilizzo di una collocazione linguistica e di due termini dal significato simile.

I corpora del web, grazie alla ricchezza del loro materiale e alla lingua più nuova e sempre aggiornata rispetto a quelli tradizionali, emergono in particolar modo per l'analisi di fenomeni linguistici quali anglicismi, termini gergali e tecnicismi. Questi sono presenti in misura limitata, specialmente gli anglicismi più recenti, nel NKRJa, il quale risulta meno aggiornato da un punto di vista linguistico. Inoltre, se da un lato i web corpora sono meno indicati per ricerche di tipo grammaticale, si sono dall'altro lato rivelati fondamentali per lo studio della diffusione di alcuni fenomeni grammaticali quali il genitivo partitivo e l'integrazione nel sistema aspettuale dei verbi bi-aspettuali.

Oltre a queste considerazioni generali sui corpora del web, è doveroso spendere alcune parole sui singoli corpora appartenenti a questa categoria. Per quanto riguarda l'osservazione generale del comportamento di una parola e le sue collocazioni, emerge più di tutti Sketch Engine, grazie alla sua funzione *word sketch*. Questa permette di avere in maniera automatica una catalogazione di tutti i comportamenti di una parola tratta da una quantità di materiale davvero consistente, rendendolo

così uno strumento innovativo e unico nel suo genere. In molti casi, infatti, come abbiamo potuto vedere nel corso della nostra ricerca, la grande mole di risultati fornita dai web corpora risulta inutile in quanto una singola persona non sarà in grado di analizzarli tutti manualmente. Parlando invece del GIKRJa, una grande speranza è riposta in questo corpus dal momento che ha tutte le carte in regola per essere uno strumento davvero promettente. Esso permette infatti di unire la ricchezza del materiale web e la possibilità di svolgere ricerche complesse dal punto di vista grammaticale, il che consentirebbe di svolgere analisi approfondite di fenomeni grammaticali tra i più rari o i più recenti. Questo sarà però possibile solamente una volta che il corpus sarà perfezionato. Ciò che di negativo è però emerso dal suo utilizzo, oltre ai tempi piuttosto lunghi di elaborazione delle ricerche e alla formulazione di query complesse poco intuitiva, è il fatto che il settaggio di parametri grammaticali per ricerche sia complesse ma anche molto semplici non funziona. La causa principale di questo problema risiede molto probabilmente nella difficoltà di perfezionare l'annotazione morfosintattica di una tale quantità di materiale. Se una volta che il corpus sarà uscito dalla fase sperimentale questo aspetto sarà migliorato almeno in parte, il GIKRJa sarà sicuramente uno tra gli strumenti più interessanti per quanto riguarda lo studio della lingua russa contemporanea. Al di là di queste problematiche, il corpus si è infatti rivelato il migliore e il più performante per la ricerca di anglicismi e slang tra i più rari, e, grazie all'annotazione meta-testuale più accurata rispetto a quella morfosintattica, anche nell'analisi cronologica di un fenomeno linguistico.

Ciò che infine possiamo dire riguardo ai motori di ricerca è che questi possono essere utili, nonostante i loro limiti, per trarre considerazioni generali sull'uso di termini e collocazioni ma soprattutto per ricerche che prevedono la comparazione della frequenza di determinati fenomeni linguistici, in particolare quelli che ancora sono presenti in maniera esigua nei corpora tradizionali.

BIBLIOGRAFIA

- A. Zaliznjak, A. D. Šmelev (2000). *Vvedenie v russkuju aspektologiju*, Moskva: Jazyki russkoj kul'tury.
- Asinovskij A., Bogdanova N., Rusakova N., Ryko A., Stepanova S., Šerstinova T. (2009). *'The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation'*, in *International Conference on Text, Speech and Dialogue*, pp. 250-257.
- Atkins S., Clear J., Osler N. (1992) 'Corpus Design Criteria', in *Literary and Linguistic Computing* 7(1), pp. 1-16.
- Baker P., Hardie A., McEnery T. (2006). *A Glossary of Corpus Linguistics*, Edinburgh: Edinburgh University Press.
- Barbera M. (2013). *Linguistica dei corpora e linguistica dei corpora italiana*. *Un'introduzione*, Milano: Qu.A.S.A.R.
- Barbera M., Corino E., Onesti C. (2007). 'Cosa è un corpus? Per una definizione più rigorosa di corpus, token, *markup*', in *Corpora e Linguistica in rete*, a cura di Barbera M., Corino E., Onesti C. Perugia: Guerra edizioni, pp. 25-88.
- Baroni M., Bernardini S. (2004). 'BootCat: Bootstrapping corpora and terms from the web', in *Proceedings of LREC 2004*, Lisbon: ELDA, pp. 1313-1316.
- Baroni M., Bernardini S. (2006). *Wacky! Working Papers on the Web as Corpus*, Bologna: Gedit.
- 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, pp. 209-26.
- Baroni M., Ueyama M. (2006). 'Building general- and special-purpose corpora by Web crawling', in *Proceedings of the NIJL International Symposyum, Language Corpora: Their Compilation and Application*, pp. 31-40.
- 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), pp. 79-93.
- Bergh G. (2005). 'Min(d)ing English language data on the Web. What can Google tell us?', in *ICAME Journal*, vol. 29, pp. 25-46.

