АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ СТИМУЛЬНЫХ ПРЕДЛОЖЕНИЙ ДЛЯ СИНТАКСИЧЕСКИХ ЭКСПЕРИМЕНТОВ

Студеникина К.А. (xeanst@gmail.com)
Московский государственный университет им. М.В. Ломоносова (Россия, Москва)

Данная работа находится на пересечении интересов теоретической и прикладной лингвистики, поскольку в ней методами компьютерной лингвистики решаются вопросы, актуальные для применения экспериментального метода в теоретическом синтаксисе. На примере конкретного лингвистического эксперимента рассматривается возможность автоматизации процесса создания стимульного материала путём осуществления морфологических и синтаксических трансформаций текста с использованием фреймворка Udapi. Также будет продемонстрировано, как полученные предложения могут быть использованы для улучшения качества обработки текстов языковыми моделями.

Ключевые слова: автоматическая генерация, экспериментальный синтаксис, синтаксические трансформации, синтаксический парсинг, Udapi

AUTOMATIC GENERATION OF STIMULI FOR SYNTACTIC EXPERIMENTS

Studenikina K.A. (xeanst@gmail.com)
Lomonosov Moscow State University (Moscow, Russia)

This study deals with both theoretical and computational linguistics issues as it considers problems of experimental method in theoretical syntax using computational methods. I examine the possibility of experimental stimuli automatic generation by means of morphological and syntactic transformations in framework Udapi by implementing the proposed algorithm in a particular syntactic experiment. I also consider how obtained stimuli sentences could be applied to additional training of language models and improving the quality of syntactic parsing.

Keywords: automatic generation, experimental syntax, syntactic transformations, syntactic parsing, Udapi

1. Введение: об эксперименте в лингвистике.

Экспериментальный метод научного исследования широко распространён в лингвистике. Он позволяет исследователю манипулировать некоторыми факторами, называемые *независимыми* переменными, так или иначе изменяя их значения или *уровни*, и изучать их влияние на другой фактор, именуемый *зависимой* переменной. Конкретные значения данных факторов определяются целью, гипотезой и методикой эксперимента.

Если говорить о синтаксических экспериментах, то в них в качестве стимульного материала используются предложения и словосочетания, в качестве независимых переменных, которые варьируются для стимульного материала – определённые характеристики слов, словосочетаний и предложений: морфологические характеристики (род, падеж, число, наклонение, время), тип передвижения (вопросительное, относительное), тип зависимой клаузы (инфинитивная, клауза с союзом *что*) и т.д. В Таблице 1, приведённой по работе (Фёдорова 2008: 11), представлен пример исследования: две независимых переменных с двумя уровнями, сочетание которых даёт четыре комбинации, называемые экспериментальными условиями. Данной схеме соответствует экспериментальный блок из четырёх предложений (по количеству условий), внутри каждого блока предложения максимально похожи и представляют одну лексикализацию.

		Независимая переменная Y	
		Уровень 1	Уровень 2
Независимая переменная X	Уровень 1	X1 – Y1	X1 – Y2
	Уровень 2	X2 – Y1	X2 – Y2

Таблица 1. Пример экспериментального исследования

Таким образом, для подготовки экспериментального блока исследователю необходимо подобрать значительное количество предложений определённой структуры и осуществить с ними ряд одинаковых трансформаций. Поскольку отбор предложений основан сугубо на их формальных характеристиках, как морфологические признаки и синтаксические роли членов предложения, представляется возможной автоматизация процесса подбора стимулов для синтаксических экспериментов. Насколько нам известно, ранее подобные исследования не проводились. Далее будет описан алгоритм, выбранный нами для осуществления автоматической генерации стимулов.

2. Автоматическая генерация экспериментальных стимулов: проблемы и пути решения.

Как было сказано во введении, стимульный материал представляет собой набор экспериментальных блоков, которые различаются набором лексем (лексикализациями), но имеют одинаковую морфологическую и синтаксическую структуру. Каждое предложение в блоке соответствует определённому экспериментальному условию, следовательно, количество предложений в блоке определяется количеством условий.

2.1. Проблемы, связанные с поиском стимульных предложений в корпусе текстов

Наиболее прямолинейным способом автоматизации подбора стимулов может показаться поиск нужных предложений в корпусе текстов по заданному морфосинтаксическому шаблону. Однако поиск готовых стимулов в корпусе невозможен по двум причинам.

