Домашнее задание № 2.
Корпусная лингвистика, статистика, языковые модели

Отчет

Задание В:

Изучить возможности системы SketchEngine (https://www.sketchengine.co.uk), предварительно зарегистрировавшись в ней, и на основе входящего в нее корпуса RuTenTen11 исследовать особенности употребления слова и его смысла, выбрав одно из слов, указанных в варианте А (или взять свое). Для этого: 1) Найти и сравнить значения и толкования слова в словарях Даля, Ушакова, Ожегова, Кузнецова (см. http://dic.academic.ru/) 2) Исследовать употребление этого слова в корпусе RuTenTen11 3) Проанализировать график употребления слова (https://books.google.com/ngrams). 3) Посмотреть использование слова на yandex.ru (взяв подходящий раздел, например, новости, культура, наука) Отчет (2-4 страницы): Характеристика функций SketchEngine; описание сходства-различий в употреблении выбранного слова в разных словарях, а также в реальном употреблении (корпус, Яндекс-тексты); выводы.

1 Введение

В данном отчете приведено исследование употребления слова "**загружать** и его изменения во времени. Изучены словарные значения из словарей разных периодов, современные примеры употребления с помощью SketchEngine и wordstat.yandex, изменение частоты и смыслов употребления со временем в книгах Google.Books и Rucorpora.ru. Кроме того, приведено подробное описание работы и возможностей всех использованных систем.

2 Значение и толкование из словарей слова “загружать”

- Толковый словарь Даля (1863-1866)[1]
  ЗАГРУЖАТЬ, загрузить что
  1. перегружать, класть груза более, чем должно, более, чем лошадь или судно подымает;
  2. о местности: завалить, загромоздить, занять весь простор. -ся, ·возвр. и страд. по смыслу речи. Загруженье ср., ·окончат. загрузка жен., ·об. действие по гл.
- Толковый словарь Ушакова (1934-1940)
  ЗАГРУЗИТЬ загружу, загрузишь, совер. (к загружать) .
  1. *что*. Наполнить грузом до какого-нибудь предела, нагрузить полностью. Загрузить все подвалы картошкой.
  2. перен., *кого-что*. Заполнить работой, делом до необходимого предела; предоставить работу в нужном количестве (неол.). Загрузить рабочий день служащих. Загрузить преподавателей.
- Толковый словарь Ожегова (1949-1992)
  ЗАГРУЗИТЬ, -ужу, -узишь и -узишь; -уженный и -ужённый (-ён, -ена); совер.
  1. см. грузить.
  2. перен., кого (что). Заполнить работой, дать работу в нужном количестве, занять 2 (в 4 знач.). З. рабочий день. З. преподавателей.

[1] ссылки на источники приведены в гиперссылках в названиях словарей

- Большой толковый словарь русского языка. Гл. ред. С. А. Кузнецов. (1998, ред. 2014г) ЗАГРУЗИТЬ -гружу, -грузишь и -грузишь; загруженный; -жен, -а, -о и загружённый; -жён, -жена, -жено; св.
  1. *что (чем).* Наполнить грузом, запасами, товарами. З. машину кирпичом. З. судно по самый борт. З. подвал картошкой. З. сумки. З. желудок (заполнить, набить, обычно не слишком полезным и калорийным).
  2. *кого (что).* Обеспечить работой, занять работой, делами в нужном количестве. З. станок. З. завод. З. учителей. Продавцы загружены лишь на пятьдесят процентов. // Заполнить работой (время). З. рабочий день служащих. Последний месяц года всегда загружен.
  3. *что чем.* Заложить руду, топливо и т.п. в печь, домну и т.п. З. в открытую шахту печи руду и древесный уголь.
  4. *Информ.* Скопировать данные из внешнего запоминающего устройства в оперативную память компьютера.

