РЕСУРС: Корпус датского языка KorpusDK

Выполнили: Евтодиева Анна (aniatta1999@gmail.com), Дарья Горнштейн (dvgornshteyn@edu.hse.ru), Туркина Евгения (jenkaturkina@yandex.ru)

Введение

KorpusDK содержит набор электронных текстов, полученных из разных источников (общенациональные сми, газеты и журналы, издательства, компании, школы, веб страницы, частные лица).
Тексты, собранные в течение 12 лет (с 1990 по 2002), были специально обработаны для лингвистических исследований. Корпус оснащен многозадачными инструментами поиска, поэтому существует возможность очень точно искать примеры конкретных терминов.

1. Дизайн

У Корпуса приятная зелено-коричневая цветовая гамма, неотвлекающая и не режущая глаз.

Руководство пользователя расписано удобно - выделены ключевые слова, на разделы, о которых рассказывается в руководстве, можно сразу перейти по ссылке - достаточно нажать на название интересующего раздела в тексте. Такое устройство сайта позволяет быстрее понять, где что расположено и как функционирует.

При переходе между вкладками внутри корпуса (например, KorpusDK -> The Danish Dictionary), шапка и заголовки разделов меняют цвет, что помогает ориентироваться: каждому разделу сайта присвоен определенный цвет.

Возможно, некоторые кнопки стоило бы немного увеличить в размере - например, печать или стрелки с выпадающими списками. Помимо того, что эти кнопки крошечные, они окрашены в светло-серый цвет, из-за чего почти сливаются с фоном сайта.

На странице поиска есть подсказки, но на них необходимо нажимать и переходить в новую вкладку - это не очень удобно. Гораздо более функциональны были бы всплывающие окна.

В целом, сайт выглядит довольно устаревшим, угловатым, ему не хватает мобильности, чтобы идти в ногу со временем. Дизайн нуждается в модификации. Однако он очень понятен интуитивно - знаки и кнопки имеют понятный практически любому пользователю интерфейс (стрелочки, вопросительные знаки, иконки листа, принтер и т.д.), все они расположены на самом видном месте, и при этом их совсем немного - ровно столько, сколько необходимо. Это важно, если мы на первых порах не хотим потеряться в обилии функций и подфункций корпуса.

2. Onboarding (глазами новичка)

Ссылка на него располагается во многих источниках, в том числе и в Национальном корпусе русского языка. При открытии сайта KorpusDK сразу же бросаются в глаза кнопки «Home», “The Danish Dictionary” (словарь современного датского языка), «Dictionary of the Danish Language» (исторический словарь датского языка) и, собственно, сам «KorpusDK», который и будет нас в дальнейшем интересовать. Справа вверху без труда замечаем, что можем перевести страницу на английский язык для удобства работы, а также имеем возможность посмотреть карту сайта – это именно то, что нужно новичку, впервые открывшему датский корпус. При переходе на страницу Корпуса, мы сразу же видим в центре строку поиска. Главная страница не перегружена лишней информацией и ссылками, что позволяет нам сразу же обратить внимание на руководство пользователя.

Есть образцы запросов (Example 1 и 2), но нам всё и так интуитивно понятно - почти все кнопки с функциями подписаны, нам остается лишь вбить нужное слово в строку и щёлкнуть по кнопкам с интересующими параметрами. Их мы также без труда находим на самом верху страницы.

3. Помощь пользователю

Сразу же на главной страницу расположена ссылка на гид по корпусу. В этом разделе рассматриваются все варианты поиска в корпусе. В левом меню отображаются теги, по которым удобно ориентироваться во вспомогательных статьях.

Гид поясняет как работать в корпусе, используя подробную инструкцию со скриншотами о каждом виде поиска, исправленных выражениях, списке сокращений и ловушках. Так же, в разделе есть развернутый FAQ и факты о корпусе.

Стоит отметить, что в в правом углу отображается блок с ссылками на разноплановый поиск. Удобное дополнение. Пользователь может сразу применить полученные знания на практике. Только эта возможность доступна при переводе сайта на датский язык. Подсказки при непосредственном поиске достаточно неудобные (растягиваются в блоке практически до конца страницы).

Видеохэлпы и другие инструкции с внешних источников отсутствуют.

4. Функционал

Состав корпуса

Объем корпуса составляет 56 млн слов, что, как нам кажется, немало для такого непопулярного языка как датский. Тексты, представленные в корпусе, довольно разноплановые: художественные, пресса, юридические, религиозные.

В комплект датского корпуса входят следующие корпуса:

- Korpus 2000: Состоит из текстов периода 1998-2002 годов.
- Korpus 90: Состоит из текстов с 1983-1992 годов.