В первую очередь, параметры, по которым отличаются предложения внутри блока, задаются самим лингвистом в качестве уровней независимых переменных. Они зависят от цели исследования и варьируются в разных экспериментах: например, при изучении нарушений в согласовании это может быть число именных групп (Bock et al. 1991), при изучении островных ограничений на передвижение – наличие островной структуры и тип передвигаемой составляющей (Goodall 2015). Поскольку в корпусе собраны предложения, взятые из живой письменной и устной речи, в нём отсутствуют предложения, имеющие максимально схожий лексический состав и различные морфосинтаксические характеристики.

Кроме того, стимульные предложения для синтаксических экспериментов зачастую имеют либо сложную структуру, как несколько вложенных клауз (1), либо структуру, которая считается носителями неграмматичной, как множественные wh-вопросы с нарушением эффектов превосходства (2). Стимульные предложения такого типа или совершенно отсутствуют в корпусе, или их можно найти довольно редко, и найденного количества, скорее всего, будет недостаточно для проведения полноценного эксперимента.

(1) [The nurse from the clinic] supervised [the administrator [who scolded the medic]] while a patient was brought into the emergency room.

'Медсестра из клиники наблюдала за администратором, который ругал медика, пока пациента вели в отделение неотложной помощи.' (Hofmeister et. al. 2010)

(2) Pat wondered what who read.

'Пэт поинтересовалась, кто что читает.' (Arnon I. 2006)

2.2. Создание экспериментального блока с помощью фреймворка Udapi

Предлагаемый алгоритм автоматической генерации стимулов состоит из двух этапов. Первый этап включает поиск базового предложения в корпусе по морфосинтаксическому шаблону. Искомые предложения не являются готовыми стимулами, а представляют собой исходную конфигурацию, из которой уже можно получить предложения для всех экспериментальных условий. Второй этап предполагает осуществление этих конфигураций с помощью морфологических и синтаксических трансформаций с использованием инструментов автоматической обработки текста. Результатом становится готовый экспериментальный блок: количество предложений в нём равно количеству экспериментальных условий, они представляют собой одну лексикализацию, то есть имеют максимально похожий лексический состав.

Реализация описанного алгоритма возможна при помощи программного интерфейса и фреймворка Udapi (Popel et al. 2017), реализованного на основе проекта универсальных зависимостей Universal Dependencies (UD, Nivre et al. 2016). В ходе данного проекта осуществляется создание трибанков – корпусов, синтаксически аннотированных в рамках грамматики зависимостей, – для различных языков мира с использованием унифицированного формата аннотации, что позволяет облегчить межъязыковую обработку естественного языка и даёт возможность проводить сравнительные лингвистические исследования. Аннотирование включает в себя разбиение на предложения, токенизацию, частеречную разметку, лемматизацию и синтаксическую разметку.

Программный интерфейс Udapi используется для операций с форматом UD: визуализации деревьев зависимостей, преобразования форматов, поиска, трансформаций, парсинга зависимостей, оценки качества и т.д. Использование Udapi для нашей задачи обусловлено тем, что данный фреймворк позволяет осуществлять операции над деревом зависимостей, а именно обращаться к классам для представления синтаксических данных. Для каждого предложения в Udapi определены классы `Root` (специальный искусственный узел для корня дерева, который добавляется в вершину дерева в формате ConLL-U) и `Node` (соответствует узлу в дереве зависимостей). Методы класса `Node` позволяют обращаться к параметрам узла (`node.form`, `node.lemma`), осуществлять передвижение узла в дереве (`node.shift_after_node(x)`, `shift_before_subtree(x)`) и его удаление (`node.remove`), а также обращаться к узлу, находящемуся выше или ниже в дереве относительно данного узла (`node.parent`, `node.children`). Всё это позволяет рассматривать предложение как дерево зависимостей, его части – как узлы данного дерева, а, значит, делает возможными манипуляции с синтаксической структурой предложения.

Таким образом, с помощью методов класса `Root` и `Node` фреймворка Udapi возможны необходимые для создания стимульных предложений операции: поиск исходного предложения по морфосинтаксическому шаблону и применение к нему морфосинтаксических трансформаций.

3. Реализация автоматической генерации стимулов на примере синтаксического эксперимента с множественными *wh*-вопросами

Описанный в предыдущем разделе алгоритм автоматического подбора стимульных предложений был реализован нами на примере синтаксического эксперимента с множественными wh-вопросами, то есть частными вопросами с несколькими вопросительными словами.

3.1. Множественные wh-вопросы с точки зрения теоретического синтаксиса

Одним из основных параметров варьирования в множественных wh-вопросах в разных языках является отсутствие или наличие ограничений на порядок wh-слов или эффектов превосходства ((3), Chomsky 1973). Существуют факторы, ослабляющие эффекты превосходства, например, дискурсивная связанность wh-слов (d-linking). Если wh-слово дискурсивно связано, ответ выбирается из заданного множества (Pesetsky 1987), и эффекты превосходства могут нарушаться (4).

