RuSkELL: ONLINE LANGUAGE LEARNING TOOL FOR RUSSIAN LANGUAGE

Ольга Буйволова (bublixa@gmail.com), Ольга Культепина (okultepina@gmail.com), Анна Малолетняя (maloletnyaya@gmail.com)
Высшая школа экономики, Москва, Россия

RuSkELL ("Russian + Sketch Engine for Language Learning") is a new online resource addressed to researchers and learners of Russian. The prototype of the resource is English SkELL (https://skell.sketchengine.co.uk/run.cgi/skell). RuSkELL is based on a specially pre-processed corpus and a sketch grammar written in CQL (corpus query language). Its interface allows users to search for language examples with a query word, extract its frequent collocates and show words by similarity. The aim of the paper was to adapt the existing SkELL tool to Russian, improve its performance and make it friendlier for Russian users. This research presents the solutions of several problems: modifying sketch grammar rules to exclude irrelevant output and partly resolve homonymy for certain Russian grammatical forms providing collocation groups with easy-to-understand Russian labels. In its improved version, RuSkELL is expected to become a reliable and flexible language resource catering to students, teachers, researchers and lexicographers of the Russian language.

Key words: language learning, sketch grammar, online language tool, collocations

Введение

В настоящей работе мы презентуем подготовку RuSkELL (Russian + Sketch Engine for Language Learning) онлайн-сервиса для изучения русского языка. Прототип RuSkELL - английская версия SkELL [Baisa, Suchomel 2014]. Сервис основан на специально подготовленном корпусе и имеет уникальный интерфейс с тремя функциями, который позволяет находить примеры употреблений слова, контекстов и значимых сочетаний, а также близкие по смыслу слова, представленные в виде облака слов. Важно отметить, что все коллокации отбираются по строгим правилам, кроме того, сервис имеет собственный CQL. Пока что в нашем распоряжении находится только бета-версия, которая была подготовлена программистами вне нашей группы и в работе которой было обнаружено множество проблем. Нашими задачи были: улучшение интерфейса, оптимизация выдачи и анализ возникающих недочетов. Наша цель - не только сделать RuSkELL более удобным для пользователей, но и описать процесс конструирования лингвистических данных и адаптировать сервис под специфику русского языка. После адаптации RuSkELL станет полезным инструментом не только при обучении русскому языку, но и для лексикографии.

Корпус RuSkELL

RuSkELL - инструмент для помощи лексикографам, студентам и преподавателям в освоении русского языка. Он представляет собой аналог EngSkELL, с которым успешно работают англоязычные пользователи.

Корпус RuSkELL был специально разработан в 2011 году на основе текстов из Рунета. Загруженные из домена .ru документы прошли через фильтр GDEX (букв. Хорошие Словарные Примеры), в результате чего остались только предложения средней длины с употреблением высоко частотных слов. Таким образом из изначального массива данных непосредственно в корпусной базе RuSkELL остались 68,232,088 предложений с суммарным количеством слов 975,742,959.

RuSkELL имеет три функции, представленные во вкладках "Examples", "Word Sketch", и "Similar words" (Рис. 1). В первой вкладке "Examples" даны примеры предложений с заданным словом (или его формами). Эта функция позволяет пользователям ознакомиться

с конкретным употреблением слова и с его контекстами. Функция "Word Sketch" показывает частотные сочетания с заданным словом, а "Similar words" - список близких по значению слов, которые необязательно синонимичны заданному слову. "Word Sketch" формирует представление о грамматических коллокациях, идиомах, полисемии и др. и позволяет изучить те особенности современного русского языка, описание и объяснение которых не всегда есть в словарных статьях, справочниках или грамматических описания.