Кроме того, у слова “загрузить” есть и другие современные значения:

- Если вы загружаете какой-то механизм, значит, вы очень активно используете этот механизм, даёте ему очень много работы. (Толковый словарь русского языка Дмитриева. Д. В. Дмитриев. 2003.)
- Если вы говорите, что кто-то загружает голову, мозги какой-то информацией, вы хотите сказать, что человек излишне серьёзно думает, беспокоится о чём-то, о чём можно было бы не думать и не беспокоиться; в разговорной речи. По-моему, ты зря загружаешь себе голову этими вопросами. (Толковый словарь русского языка Дмитриева. Д. В. Дмитриев. 2003.)
- Утомить тяжелым разговором; отяготить посторонними проблемами. /возможно доп. в тв.п./ Батый меня загрузил, аж крыша поехала.
- Загрузить пациента — ввести психотропные препараты (Толковый словарь медицинского сленга)

Можно сделать вывод о том, что изначально слово употреблялось в смысле "заполнить грузом или работой" и близких к ним. С распространением интернета слово не утратило этих значений, но приобрело еще одно — "перемещение данных между носителями информации". Кроме того появилось еще несколько разговорных значений.

3 Характеристика функций SketchEngine. Анализ современного использования слова

SketchEngine — онлайн система, работающая на больших корпусах текстов разных языков и позволяющая определить типичные или наиболее часто употребляемые фразы, а также редкие или устаревшие слова и выражения.

Система работает с более чем 90 языками, содержит около 100 корпусов (до 30 миллиардов слов) текстов на различных языках. Кроме того, есть возможность с помощью специального архитектора создавать собственные корпуса из текстов, загруженных из интернета или с носителя.

SketchEngine объединяет несколько систем с различным функционалом. Основная из них — Word Sketch. Она позволяет по слову или нескольким словам определить основные словосочетания, в которых оно (они) используется. Такие словосочетания анализируются

и выдаются объединенными по группам, чтобы не пришлось просматривать все примеры, встреченные в корпусе. Word Scetch возвращает в удобном сжатом формате всю найденную информацию. Кроме того, в составе SE есть система Word Sketch Difference, позволяющая увидеть сходства и различия использования нескольких лемм. Thesaurus позволяет посмотреть синонимы к запрашиваемому слову. Это как слова со со схожим значением, так и слова, часто употребляемые в том же контексте.

Для анализа использовался корпус ruTenTen11 (Russian Web 2011). Он составлен из нескольких корпусов интернет-текстов (включая русскоязычную википедию, например) и содержит 14,553,856,113 слов. Система позволяет работать с подкорпусами, выбранными по годам, по темам и т.д. Для анализа в этом отчете используется самый полный корпус. В основном он содержит тексты, написанные после 2010, соответственно после приобретения словом "загрузить" новых смыслов. Следовательно, можно считать весь корпус современным в интересующем нас смысле.

По лемме в запросе Word Sketch возвращает *collocations* — примеры использования слова в разных качествах (объект/субъект, например) или разных смыслах, самые частые и самые типичные комбинации для этого слова. Для каждого примера есть возможность просмотреть контекст — *concordance*. Словосочетания можно искать по n-граммам во всем корпусе. Примеры можно отсортировать по частоте, с которой они встречаются, или по 'типичности' такого словосочетания (SketchEngine подсчитывает специальный score, отражающий такую типичность). Помимо этого, SketchEngine предлагает синонимы для выбранного слова и похожие в использовании (в схожих контекстах) слова.

Grammar relation	**Collocate**	**Freq**	**Score**	**Grammar relation**	**Collocate**	**Freq**	**Score**
subject	**10219**	**3.17**		**adv_modifier**	**59037**	**18.29**	
	пользователь	772	2.93		можно	13784	4.6
	браузер	190	4.42		полностью	3115	5.27
	файл	171	1.48	**verb_post_inf**	**46200**	**14.31**	
	процессор	119	2.5		мочь	11715	3.03
	сейв	38	6.67		быть	6946	1.19
object4	**88330**	**27.37**			позволять	5805	4.8
	файл	10872	7.41	**pp_в**	**20688**	**6.41**	
	фотография	3284	5.71		память	1761	4.3
	программа	2324	2.5		машина	772	2.15
post_prep	**64526**	**19.99**			телефон	548	2
	в	27254	0.81		формат	322	2.74
	на	15177	0.98		печь	279	3.95
	с	6763	0.55	**и/или**	**13674**	**4.24**	
	из	4157	1.26		устанавливать	1536	3.94
	до	1712	1.25		выгружать	522	8.84
	через	1402	2.33		просматривать	507	5.26
	при	1113	0.18	**pp_на**	**10188**	**3.16**	
	под	663	0.42		сайт	2006	2.9
	около	219	1.31		компьютер	1385	4.05
	сверх	103	4.41		сервис	207	1.84
	высокий	66	1.85		мощность	173	1.37
	вместо	60	0.69		борт	124	2.92
	свыше	58	2.58		комп	70	4.71
					файлообменник	49	6.85