Тексты составляют подмножество корпуса Датского словаря.

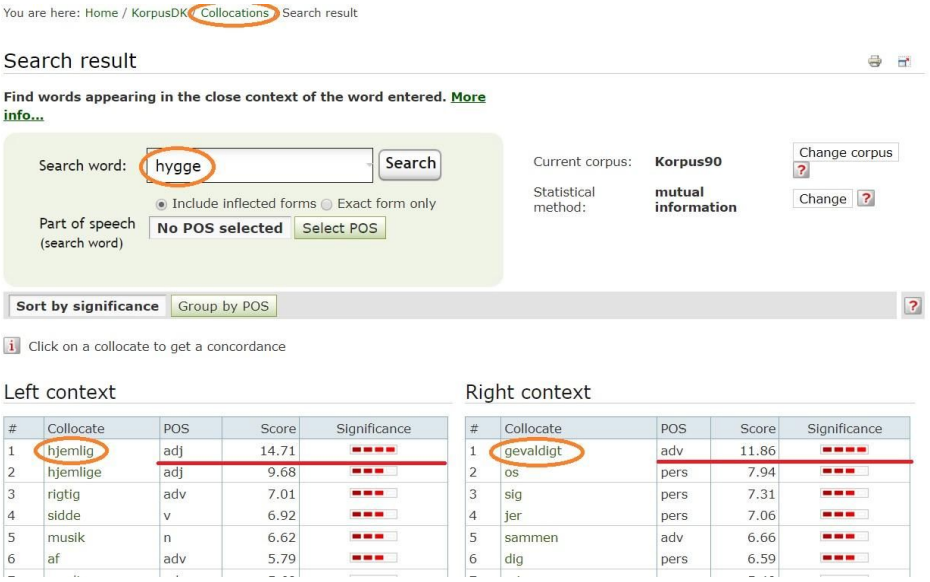
- KorpusDK: Состоит из Korpus 2000 и Korpus 90 совместно.

Возможности поиска

Поиск по данному корпусу подразделяется на три типа: *Concordance, Collocation (Standard, Extended, Formal search), Set phrases.*

1. **Collocations** – слова, которые статистически более часто встречаются в языке и речи вместе. Однако, допускается более или менее свободный порядок. Выдача результатов - по правому и левому контекстам. Таблица предоставляет сведения о частоте встречаемости коллокации /score/ и о степени сочетаемости слов друг с другом /significance/.

Попробуем найти наиболее значимые коллокации небезызвестного всем слова hygge – уют.



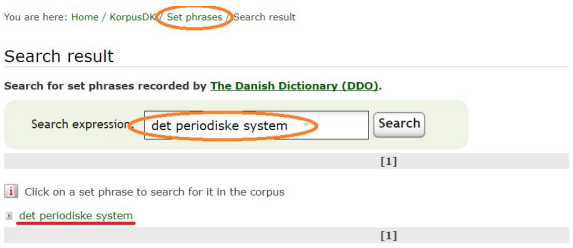
Итак, в левом контексте наиболее часто коллоцирующееся с hygge – hjemlig– «домашний» (прилагательное), в правом контексте видим gevaldigt – «чрезвычайно» (наречие).

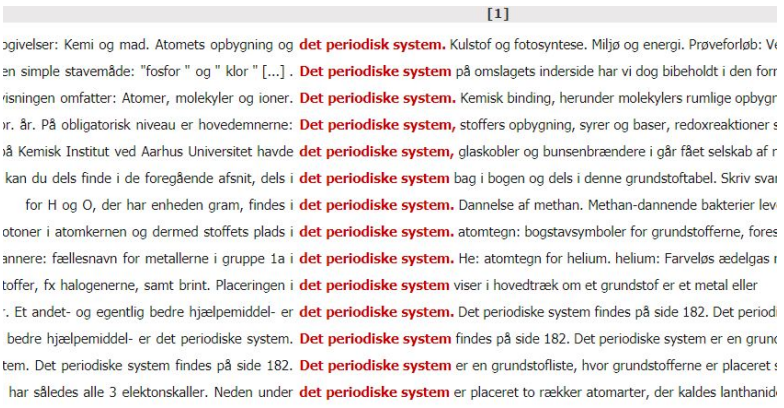
Более того, если мы нажмем на одно из слов из правого или левого контекста, корпус сделает выдачу примеров со всей коллекцией.

Отметим, что корпус предоставляет сортировку по правому и левому контекстам и по самому запросу, по начальной и конечной буквам. Также возможна функция выравнивания по правому или левому слову.

2. **Set-phrases** – два или более слова, сочетание которых более или менее фиксированно. К ним, в основном, относятся идиомы, фразеологизмы, пословицы, какие-либо термины, устойчивые выражения.