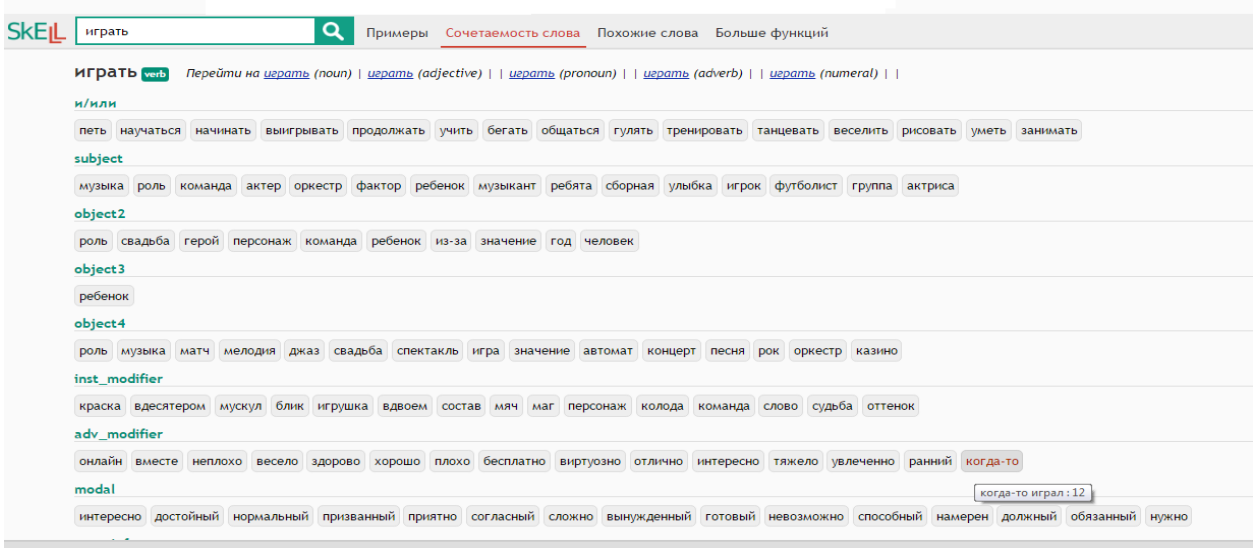

Рис. 1. Интерфейс RuSkELL

RuSkELL показывает релевантные частотные сочетания и примеры употребления, которые имеют широкое распространение в современном русском языке.

Выдача примеров и сочетаний с искомым словом определяется при помощи Sketch Grammar (Рис. 2). В нем записаны правила автоматического определения грамматических отношений между лексическими единицами, и в соответствии с ними формируется список примеров и сочетаний к заданному слову.

```
=subject/subject_of
#2:noun_nom [pos="R"]? 1:"Vmi.*"
#1:"Vmi.*" [pos="R"]? 2:noun_nom
2:noun_nom [pos="A" & case="[ng]"]? noun_gen{0,2} [pos="Q"]{0,3}
[tag="Afpns.s"|pos="R"|tag="Afc.*"]{0,3} 1:"Vmi.*"
1:"Vmi.*" [pos="Q"]{0,3} [tag="Afpns.s"|pos="R"|tag="Afc.*"]{0,3} adj_nom{0,3}
2:noun_nom
```

Рис. 2. Правило для группы “Object2” в Sketch Grammar

На сегодняшний день немного работ, посвященных системе SkELL: [Baisa, Suchomel], [SkELL], [Kilgarriff et al. 2014]. В указанных исследованиях подробно разбираются механизмы работы Sketch Engine и английского аналога SkELL, а также анализируются особенности существования лексикографии в современном онлайн-пространстве.

Данная работа берез за основу тезисы, подготовленные В.Ю. Апресян, Vit Baisa и нашей проектной группой для участия в XVII EURALEX International Congress (6-10 September 2016, Tbilisi, Georgia) [Baisa et el. 2015].

RuSkELL: диагностика проблем

Несмотря на несомненные преимущества RuSkELL, многие ошибки в его работе мешают эффективному использованию сервиса.

Во-первых, в отличие от своего английского аналога, интерфейс RuSkELL недостаточно понятен и удобен для пользователей, что показал проведенный нами опрос. Слишком большая выдача и большое количество групп в “Word Sketch” усложняют процесс поиска и обращения к конкретным примерам и путают пользователей.