Таблица 1: Word Sketch by Word Engine (выборка)

Итак, используем SketchEngine для изучения использования слова 'загружать'. Так как

даже в сжатом виде (когда примеры объединяются по смыслу в кластеры) таблица имеет около 200 строк, приведены первые несколько граф, остальные описаны кратко. Автоматически выданные примеры стандартно сортируются по score (типичности), в приведенной таблице результаты отсортированы по частоте. В таком виде таблица кажется более информативной, и разделение на объект и субъект (об этом далее) более точное.

Результаты приведены в таблице 1. В словосочетаниях 'загрузить' используется с **субъектом** действия (кто загружает: пользователь, процессор,..), с объектом **действия** (что загружается: файл, фотография, компьютер). Самые частые предлоги, используемые со словом в современном корпусе перечислены в графе **post_prep**. Видно, что самый частый предлог 'в' может использоваться в нескольких смыслах слова 'загружать', а следующие по частоте 'на', 'с', 'из', в основном используются в смысле перемещения информации. В то же время предлоги с наибольшим score (наиболее типичные): 'сверх', 'свыше' — чаще используется в смысле физического заполнения чего-либо.

В следующих графах таблицы приведены уточнения к слову 'загружать' разных частей речи (как, на сколько сильно, ...), а так же примеры использования с конкретными предлогами или существительными. Например, 'загружать в' часто встречается как с леммами 'память', 'телефон' (современное значение) так и с 'машина', 'печь' (старое значение).

Можно сделать вывод о том, что слово не утратило своих предыдущих значений с появлением новых.

4 График употребления в google.books

В этой части исследования проанализируем график употребления слова по подсчетам в Google.Books. Система подсчитывает относительную частоту употребления слова в книгах по годам. На графике 1 представлена частота употребления слова 'загружать' в книгах с 1880 по 2008 года (данные существуют только до 2008 года).

Рис. 1: Частота употребления слова 'загружать' в книгах с 1880 по 2008 г

Кроме того, Google позволяет посмотреть примеры употребления слова в Google.Books по нескольким периодам, выделяемым автоматически. **В период 1995-2008** в приведенных примерах слово употребляется в следующих словосочетаниях: 'загружать обновления',

'загружать рисунок', 'загружать сайты', 'загружать в фоновом режиме',... Среди книг, приведенных в пример, в основном самоучители по компьютерным программам и использованию интернета и современный англо-русский словарь.

В период 1959-1994 в примерах слово употребляется в таких словосочетаниях: 'вагоны стали загружать комбинированно', 'загружать в камеру для электрофореза бумагу', "Новый шофер отказывается 'загружать' трупом машину", ...В основном в примерах приведены справочники для разных профессий с описанием производственных действий.

В период 1936-1958 встретились следующие примеры употребления: 'Загружать политическую экономию вопросами хозяйственной политики значит загубить её, как науку', 'Загружать в риги или овины только выстоявшийся (дозревший) в поле лён', 'загружать в овин', 'бежной конец невода следует загружать больше всего', 'Загружать мороженой птицей камеры хранения', ...Как можно догадаться из этих примеров, основном в примерах встречаются руководства: по экономике, по растениеводству, рыболовству и т.д.

В самый ранний период, **до 1935г** примеры следующие: 'хорошо ли загружать телеграфъ словами «въ», «ли» и т. п.', 'загружать десятки и сотни чановъ такимъ вьсомъ', 'загружать предисловия цитатами', 'чтобы не загружать и не обременять умовъ', ... Орфография в книгах сохранена, и в примерах много дореформенных книг. В основном это учебники и пособия.