- Bernardini S. (2006). ‘Corpora for translator education and translation practice. Achievements and challenges’, in *Proceedings of LREC 2006,* pp. 17-22.
- Biber D. (1993). ‘Representativeness in Corpus Design’, in *Literary and Linguistic Computing*, Volume 8, Issue 4, pp. 243–257.
- Biber D., Conrad S., Reppen R., Aitchison J. (1998). *Corpus Linguistics: Investigating Language Structure and Use,* Cambridge: Cambridge University Press.
- Bocharov V., Bichineva S., Granovskii D., Ostapuk N., Stepanova M. (2011). ‘Quality assurance tools in the opencorpora project’, in *Komp’juternaja lingvistika i intellektual’nye technologii*, pp. 101-109.
- Bowker L., Pearson J. (2002). *Working with Specialized Language. A Practical Guide to Using Corpora,* London-New York: Routledge.
- Brezina V. (2012). ‘Google Scholar as a linguistic tool: new possibilities in EAP’, in *The future of Applied Linguistics: Local and Global Perspective,* a cura di C. Gkitsaki, R. Baldauf, Newcastle upon Tyne: Cambridge Scholars Publishers, pp. 26-48.
- Che Chen, N. (2013). ‘Uzus, norma i sistema v kontekste sovremennogo russkogo jazyka: na materiale internet-kommunikacii’, in *Mir russkogo slova 3,* pp. 33-43.
- Crystal D. (2011). *Internet Linguistics. A Student’s Guide,* London: Routledge.
- De Schryver G. (2002). ‘Web for/as corpus: a perspective for the African languages’, in *Nordic Journal of African Studies,* vol. 11(2), pp. 266-82.
- Dizionario Corriere della Sera https://dizionari.corriere.it/
- Dizionario Edizioni Giuridiche Simone https://www.simone.it/newdiz/
- Dobrušina E. R., Poljakov A. E. (2013). ‘Korpus cerkoslavjanskogo jazyka: vozmožnosti, metody cozdanija, perspektivy’, in *Vestnik PSTGU* vol. 3: filologija, ed. 1 (31), pp. 32-44.
- Ferraresi A. (2009). ‘Google and beyond: web as corpus methodologies for translators’, in *Tradumàtica: traducció i tecnologies de la informació i la comunicació,* vol. 7.
- Fletcher W. (2007). ‘Concordancing the Web. Promise and problems, tools and techniques’, in *Corpus Linguistics and the Web*, a cura di M. Hundt, N. Nesselhauf, C. Biewer, Amsterdam: Rodopi, pp. 25-46.
- Francis W. Nelson. (1982). ‘Problems of assembling and computerizing large corpora’ in Computer *Corpora in English Language Research*, a cura di S. Johansson, Bergen: Norwegian Computing Centre for the Humanities, pp. 7-24.

- Gamallo Otero P., Gonzalez Lopez I. (2008). 'Wikipedia as a Multilingual Source of Comparable Corpora', in *Workshop on Building and Using Comparable Corpora, LREC* 2010, pp. 19-26.
- Gatto M. (2014). *Web As Corpus. Theory and Practice*. London-New York: Bloomsbury.
- Grishina E. (2010). 'Multimodal Russian Corpus (MURCO): First steps', in *7th Conference on Language Resources and Evaluation*, Valletta, Malta, pp. 2953-60.
- Grishina E. (2011). 'Mul'timedijnyj russkij korpus (MURKO): sovremennoe sostojanie i perspektivy razvitja', in *Trudy Meždunarodnoj Konferencii "Korpusnaja lingvistika – 2011*", pp. 138-144.
- Grišina E., Savčuk S. (2009). 'Korpus ustnyh tekstov v NKRJa: sostav i struktura', in: *Nacional'nyj korpus russkogo jazyka: 2006—2008. Novye rezul'taty i perspektivy*, a cura di Vladimir A. Plungjan. Sankt Peterburg: Nestor-istorija, pp. 129–149.
- Henzinger M., Lawrence S. (2004). 'Extracting Knowledge from the World Wide Web', in *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5186-91.
- Hunston S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Johansson S. (1991). 'Computer Corpora in English Language Research', in *English Computer Corpora. Selected Papers and Research Guide*, a cura di Johansson S., Stenström A., Berlin-New York: Mouton de Gruyter.
- Kachkovskaia T., Kocharov D., Skrelin P., Volskaya 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. 1949-54.
- Kennedy G. (1998). *An introduction to corpus linguistics*. London-New York: Longman.
- Kilgarriff A. (2001). 'Web as corpus', in *Proceedings of the Corpus Linguistics Conference (CL 2001), University Centre for Computer Research on Language Technical Paper*, vol. 13, Special Issue, Lancaster University, pp. 342-44.
- Kilgarriff A., Grefenstette G. (2003). *Introduction to the Special Issue on the Web as Corpus*, in Computational Linguistics vol. 29(3), pp. 333-47.
- Kilgarriff A., Rychly P., Smrz P., Tugwell D. (2004). 'The Sketch Engine', in *Proceedings of Euralex*, Lorient, France, July 2004, pp. 105-116.
- Klyšinskij E. S., Lukaševič N. Y. (2018). 'Coprus of syntactic co-occurrences: a delayed promise', in *Artificial Intelligence and Natural Language*, pp. 121-127.