(3) a. Who read what?
b. *What did who read?
'Кто что прочитал?'
(4) a. Which man did you persuade to read which book?
b. Which book did you persuade which man to read?
'Какого человека ты убедил прочитать какую книгу?'

Для русских множественных wh-вопросов наблюдаются противоречия как в теоретических подходах, так и в анализируемых эмпирических данных. В работе (Stepanov 1998) постулируется отсутствие эффектов превосходства. В экспериментальном исследовании напротив (Meyer 2004) утверждается наличие эффектов превосходства.

Однако описанные выше работы нацелены на изучение множественных вопросов, где в качестве вопросительных слов выступают субъект и объект. Для выяснения, обоснованы ли выявленные закономерности различием подлежащего и дополнения или структурным приоритетом, следует проверить, возникнет ли тот же эффект, если снять противопоставление подлежащего и дополнения и оставить только с-командование.

3.2. Экспериментальное исследование множественных wh-вопросов и автоматическая генерация стимулов

Конфигурация, взятая в роли исходной для стимульных предложений, представляет собой предложение с зависимой инфинитивной клаузой и двумя объектами: один из них одушевлённый и является зависимым глагола в главной клаузе, другой – неодушевлённый и зависит от инфинитива.

(5) субъект глагол-FIN объект-ANIM глагол-INF объект-INAN
Маша заставила друзей купить шоколадки.

Далее в исходном предложении необходимо осуществить вопросительный вынос двух объектов с учётом экспериментальных факторов: порядка слов: прямой (одушевленный + неодушевлённый объект) vs. обратный (неодушевлённый + одушевлённый объект), дискурсивной связанности одушевленного объекта: не связан (кого) vs. связан (каких X), дискурсивной связанности неодушевленного объекта: не связан (что) vs. связан (какие X). Пример экспериментального блока представлен в Таблице 2.

Пример	Порядок слов	Дискурс. связ. wh-объекта-ANIM	Дискурс. связ. wh-объекта- INAN
Кого что Маша заставила купить?	прямой	-	-
Что кого Маша заставила купить?	обратный	-	-
Каких друзей что Маша заставила купить?	прямой	+	-
Что каких друзей Маша заставила купить?	обратный	+	-
Кого какие шоколадки Маша заставила купить?	прямой	-	+
Какие шоколадки кого Маша заставила купить?	обратный	-	+
Каких друзей какие шоколадки Маша заставила купить?	прямой	+	+
Какие шоколадки каких друзей Маша заставила купить?	обратный	+	+

Таблица 2. Трансформации исходного предложения по экспериментальным условиям

Автоматическая генерация стимульных предложений заключалась в следующем. В начале осуществлялся поиск нужных базовых предложений по морфосинтаксическому шаблону (5). Поиск осуществлялся в подкорпусе автоматически собранного в рамках проекта deeppavlov корпуса новостных текстов (Burtsev et al. 2018), содержащем глагольные инфинитивы. Подкорпус, размеченный с помощью анализатора[1], основанного на энкодере BERT (Devlin et al. 2018) и рекуррентной нейронной сети, содержит 101000 предложений в формате UD. Морфосинтаксический шаблон для поиска, реализованный с помощью программного интерфейса Udapi, заключался в следующем: (i) среди всех предложений находятся те, которые имеют зависимый от корня финитный глагол, (ii) среди отобранных находятся те, где данный глагол имеет определённые зависимые: субъект, одушевлённый объект-существительное, глагольный инфинитив, (iii) в конце остаются предложения, где инфинитив имеет в качестве зависимого неодушевлённый объект-существительное. В результате данного поиска была отобрано 326 предложений нужной базовой конфигурации.

На следующем этапе осуществлялись вопросительные трансформации. Программная реализация включает две функции для вопросительных слов: для одушевлённого и неодушевлённого объекта. Функция для одушевлённого объекта имеет параметр `d-linking` для дискурсивной связанности: при значение `False` субъект дискурсивно не связан, при значении `True` – дискурсивно связан. Функция для неодушевлённого объекта имеет параметр `d-linking`, а также параметр `superiority` для эффектов превосходства: при значении `True` порядок слов прямой, при `False` – обратный. Данная функция возвращает wh-узел для неодушевлённого объекта, в частности его расположения относительно одушевлённого wh-объекта. Поскольку существительные в дискурсивно связанных wh-составляющих должны стоять во множественном числе, для изменения существительных по числу был использован морфологический анализатор pymorphy2 (Korobov 2015).