Можно сделать **вывод**, что смысл, в котором употребляется слово, изменился со временем. В целом эти изменения совпадают с соответствующими по эпохам толковыми словарями. В начале XX века слово 'загружать' в основном использовалось в смысле излишка груза для вагонов, нагрузки для механизмов, или, в переносном смысле, текстов или умов. В середине века чаще в употреблении появляется смысл помещения чего-либо в аппарат, камеру, хозяйственную постройку. В начале XXI века, с распространением интернета, чаще слово употребляется в смысле перемещения информации между носителями. Хотя все предыдущие накопленные смыслы слово сохраняет за собой.

Любопытно разграничение периодов, автоматически выделенных системой. Google довольно точно отделил периоды появления новых примеров употребления.

Кроме описанного выше, Google.Books позволяет искать употребление n-грамм. Правда, как и в случае употребления одного слова, он ищет словосочетание буквально, без учета изменения склонений, порядка слов или с некоторым промежутком. (В то время как SketchEngine учитывает все эти изменения при поиске словосочетаний).

Приведенный график показывает, как изменялась частота употребления слова, но не отражает, какой процент употреблений имел какой смысл. Чтобы отследить изменение частоты употребления слова в каком-нибудь смысле, построим такой график для биграммы. Изменение частоты употребления словосочетания, фиксирующего смысл слова, может отражать изменение употребления слова в этом смысле. Например, на графике 2 изображено изменение частоты употребления биграммы 'загружать машину'. Видно, что сейчас такое словосочетание употребляется реже, чем во второй половине XX века, но график сильно зашумлен из-за небольшого количества публикаций с таким словосочетанием, и нельзя сказать, что уменьшение употребления этого словосочетания статистически значимо.

Для поиска биграмм 'загрузить картинку' или 'загрузить обновление' не хватило данных. Можно проследить рост употребления слова 'загрузить' в смысле перемещения информации, построив график для английского слова 'download' в той же системе. Это прямой перевод слова в интересующем нас смысле, при этом других толкований у слова 'download' нет, и можно приближенно считать, что употребление русского слова изолированно от других значений изменялось так же. На графике 3 видно, как слово вошло в обращение с появлением

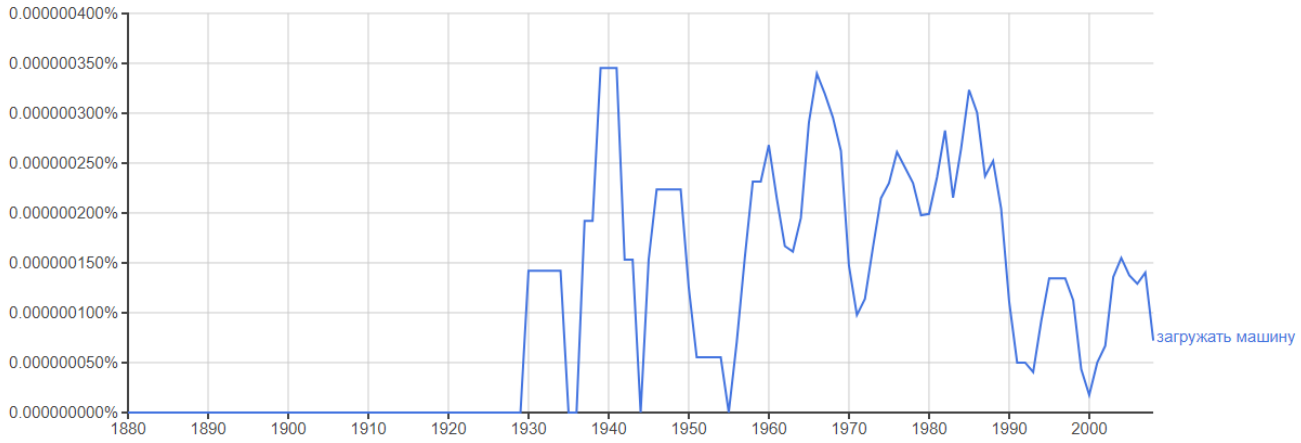

Рис. 2: Частота употребления словосочетания 'загружать машину' в книгах с 1880 по 2008 г (сглаживание по 4 годам)