- Kopotev M. V., Mustajoki A. (2003). 'Principy sozdanija Chel'sinskogo annotirovannogo korpusa russkich tekstov (CHANKO) v seti Internet', in *Naučno-techničeskaja informacija, vol. 2 Informativnye processy i sistemy*, n. 6 pp. 33-37.
- Kopotev M., Escoter L., Kormačeva D., Pierce M., Pivovarova L., Yangarber R. (2015). 'CoCoCo: Online extraction of Russian multiword expressions', in *5th Workshop on Balto-Slavic Natural Language Processing*, Hissar, Bulgaria, pp. 43-45.
- Kotov A., Budjanskaja E. (2012). 'The Russian emotional corpus: communication in natural emotional situations', in *Computational Linguistics and Intellectual Technologies. Proceedings of the international conference Dialogue*. Moscow 11:18 (2012), pp. 296-306.
- Kutuzov A., Kopotev M., Sviridenko T., Ivanova L. (2016). 'Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints', in *arXiv: Computation and Language*, pp. 3-10.
- Kutuzov A., Kunilovskaja M. (2017). 'Size vs. structure in training corpora for word embedding models: Araneum Russicum Maximum and Russian National Corpus', in *AIST 2017: Analysis of Images, Social Networks and Texts*, pp. 47-58.
- Leech G. (1992). 'Corpora and theories of linguistic performance', in *SVARTVIK*, pp. 105-122.
- Leech G. (2007). *New resources or just better ones? The holy grail of representativeness*, in *Corpus Linguistics and the Web*, a cura di M. Hundt, N. Nesselhauf, C. Biewer, Amsterdam: Rodopi, pp. 133-149.
- 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.
- Lüdeling A., Evert S., Baroni M. (2007). Using Web data for linguistic purposes, in *Corpus Linguistics and the Web*, a cura di M. Hundt, N. Nesselhauf, C. Biewer, Amsterdam: Rodopi, pp. 7-24.
- 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"*, Moscow, pp. 427-40.
- Manning C. D., Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge (Massachusetts)-London (England): The MIT Press.
- McEnery T. (2003). 'Corpus Linguistics', in *MITKOV*, pp. 448-63.

- McEnery T., Gabrielatos C. (2006). ‘English Corpus Linguistics’, in *AARTS -MCMAHON*, pp. 33-71.
- McEnery T., Xiao R., Tono Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.
- Mitrenina O. (2014). ‘The corpora of Old and Middle Russian texts as an advanced tool for exploring an extinguished language’, in *Scrinium*, vol. 10, pg. 455-61.
- Mohammadi M., Ghasem A. (2010). ‘Building parallel bilingual corpora from Wikipedia’, in *Proceedings of the 2012 Second International Conference on Computer Engineering and Applications*, pp. 264-8.
- Morley A. (2006). ‘WebCorp: a Tool for Online Linguistic Information Retrieval and Analysis’, in *The Changing Face of Corpus Linguistics*, a cura di B. Morley, Amsterdam: Rodopi, pp. 283-96.
- Nacional'nyj Korpus Russkogo Jazyka http://www.ruscorpora.ru/new/corpora-intro.html
- Noseda V. (2017). *Corpora paralleli e linguistica contrastiva: ampliamento e applicazioni del corpus italiano russo nel Nacional’nyj Korpus Russkogo Jazyka*, A.A. 2015/16, Università Cattolica del Sacro Cuore, Milano.
- Noseda V. (2018). ‘La corpus revolution russa e il corpus parallelo italiano-russo’, in *L’analisi linguistica e letteraria XXVI (2018)*, Università Cattolica del Sacro Cuore, Milano, pp. 115-32.
- Oja D., Parsons J. J. (2012). *Computer Concepts: Illustrated*, Stanford, CT: Cengage Learning.
- Piperski A. C. (2013). ‘General’nyj Internet-korpus russkogo jazyka i ponjatie representativnosti v korpusnoj lingvistike’, in *Sovremennye problemy nauki i obrazovanija*, vol. 5.
- Renouf A, Kehoe A., Banerjee J. (2007). ‘WebCorp: an integrated system for web text search’, in *Corpus Linguistics and the Web*, a cura di M. Hundt, N. Nesselhauf, C. Biewer, Amsterdam: Rodopi, pp. 46-67.
- Rosenbach A. (2007). ‘Exploring constructions on the Web: a case study’, in *Corpus Linguistics and the Web*, a cura di M. Hundt, N. Nesselhauf, C. Biewer, Amsterdam: Rodopi, pp. 167-90.
- Rubtsova Y., Zagorulko Y. A. (2014). ‘An approach to construction and analysis of a corpus of short Russian texts intended to train a sentiment classifier’, in *Computer Science 37 (2014)*, pg. 107-16.
- Šagalova I., Ekaterina N. (2017). *Slovar’ novejšich inostrannych slov’*, Moskva: AST-Press.