После опроса нами были предложены новые варианты названий для ряда групп. В дальнейшем мы доработала их до оптимального вида, включив в названия особый символ %w, который при выдаче автоматически заменяется на заданное в поиск слово.

Старые названия	Новые названия
subject of %w/verbs with %w as subject	подлежащее при %w/глагол с %w в роли подлежащего
object2/object2_of	дополнение в родительном падеже при %w / глаголы с %w в роли дополнения в родительном падеже
object3/object3_of	дополнение в дательном падеже при %w / глаголы с %w в роли дополнения в дательном падеже
object4/object4_of	дополнение в винительном падеже при %w / глаголы с %w в роли дополнения в винительном падеже
inst_modifier/inst_modifies	дополнение в творительном падеже при %w / глаголы с %w в роли дополнения в творительном падеже
gen_modifier/gen_modifies	%w подчиняет существительное в родительном падеже / %w подчиняется существительному в родительном падеже
a_modifier/modifies	определение при %w/ существительное с %w в роли определения
adv_modifier / adv_modifies	обстоятельство при %w / глаголы с %w в роли обстоятельства
adv_modifier_verb / adv_modifies_verb	обстоятельство при %w / глаголы с %w в роли обстоятельства
adv_modifies_adj/adj_modifies _adv	обстоятельство при %w / прилагательное с % в роли обстоятельства

Таб. 1. Новые названия групп

Во-вторых, мы столкнулись со слишком большой выдачей коллокаций во вкладке “Word Sketch”. Некоторые из них нерелевантны, и наша задача - минимизировать их количество.

Многие примеры были некорректно отнесены в группы коллокаций из-за грамматической омонимии. Например, глагол “верить” управляет дательным и винительным падежами, также есть предложное управление со значением стимула “верить во что-то”. На запрос на глагол “верить” мы видим группу Object2 (“ложь”, “библия”, “информация”, “власть”), управляющую родительным падежом. Хотя в данном случае речь идет о значении “верить чему-то”, т.е. Object2 является неким источником информации, что неправильно, потому что данные примеры должны быть в группе Object3. Из-за того, что RuSkELL не умеет различать омонимию, пользователи сталкиваются с подобными случаями. Некоторые из таких ошибок с неверной выдачей в группах возможно исправить, сформировав особое правило для автоматической системы распределения (т.е. для sketch grammar). Тем не менее, часть подобных сбоев относится к специфике грамматики русского языка. Поэтому определенный процент излишней выдачи останется неизменным.
На данный момент этот вопрос находится в разработке, и мы тестируем различные способы сокращения “мусора” в выдаче, например, задав отдельные глаголы списком.
В ходе тестирования RuSkELL, мы также обнаружили недостатки в выдаче при поисковом слове-наречии, вроде *плохо, хорошо, точно*. Смешение частей речи в одну группу под общим названием adv_modifier делало неэффективном пользовательский поиск.
Именно поэтому для наречий мы разделили исходное правило в sketch grammar, создав две группы сочетаемости - отдельно с глаголом и отдельно с прилагательным. Первое правило в разработке выглядит подобным образом:

```
*DUAL
=adv_modifies_verb
1: "Vm.*" 2: [pos="R"]
2: "Vmi.*" 1:[pos="R"]
2: "Vmp.*" 1:[pos="R"]
2: "Vmn.*" 1:[pos="R"]
```

Ср.: query in ruTenTen: (meet[pos="R"] [tag="Vm.*"] 0 1))
Это правило включает отношения с глаголами в индикативе, причастие, инфинитив. Наречия находится в препозиции, что снижает вероятность появления нерелевантных сочетаемостей.
Следующее правило показывает коллокации между наречием и прилагательным:

```
*DUAL
=adv_modifies_adj
1: [pos="A"] 2: [pos="R"]
2:[pos="R"] 1:adj_nom
2:[pos="R"] 1:adj_gen
2:[pos="R"] 1:adj_acc
2:[pos="R"] 1:adj_dat
2:[pos="R"] 1:adj_inst
2:[pos="R"] 1:adj_loc
```