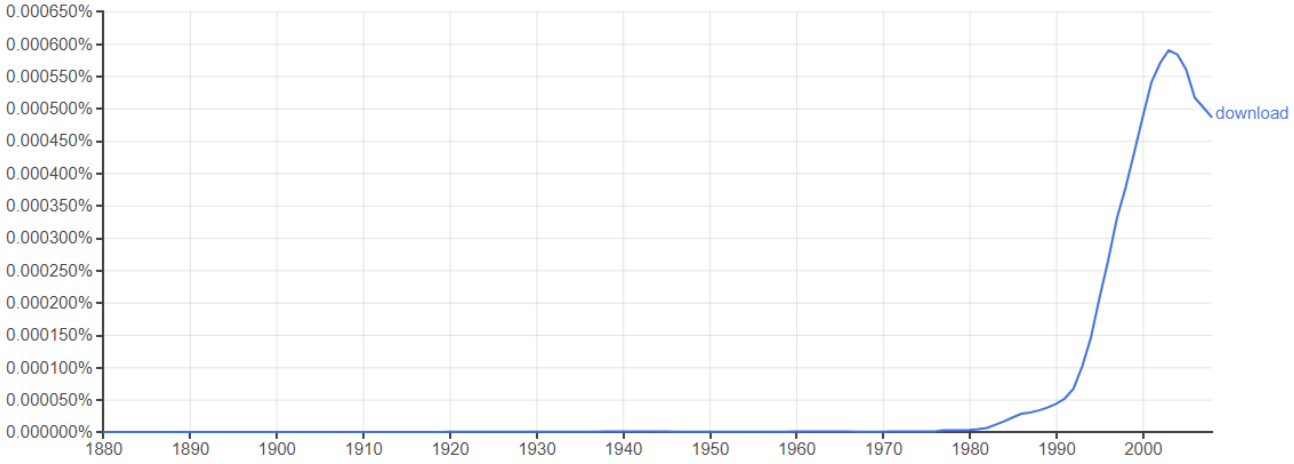

Рис. 3: Частота употребления слова 'download' в книгах с 1880 по 2008 г (сглаживание по 2 годам)

интернета и развивалось с более широким его распространением.

Вывод: Google позволяет проследить изменение частоты и смысла употребления конкретного слова по времени. Но он ищет слово в тексте без изменений, в то время как SketchEngine, например, ищет все образованные словом словоформы. Это затрудняет также поиск по словосочетаниям, потому что часто слова склоняются, употребляются в разном порядке или не подряд, и не 'ловятся' n-граммой.

График употребления слова показывает, что слово употребляется сейчас примерно с той же частотой, что и раньше. Но, по косвенным признакам, мы можем понять, что в старом смысле ('загрузить машину') употреблений становится относительно меньше, а в новом ('download') — относительно больше.

5 График употребления на Rucorpora.ru

График употребления, подобный построенному Google.Books, можно построить с помощью национального корпуса русского языка rucorpora.ru. Такой график приведен на рисунке 4. Можно увидеть, что он сравним с графиком Google, хотя есть и некоторые различия. Нельзя