- Sampson G., McCarthy D. (2004). *Corpus Linguistics. Readings in a Widening Discipline*. London-New York: Continuum.
- Šarov S. (2003). 'Methods and tools for development of the Russian Reference Corpus', in *Proceedings of Corpus Linguistics Conference*, April, 2003, Lancaster, UK.
- Šarov S. (2006). 'Creating general-purpose corpora using automated search engine queries', in *Wacky! Working Papers on the Web as Corpus*, a cura di Baroni M., Bernardini S., Bologna: Gedit, pp. 63-98.
- Selegej D. V., Šavrina T. O., Selegej V. P., Šarov S. A. (2016). 'Avtomatičeskaja morforazmetka korpusov russkojazyčnych social'nych media: obučenie i ocenka kačestva', in *Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ežegodnoj meždunarodnoj konferencii «Dialog»* 15, pp. 589–604.
- Šerstinova T. (2009). 'The structure of the ORD speech corpus of Russian everyday communication', in *Text, Speech and Dialogue, 12th International Conference, TSD 2009*, Pilsen, Czech Republic.
- Sinclair J. (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing and the development of the Collins COBUILD English Language Dictionary*. London-Glasgow: Collins ELT.
- Sinclair J. (1991). *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.
- Sinclair J. (2003). *Reading Concordances: An Introduction*. London-New York: Pearson/Longman.
- Sinclair J. (2005). 'Corpus and text. Basic principles', in *Developing Linguistic Corpora: a Guide to Good Practice*, a cura di Wynne M., Oxford: Oxbow Books, pp. 1-16.
- Sinclaire J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Skrelin P., Volskaja N., Kočarov D., Evgrafova K., Glotova O., Evdokimova V. (2010). 'A fully annotated corpus of Russian speech', in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 17-23 May 2010, Valletta, Malta.
- T. McEnery, A. Hardie (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- T. McEnery, A. Wilson (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Tognini-Bonelli E. (2001), 'Corpus Linguistics at Work', in *Studies in corpus linguistics*, vol. 6, Amsterdam: John Benjamins Publishing Company.

- Trofimova G. N. (2011). *Jazykovoj vkus internet-èpochi v Rossii. Funkcionirovanie russkogo jazyka v Internete: konceptual'no-suščnostnye dominanty*, (2° ed.) Moskva: Rossijskij Universitet Družby Narodov.
- Von Waldenfels R., Daniel M., Dobrušina N. (2014). 'Why Standard Orthography? Building the Ustya River Basin Corpus, an Online Corpus of a Russian Dialect', in *Komp'juternaja lingvistika i intellektual'nye technologii: po materialam ežegodnoj meždunarodnoj konferenzii "Dialog"*, pp. 720-28.
- Zacharov V. (2013). 'Corpora of the Russian language', in *16th International Conference Text, Speech and Dialogue 2013*, Pilsen, Repubblica Ceca, pp. 1-13.
- Zanettin F. (2012). *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*, Manchester: St. Jerome.
- Zanni S. (2007). 'Corpora elettronici e copyright. Lo status legale della questione', in *Corpora e Linguistica in rete*, a cura di Barbera M., Corino E., Onesti C. Perugia: Guerra edizioni, p. 119-26.

РЕЗЮМЕ

Темой данной дипломной работы является сравнение традиционных корпусов с интернет-корпусами русского языка.

Глава 1

Поскольку наша работа касается корпусов, она начинается с вводной главы о корпусной лингвистике. Корпусная лингвистика — это метод исследования языка основан на наблюдении реальных данных. Такие данные являются текстами в электронном формате, составляющими корпуса. На самом деле, корпусная лингвистика не область лингвистики, как некоторые авторы говорят, но, согласно авторам McEnery и Wilson, это - метод, применимый к любой области лингвистики.

Английский лингвист Geoffry Leech утверждает, что корпусная лингвистика является научной дисциплиной. По его словам, она позволяет исследовать лингвистические проблемы с научным методом, применяя три научных принципа: достоверность, повторяемость, извращение (falsificabilità). С лингвистическими данными, полученными от корпусов, можно привести и качественные и количественные анализы.

Мы сделали также экскурс в историю дисциплины. Очень кратко, она родилась в 19-ом веке и, несмотря на критику лингвиста Chomsky в 1950-х годах, развивалась до наших дней. В течение этих годов она была применена к разным областям, таким как преподавание иностранных языков, как к синтаксис, так и к семантике.

Следующая часть главы касается корпусов, те есть главных инструментов корпусной лингвистики. Мы начали с их определения. Национальный Корпус Русского Языка описывает лингвистический корпус как «информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме». Разметка и тонизация являются главными техническими характеристиками данных текстов.

Кроме того, объясняются все характеристики корпусов: подлинность, электронная форма, большой размер, представительность, уравновешивание, отбор, несовершенство, разметка, целевой порядок, сравнимость.