Ср.: query in ruTenTen: (meet[pos="R"] [pos="A"] -1 0))
Для обеих групп были разработаны названия, схожие с названиями для других групп, где знак %w при запросе будет автоматически заменяться на заданное в поиске наречие: обстоятельство при %w / глаголы с %w в роли обстоятельства и при %w / прилагательное с % в роли обстоятельства

Добавление новых частей речи

При добавлении новой части речи в скетч грамматику, мы руководствовались несколькими доводами. Например, частотность коллокаций, в данном случае, числительные являются информативным дополнением. В экспериментальную версию

RuSkELL были включены правила, находящие как порядковые, так и количественные числительные.

Рассмотрим порядковые числительные, у которых есть одна группа **num_modifies**. Например, ‘второй этаж’, ‘второй тайм’ и ‘вторая половина’. Количественные числительные представлены двумя группами **numeral_object2_of** и **numeral_inst**. Первая содержит лексемы с высокой частотностью, такие как: ‘два года’, ‘два раза’ и ‘два десятка’. Во второй группе есть коллокации типа ‘двумя руками’, ‘двумя годами’, ‘двумя рядами’.

Также на данный момент мы работаем над вопросом добавления депиктивных конструкций в группу inst_modifier (сочетание с Тв.п.). При удачной разработке правила в этой группе можно будет увидеть такие сочетания, как *выглядеть усталым* или *казаться странным*.

Заключение

Конечным результатом работы проекта будет являться обновленный RuSkELL с максимально релевантной выдачей примеров и, возможно, с дополнительными функциями, которые помогут понять некоторые особенности грамматики русского языка (например, определенные элементы разметки на специфических устойчивых сочетаниях).

Литература

Апресян 1967 - Апресян Ю.Д. Экспериментальное исследование русского глагола. М.: Наука, 1967.
Зализняк 2006 - Зализняк А.А. Многозначность в языке и способы ее представления. М.: Языки славянских культур, 2006.
Ляшевская, Шаров 2009 - Ляшевская, О. Н. Шаров, С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
НКРЯ - Национальный корпус русского языка http://www.ruscorpora.ru/.
Падучева 1997 - Падучева Е.В. Родительный субъекта в отрицательном предложении: синтаксис или семантика? // Вопросы языкознания, 1997, № 2.
Якобсон Р.О. К общему учению о падеже: Общие значения русских падежей. 1936.
Atkins et al. 2008 - Atkins, BT Sue, and Michael Rundell. The Oxford guide to practical lexicography. Oxford University Press, 2008. [Электронный ресурс] Доступ: http://site.ebrary.com/lib/hselibrary/reader.action?docID=10246237&ppg=6 ;
Baisa, Suchomel - Baisa, Vít, Suchomel, Vít . SkELL: Web Interface for English Language Learning. [Электронный ресурс].
Доступ: https://www.researchgate.net/publication/275007862_SkELL_Web_Interface_for_English_Language_Learning ;
Baisa et al. 2015 - Baisa V., Apresjan V., Buivolova O., Kultepina O., Maloletnjaja A., Iskhakov T., Suchomel V. RuSkELL: Online Language Learning Tool for Russian Language. 2015
Kilgarriff et al. 2014 - Kilgarriff Adam , Baisa Vít, Bušta Jan, Jakubíček Miloš, Kovář Vojtěch, Michelfeit Jan, Rychlý Pavel, Suchomel Vít. The Sketch Engine: ten years on. Lexicography. July 2014, Volume 1, Issue 1, pp 7-36
Rusgram - Проект корпусного описания русской грамматики (http://rusgram.ru)
SkeLL - What is SkELL? https://www.sketchengine.co.uk/skell/
Wierzbicka 1980 - Wierzbicka A. The case for surface case. Linguistica extranea. 1980.