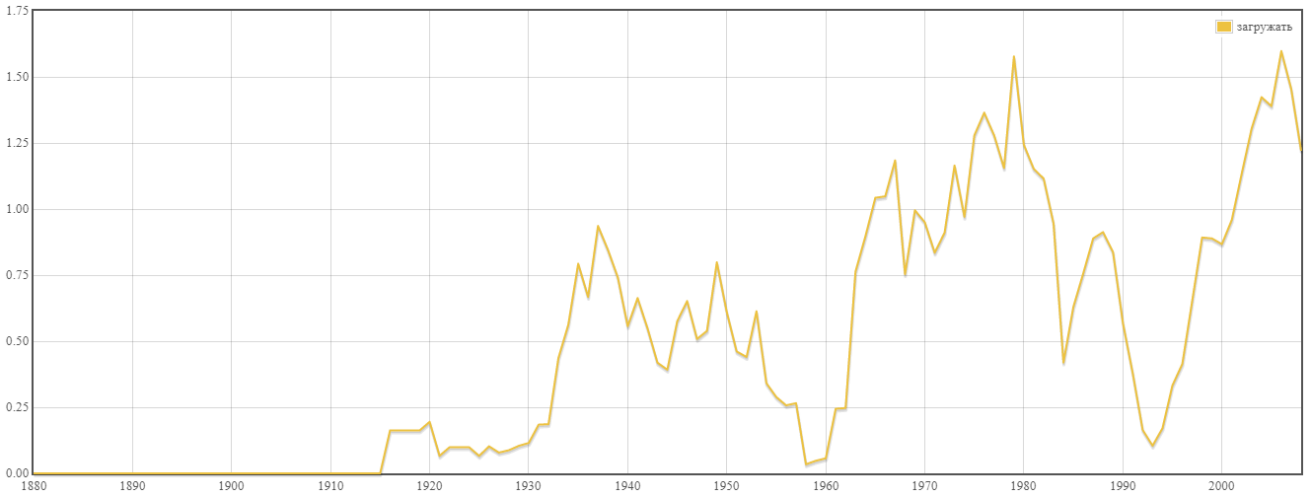

Рис. 4: Распределение по годам (частота на миллион словоформ) слова загружать, источник: rucorpora.ru

сказать, что частота изменения употребления слова как-то значительно изменилась с появлением интернета, а изменения контекстов невозможно увидеть на таком графике. Так же по таким графикам нельзя определить, являются ли различия статистически значимыми.

6 Использование слова на yandex.ru

Яндекс разработал систему wordstat, подсчитывающую частоту, с которой слово встречается в различных запросах. Система разработана для продвижения и оптимизации рекламы пользователей: подсчет количества запросов по определенным словам и словосочетаниям в месяц дает представление о том, сколько просмотров может получить рекламный банер, выдаваемый вместе с ответами на такие запросы.

В таблице 2 представлена частота соответствующих запросов в месяц. Как видно, самое частое употребление этого слова в запросах связано с новым смыслом слова. Запросов связанных со стандартными прошлыми вариантами толкования слова нет в первых 30, показываемых в этой статистике.

Кроме того, Яндекс показывает динамику соответствующих запросов в течении года. По ней нельзя заметить никакой сезонности запросов, содержащих слово 'загружать' в какой-либо форме. Действительно, ни одно из обсуждаемых употреблений слова не связано с определенным периодом года.

Далее рассмотрим первые ссылки, которые Яндекс.Поиск выдает по запросу 'загружать' в разных категориях. Так как в запросе нет уточнения, скорее всего алгоритм предложит какие-то типичные ссылки в самом начале.

В категории поиска по всем новостям, среди первых ссылок новость о пробках ('самую загруженную магистраль Новосибирска'), об Аэрофлоте ('компания загрузила на рейс питание'), о загруженности вокзалов ('Самым загруженным вокзалом Украины')? о музыке('музыканты смогут загружать произведения'). Можно заметить, что в первых ссылках слово в основном используется в стандартном значении, но встречается и в новом значении.

Статистика по словам	Показов в месяц
загрузить	911 017
загрузить фото	84 057
загрузить видео	67 306
загрузить файл	57 023
загрузить бесплатно	54 227
загрузить яндекс	37 549
+не удалось загрузить	33 242
загрузить игры	31 916
загрузить картинку	31 903
+как загрузить компьютер	30 302
загрузить инстаграм	27 186
загрузить музыку	26 761
загрузить через	26 544
загрузить +на телефон	26 090
загрузить +на диск	24 485

Таблица 2: Частота запросов со словом 'загрузить'

7 Краткие выводы

Проведено исследование значений слова 'загружать' и изменение его употребления во времени. Изначально (в конце 19 века) слово имело смысл и употребление слова 'заполнять' (напр. 'загружать телегу', 'загружать работой'). С появлением различных справочников и технологий широкое распространение получил смысл 'поместить что-то во что-то' (напр. 'загружать овес в овин', 'загрузить материал в печь'). С появлением интернета появляется новое значение для слова загружать — 'перемещать информацию' ('загрузить картинку из интернета', 'загрузить компьютер', 'загрузить фото на сайт'). Это употребление слова становится более частотным по сравнению с другими значениями. Кроме того, в современной разговорной речи используются новые словосочетания ('загрузить кого-то своими проблемами'). При появлении новых значений все предыдущие сохранились, и сейчас используются тоже. Частота употреблений слова со временем изменялась за период, который мы можем проследить, но изменения похожи на случайные и не связаны с изменением смысла слова. По имеющимся данным об использовании слова за прошедшие года (около 100 лет) нельзя сказать, что появление новых значений увеличило частоту употребления слова в текстах.