Мы объяснили также разные типы корпусов и, для несколько из них, проводили примеры русских корпусов. Существуют корпуса письменного языка (НКРЯ) и устной речи (Корпус Устной Речи); мультимедийные корпуса (Мультимедийный Корпус МУРКО); общие корпуса (НКРЯ) и специализированные корпуса (Регенсбургский Диахронический Корпус Русского Языка); одноязычные корпуса (НКРЯ) и многоязычные корпуса, среди которых существуют параллельные корпуса (параллельный корпус русско-итальянский Национального Корпуса Русского Языка) и сопоставимые корпуса; синхронические корпуса и диахронические корпуса (Регенсбургский Диахронический Корпус Русского Языка); мониторинг корпуса; корпуса учащихся (Корпус Русских Учебных Текстов).

Более того, мы задержались на нескольких технических аспектах корпусов, таких как токенизациа, разметка, лемматизация, конкорданции и коллокации.

В конце, выделяются разные сферы применения корпусной лингвистики. Это не только лингвистические дисциплины, как сравнительно-историческое языкознание, анализ речи, семантика, социолингвистика, грамматика и т.д., но и другие дисциплины. Корпусная лингвистика очень полезна в применении к гуманитарным дисциплинам и общественным наукам, занимающимся изучением текстов, как на пример литература, религия, история, психология.

Глава 2

Во второй главе речь идёт о русских корпусах. Говорится не только о традиционных корпусах как НКРЯ, но и о новых, экспериментальных интернет-корпусах.

Мы начали с НКРЯ, его истории, характеристик и под-корпусов. НКРЯ является главным корпусом русского языка. Он был создан в 2004 года при Российской академии наук в Москве, и его создание поддержал поисковая система Яндекс. Корпус содержит 700 миллионов слов и его тексты были написаны между 11-ым и 21-ым веками. Тексты принадлежат жанрам повествовательным, журналистским, академическим, экономическим, научным и т.д. Корпус имеет разметку разных типов: мета-текстуальную, морфологическую, семантическую, синтактическую, акцентуальную и поэтическую. В корпусе включены 12 под-корпусы, каждый из которых содержит разные типы текстов: основой корпус, синтактический корпус, газетный корпус, параллельные корпуса, корпус диалектных текстов, корпус поэтических текстов, обучающий корпус, корпус устной речи, акцентологический корпус, мультимедийный корпус, исторический корпус и

мультимедийный параллельный корпус Мультпарк. Среди параллельных корпусов, мы обратили внимание особенно на русско-итальянский корпус. До 2015 года, этот подкорпус был очень маленьком, но в этом году его увеличили. Сегодня он содержит 4,5 миллионов слов.

Ещё другой корпус стоит отметить, то есть Генеральный Интернет-Корпус Русского Языка. Это «мега» интернет-корпус, содержащий 20 миллиардов слов. Он был открыт в 2012 году, но сегодня – ещё в разработке и, чтобы его употреблять, надо потребовать от создателей данных доступа. Корпус основан на материале, взятом из интернета и его цель является представить языка веба. Данный язык – разговорный, непосредственный и владеет свойствами и разговорной и письменной речи. ГИКРЯ отличается от других интернет-корпусов тем, что имеет не только мета-текстуальную, но и морфологическую разметку. Поэтому, данный корпус является одним из самых интересных лингвистических инструментов исследования русского языка.

Кроме этих двух, можно упомянуть много других корпусов разных типов:

Интернета корпуса. Araneum Russicum (Aranea), ruTenTen (Sketch Engine), ruWaC (WaCky)

Корпуса учащихся. *Child Language Data Exchange System* (*CHILDES*), Корпус Несовершенного Перевода, Корпус Русских Учебных Текстов (КРУТ), Русский Учебный Корпус

Исторические и диалектные корпуса. Корпус Великих Четьих-Миней, параллельный корпус переводов Слова о Полку Игореве, Манускрипть, Регенсбургский Диахронический Корпус, Санкт-Петербургский корпус агиографических текстов (СКАТ), *Ustja River Basin Corpus*

Корпуса разговорной речи. *Corpus of Russian Professionally Read Speech* (*CORPRES*), *CoRuSS*, Один Речевой День, Учебный Мультимодальный Корпус (УМКО)

Корпуса невербального общения. *Russian Emotional Corpus*,

Другие Корпуса. Хельсинский Аннотированный Корпус (ХАНКО), корпус библиотеки Мошкова, *Integrum World Wide*, Компьютерный корпус текстов русских газет конца 20-ого века, Корпус Синтактических Комбинации (КоСиКо), *Leeds University Corpora*, *OpenCorpora*, *OPUS*, *RuTweetCorp*

Другие *corpus-based* ресурсы. Частотный русский словарь, Частотный словарь современного русского языка, *CoCoCo*, частотный словарь корпуса *ruWaC*, Лексико-

грамматический частотный словарь, *FrameBank,* Частотная русская грамматика, Статистический словарь языка русской газеты.

Глава 3

В третей главе, мы ввели понятие «веб как корпус», согласно которому, веб стал источником лингвистических данных как альтернатива традиционным корпусам. Эта идея впервые появилась в новом тысячелетии в статье вычислительного лингвиста Adam Kilgarriff. Веб считается крупной спонтанной коллекцией действительных текстов в электронном формате, в которую всякий может входить просто и бесплатно. Достоинствами данного средство с лингвистической точкой зрения являются большие разметки, разнообразные жанры и стили, новейший материал, непосредственность и, наконец, тот факт, что в нём присутствуют все языка мира.