Введём, например, какой-нибудь термин - det periodiske system.

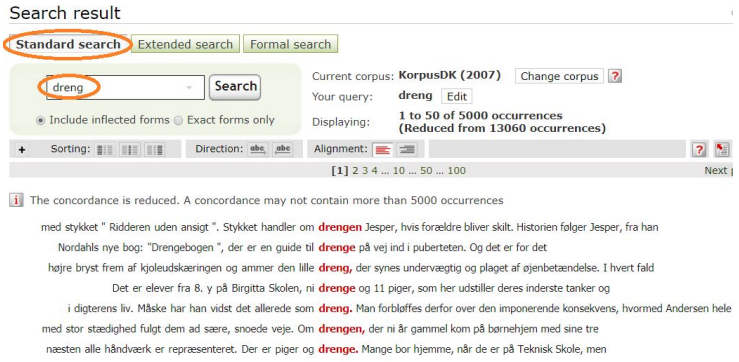




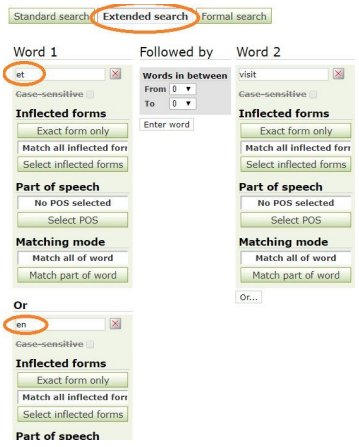
Однако существует ограничение:
корпус выдаёт лишь *set-phrases*, зафиксированные главным датским словарём (DDO).

3. **Concordance** выдаёт список всех употреблений данного слова в контексте. В Concordance'е существуют 3 вида запросов:

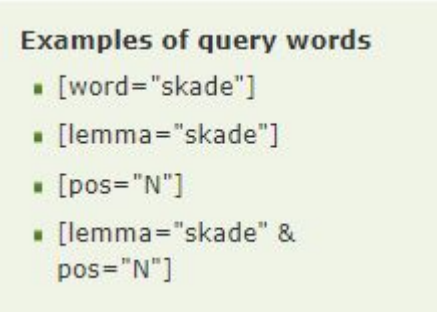
- *Standard search* – самый простой и удобный в использовании вид. В окно поиска вводится одно или несколько слов; возможен поиск по лемме и по точной форме слова. Также предоставляется возможность поиска слов различной длины по его началу, середине или концу с помощью универсальных метасимволов в Wildcards.



- *Extended Search* удобен для запросов из нескольких слов. Помимо изменяемой и неизменяемой форм и выбора части речи, расширенный поиск предоставляет выбор изменяемых частей (вручную),а также разрешает задавать поиск по какой-либо части слова (matching mode). Кроме того, для каждого слова есть функция OR, функция case-sensitive (которая считывает точно то, что введено в окне, вплоть до регистра), количество слов между и, конечно, followed by, т.е. добавление слов.



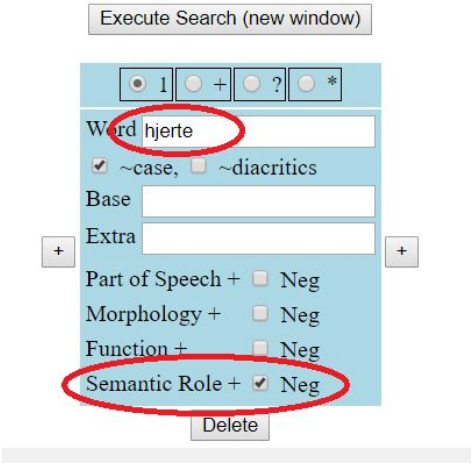
- *Formal Search,* так же как и Standard, может включать в себя 1 или несколько слов. Разница заключается в формате запроса. Слово вводится в квадратных скобках и имеет при себе 1 или более атрибутов, показывающих тип запрашиваемой информации, это могут быть word, lemma, pos или ortho. Поиск подходит и для нескольких слов, если они записаны в правильной последовательности.



Слои разметки

Согласно информации о разметке на сайте корпуса, корпус снабжён *морфологической* (частеречной), *синтаксической* и *мета-разметкой*.

- **Морфологическая и мета-разметка** (включающая название документа, источник, дату публикации, имя автора, дату рождения автора, а также целые абзацы, “окружающее” тот или иной запрос) легкодоступны в KorpusDK:



- Однако на **синтаксическую разметку** дана ссылка на “побочный” корпус. Попытка узнать информацию о синтаксисе того или иного слова обернулась крахом: “побочный” корпус выдаёт ошибку и что с этим делать - непонятно.