С другой стороны, некоторые лингвисты утверждают, что веб невозможно считать корпусом, потому что его не создали с лингвистической целью. На самом деле, веб был создан для того, чтобы найти информацию, а не лингвистические элементы; его размеры не определены а постоянно развиваются; свой материал не организованный, а беспорядочный; его не хватает лемматизацию и морфологическую разметку; в нём присутствуют ошибки потому что свой материал никто не контролирует.

Потом, мы ещё раз рассмотрели главные характеристики корпусной лингвистики, но с точкой зрения веба:

Действительность тексты. Язык текстов, содержащих в интернете – реальный, спонтанный. Поэтому, веб так действительный как традиционные корпуса

Репрезентативность. С одной стороны, веб - репрезентативный какого-либо языка, потому что он включает разнообразные жанры и стили, но с другой, поскольку его материал не уравновешенный, не могли бы считать веб репрезентативным.

Размеры. Тот факт, что веб - такой крупный источник, является преимуществом. На самом деле, его употреблением можно получить больше результатов, чем традиционными корпусами. Но корпусная лингвистика является научной дисциплиной и поэтому надо основаться на количественные данные. Однако, размеры интернета – неизвестные.

Состав/содержание. Материал веба является очень богатым, в нём присутствуют тексты на каком-либо языке, диалекте, стиле, жанре. Тем не менее, тот материал – анархический и беспорядочный.

Достоверность результаты. По причине нескольких свойств и функций поисковых систем, результаты поиска могут быть ненадёжные.

Разметка. Веб-страницы не хватают и мета-текстуальную, и морфологическую разметку. Некоторые мета-текстуальные данные можно извлечь «вручную» от нескольких типологий текстов, такие как статья, блоги, посты социальных сетей.

Следующая часть главы - весьма интересная. Здесь речь идёт об использовании веба как корпус. Согласно авторам Baroni и Bernardini, существуют четыре подхода, которые мы здесь цитируем на английском языке:

1. **Веб как *corpus surrogate***. Это - использование веба с лингвистической целью посредством поисковых систем (Google, Яндекс) или программных обеспечений, нацеленных на лингвистическое исследование (*WebCorp*).

   Поисковые системы: позволяют искать слово или группу слов. Результаты будут нам показаны в их контексте, вместе с указанием их частоты. Благодаря ряду параметров и операторов поиска, можно вести и сложные грамматические поиски.

   Лингвистические программные обеспечения: то, что они делают это, во-первых, помогать формулировать в поисковой системе запрос, эффективный с лингвистической точки зрения, и, во-вторых, показать результаты поисковой системы в формате, более подходящем лингвистическому исследованию, как например формат «concordancer».

   Поисковые системы и лингвистические программные обеспечения являются хорошими средствами особенно для поиска предложений, коллокаций и кандидатов перевода. Однако, их главный недостаток— это то, что, поскольку поисковые системы основаны на данных и интересы пользователей, каждый поиск даёт разные результаты.

2. **Веб как *corpus shop***. Возможно создать, посредством программ как *BootCat*, специализированные сопоставимые корпуса, основанные на материале, извлёкшем от веба. Данные корпуса являются маленькими и одноразовыми, и пользователь может создать их каждый раз в зависимости от требования. Они очень полезные особенно для переводчиков.

3. **Веб как *corpus proper***. В соответствии с этим подходом, веб является корпусом, представляющим язык интернета. Следовательно, веб станет специализированным корпусом.
4. ***Mega-corpus mini-web***. *Mega-corpus mini-web* - это новое средство лингвистического исследования, объединяющее характеристики веба (большой, новейший и т.д.) и корпусов (разметка, определённые размеры и т.д.). От этого понятия родились «мега» интернет-корпуса. Они являются *corpus manager,* содержащими корпуса разных языков. Некоторые примеры данных новых корпусов являются: *Leeds Collection of Internet Corpora, Sketch Engine, Aranea, WaCky*.

В данной работе, мы приняли во внимание особенно подходы 1, 2 и 4.

Исходя из этих подходов, можно различить два понятия: *web for corpus* (1 и 3) и *web as corpus* (2 и 4). Согласно первому, веб является корпусом, а согласно второму, веб является источником, от которого извлекать материал чтобы создать корпус.

Существуют другие веб-сервисы, полезные для лингвистического исследования:

Google Books. Это сбор книг в электронном формате. Среди прочего, *Google Books* позволяет и лингвистические поиски благодаря нескольким параметрам поиска, таким как язык, дата, жанр, автор и т.д. Разметки, однако, не хватает.

Google Ngram Viewer. Предоставляет информацию о частоте употребления слов или предложений в данном периоде.

Google Scholar. Это сбор академических текстов и работает точно, как *Google Books*.

Wikipedia. Из-за того, что *Wikipedia* является многоязычной энциклопедией, её очень часто считают параллельным или сопоставимым корпусом энциклопедических текстов.

В конце данной главы мы пришли к заключению, что определение веба как корпус – относительное и мы не можем утвердить, является ли веб хорошим или плохим инструментом. Ещё, действительность интернет-корпусов зависит от типологии исследования и от знакомства пользователей с веба и его функций.

Глава 4

В заключительной главе сравниваются традиционные и интернет-корпуса русского языка на практике. Мы использовали данные типологии корпусов чтобы провести исследования по разным лингвистическим аспектам: это грамматика, коллокации, англицизмы, сленги, технические термины, подобные слова. Нашей целью является установить какую типологию корпуса лучше подойдёт для исследования этих аспектов. Мы сравнили и разные интернет-корпуса друг с другом.

Как традиционный русский корпус мы употребляли НКРЯ, а как интернет-корпуса – *ruTenTen* (*Sketch Engine*), *Araneum Russicum Maximum* (*Aranea*) и ГИКРЯ.

Грамматика

Собирательные существительные

Мы выбрали собирательное существительное «картошка», потому что итальянские люди очень часто его ошибочно употребляют во множественном числе. Кроме того, они думают, что в словосочетании «купить картошки», существительное «картошки» – в винительном падеже. Напротив, это партитив, и поэтому окончание «-и» указывает родительный падеж. С нашим исследованием мы хотим доказать, что:

1. Существительное «картошка» возможно употреблять только в единственном числе
2. Когда существительное «картошка» оканчивается на «-и» после глагола, оно не в винительном, а в родительном падеже.

Наш поиск состоит из двух фаз:

1. Наблюдать как пользуется существительное «картошка». Этот поиск можно было провести с употреблением и традиционных и интернет-корпусов.
2. Проверять, существуют ли случаи, когда существительное «картошки» употребляется в именительном или винительном падеже. Этот поиск можно было провести только с традиционным корпусом НКРЯ, потому что интернет-корпуса не позволяют сложные грамматические поиски. Также ГИКРЯ, теоретически, позволяет сложные грамматические поиски, но, практически, мы заметили, что они не работают.

Благодаря этому поиску, мы доказали, что

1. Существительное «картошка» употребляется только в единственном числе

2. Случаи, когда существительное «картошки» употребляется в именительном или винительном падеже, не существуют.

Партитив

Мы хотели доказать, что родительный партитивный, оканчивающий на «-у» в мужских существительных, сегодня употребляется очень редко по сравнению с родительным партитивным, оканчивающим на «-а». В качестве примера мы выбрали предложение «я добавил немного сахару». Мы использовали корпуса чтобы сравнить частоту употребления выражений «немного сахара» и «немного сахару». И традиционные корпуса, и интернет-корпуса, и поисковые системы показали, что первое встречается чаще, чем второе.

	немного сахара	немного сахару
НКРЯ	32	13
Sketch Engine	2.994	147
Aranea	4.811	117
ГИКРЯ	1.985	80
Google	754.000	20.300
Яндекс	3.000	182

Однако, мы заметили, что редкость выражения «немного сахару» менее явная в НКРЯ, чем в интернет-корпусах. Это потому, что в НКРЯ присутствует значительное количество текстов из прошлого и, наоборот, язык интернет-корпусов - современный и новейший.

Биаспектив

Некоторые биаспективные глаголы, со временем, входят в аспектуальную систему после того, как они приобрели префикс. Два примера таких глаголов - «поисследовать» и «продиагностировать». С помощью корпусов, мы проанализировали их развитие с течением времени и наблюдали их распространение по сравнению с формой без префикса. Результаты видны в следующих таблицах:

	Исследовать (1700-2017)	Поисследовать (1884-1953)	Процент префиксации
НКРЯ	9050	8	0,08 %
Sketch Engine	550.920	427	0,07 %
ГИКРЯ	45.743	798	1,74 %

	Диагностировать (1877-2017)	Продиагностировать (1997-2013)	Процент префиксации
НКРЯ	266	10	3,75 %
Sketch Engine	69.793	2.345	3,35 %
ГИКРЯ	5.780	558	9,65 %

На основе этих данных, мы пришли к заключению, что:

- К глаголу «диагностировать» применяется префикс в 5,58 % случаев, а к глаголу «исследовать» в 0,6 %. Поэтому, глагол «исследовать» это - почти полностью биаспективный. Напротив, глагол «диагностировать» вышел в аспектуальной системе в большей степени. Более того, префикс применяется чаще в языке веба, в частности в блогах и социальных сетях.
- Глагол «поисследовать» меньше распространился несмотря на то, что «исследовать» это - старший глагол чем «диагностировать» и он приобрёл префикс раньше.

Чтобы получить такие данные, НКРЯ был фундаментальным благодаря своей мета-текстуальной разметке. С другой стороны, интернет-корпуса нам показали распространение данных глаголов на современном русском языке и их интеграцию в аспектуальной системе.

Коллокации

Мы выбрали коллокацию «носить имя» чтобы провести, с помощью корпусами, два поиска:

1. Наблюдать контексты употребления данной коллокации
2. Искать другие коллокации с существительным «имя»

Первый поиск мы смогли провести и с НКРЯ, и с интернет-корпусами, и с поисковыми системами. Однако, мы заметили, что результаты интернет-корпусов - крайне многочисленные и нам было невозможно все просмотреть. Поэтому, результаты, полученные от НКРЯ – достаточные для нашего исследования. Мы ещё увидели, что среди результатов поисковых систем присутствуют несущественные элементы.

Второй поиск был успешно проведён с интернет-корпусом *Sketch Engine* благодаря его функции *word sketch*. Нам было достаточно искать существительное «имя» и корпус нам показал все его коллокации. Например:

Англицизмы

Мы использовали корпуса чтобы проанализировать употребление глагола «лайкать-лайкнуть». В таблице указан итог появлений глагола в разных корпусах:

	Лайкать-лайкнуть
НКРЯ	11
Sketch Engine	363
Aranea	4.596
ГИКРЯ	16.291
Google	4,7 млн

Результаты, полученные от НКРЯ недостаточные чтобы узнать общее поведение глагола. Данный англицизм является довольно недавним, а корпус – недостаточно новейшим. Для такого поиска, мы советуем употребление интернет-корпуса. Таким образом будет ясно, что данный глагол, на самом деле, довольно популярный на русском языке.

Существует, на русском, и антоним глагола «лайкать-лайкнуть», то есть «анлайкать-анлайкнуть». Это - более редкий и более недавний явление, как можно увидеть от таблицы:

	Анлайкать-анлайкнуть
НКРЯ	0
Sketch Engine	2
Aranea	1
ГИКРЯ	6
Google	410

Употребление интернет-корпусов, корме НКРЯ, нам позволил:

1. установить тот факт, что глагол «лайкать-лайкнуть» очень распространённый на русском языке, в отличие от того, что НКРЯ указал.
2. подтвердить тот факт, что глагол «анлайкать-анлайкнуть» в действительности очень редкий.

Ещё, мы искали через корпус англицизмы «оффтоп» и «оффтопик», чтобы проверить какой из них чаще используется в русском языке.

	Оффтоп	Оффторик
НКРЯ	40	14
Aranea	8.614	3.205
Sketch Engine	11.824	2.412
ГИКРЯ	19.728	4.201
Google	5,1 млн	2,6 млн

Мы установили, что «оффтоп» более распространённый.

Сленги

Мы выбрали два жаргонных слова, обозначающих «деньги», те есть «бабло» и «бабки». В таблице приведены результаты:

	Деньги	Бабло	Процент употребления сленга
НКРЯ	74.892	83	0,11 %
Sketch Engine	6 milioni	41.223	0,68 %
Aranea	7,5 milioni	43.002	0,57 %
ГИКРЯ	4,9 milioni	82.715	1,68 %
Google	463 milioni	10,8 milioni	2,33 %
Яндекс	29 milioni	9 milioni	31 %

	Деньги	Бабки	Процент употребления сленга
НКРЯ	74.892	2.881	3,8 %
Sketch Engine	6 milioni	73.327	1,22 %
Aranea	7,5 milioni	65.792	0,87 %
ГИКРЯ	4,9 milioni	98.806	2,01 %
Google	463 milioni	12,3 milioni	2,65 %
Яндекс	29 milioni	5 milioni	17,2 %

Мы проверяли, что:

- «бабки» употребляется чаще, чем «бабло».
- «бабло» чаще употребляется в разговорной речи, а «бабки» в литературном языке.
- Слово «деньги» заменяется с сленгом «бабло» в 1,07 % случаев, а с сленгом «бабки» в 2,11 %.

Подобные слова

«День» и «сутки» являются подобными словами так как оба они выражают значение «24 часа». Если смотреть в словарях, неясно, когда лучше употреблять одно или другое слово. Мы употребляли корпуса чтобы пояснить, когда используется слово «сутки» и установили,

что, для такого поиска, любой корпус подходит. Однако, благодаря ГИКРЯ, мы пришли к заключению, что слово «сутки» принадлежит более официальному языку.

Технические термины

Мы искали медицинский термин «биопсия» чтобы больше узнать о его употреблении и коллокации. От интернет-корпусов мы получили больше результатов, чем от НКРЯ. В частности, лучший интернет-корпус чтобы провести такого поиска это – *Sketch Engine,* благодаря его функции *word sketch.*

Заключение

Традиционный корпус НКРЯ является лучшим средством, чтобы провести грамматические поиски благодаря его морфо-синтактической разметке. Напротив, интернет-корпуса не подходящие для грамматических поисков. Ещё, НКРЯ лучше, чем интернет-корпуса для хронологических исследований лингвистических феноменов благодаря своей мета-текстуальной разметке.

Интернет-корпуса – это ценные инструменты благодаря их богатому и новейшему языковому материалу. Они более подходящие для анализа таких лингвистических феноменов как англицизмы, технические и жаргонные термины. В НКРЯ, эти феномены присутствуют в маленьком количестве, или отсутствуют если являются слишком недавними. Более того, интернет-корпуса являются очень полезными для исследования распространения таких грамматических феноменов, как партитивный родительный и интеграция в аспектуальной системе биаспективных глаголов.

Среди интернет-корпусов, *Sketch Engine* отличается своей функцией *word sketch* для исследования поведения и коллокаций слов. Ещё, ГИКРЯ отличается соединением богатства материала интернет-корпусов и сложных грамматических поисков традиционных корпусов. Однако, его грамматический поиск нуждается в усовершенствовании.