Далее следуют функции для восьми трансформаций, соответствующие восьми условиям в Таблице 2. Функции берут на вход узлы для субъекта, финитного глагола, инфинитива, а также результаты функций для wh-узлов. Они различаются тем, какие значения принимают параметры дискурсивной связанности и эффектов превосходства. Каждое найденное в подкорпусе предложения подвергается всем восьми вопросительным трансформациям, результатом трансформации каждого предложения становится готовый экспериментальный блок.

Наконец, осуществляется запись полученных с помощью трансформаций предложений в формате CONLL-U. Они представляют собой 326 экспериментальных блоков, то есть 326 предложений на каждое условие, всего 2608 стимулов. Поскольку для эксперимента необходимо использовать как минимум четыре лексикализации на каждое условие, для данного эксперимента потребуется 32 экспериментальных блока, включающие 256 стимульных предложений. Даже при учёте того, что не все из автоматически сгенерированных предложений будут звучать достаточно естественно и некоторые из них не могут быть использованы в эксперименте, полученное количество предложений является достаточным для проведения описанного синтаксического эксперимента.

Таким образом, описанный в данном разделе конкретный пример автоматического порождения стимулов для эксперимента с множественными *wh*-вопросами показывает, что автоматический метод подбора стимулов обеспечивает достаточное для проведения эксперимента количество предложений и может быть использован в дальнейшем в других экспериментах при создании нужного морфосинтаксического шаблона для базовых предложений и функций для необходимых трансформаций.

[1] http://docs.deeppavlov.ai/en/master/features/models/morphotagger.html#advanced-models-bert-and-lemmatized-models

3.3. Возможность применения полученных стимульных предложений для улучшения качества моделей обработки естественного языка

Как было сказано в разделе 2.1, стимульные предложения для синтаксических экспериментов зачастую имеют сложную структуру, из-за чего такие предложения не могут быть корректно проанализированы существующими моделями для обработки естественного языка. В качестве примера была взята модель[2], использованная для разметки подкорпуса для поиска предложений и основанная на энкодере BERT (Devlin et al. 2018) и рекуррентной нейронной сети. Точность её морфологического анализа на тестовой выборке корпуса SynTagRus (Дяченко и др. 2015, Droganova et al. 2018), на котором обычно происходит сравнение, 97.8%, а точность синтаксического разбора (Labeled Attachment Score) — 93.7%. Тем не менее, если применить данную модель к анализу множественных *wh*-вопросов, полученных с помощью морфосинтаксических трансформаций, синтаксический разбор будет иметь довольно высокое качество, но всё же отличное от корректной разметки. Основная проблема при парсинге данных стимульных предложений заключается в том, что модель распознаёт *wh*-составляющие как зависимые одного глагола: либо финитного в главной клаузе, либо инфинитивного.

Эта проблема, однако, может быть решена с помощью дообучения модели на полученных с помощью трансформаций примерах. При валидации на 20% обучающей выборки и обучении на оставшихся 80% качество морфологии выросло с 94% до 98%, качества синтаксического парсинга же достигает 100%, анализ предложений становится полностью корректным.

4. Заключение.

В данной работе была рассмотрена проблема автоматической генерации стимульных предложений для синтаксических экспериментов. Для формирования экспериментальных блоков необходимы (i) подбор лексикализации определённой синтаксической структуры, (ii) осуществление с лексикализацией морфосинтаксических преобразований, соответствующих экспериментальным условиям. Предполагалось, что, поскольку все эти этапы основаны на сугубо формальных характеристиках, как морфологические и синтаксические признаки, возможно автоматизировать процесс подбора стимулов для синтаксических экспериментов.

Предложенный алгоритм реализован с помощью программного интерфейса и фреймворка Udapi (Popel et al. 2017), реализованного на основе проекта универсальных зависимостей Universal Dependencies (UD, Nivre et al. 2016). Определённые для каждого предложения в Udapi классы `Root` и `Node`, а также их методы, позволяют рассматривать предложение как дерево зависимостей, его части – как узлы данного дерева. Это даёт возможность реализовывать синтаксические трансформации с синтаксической структурой предложения. Алгоритм автоматической генерации стимулов включает два этапа: во-первых, поиск исходного предложения по морфосинтаксическому шаблону, во-вторых, применение к нему морфосинтаксических трансформаций. В качестве конкретной реализации данного алгоритма была рассмотрена генерация стимулов к синтаксическому эксперименту с множественными *wh*-вопросами с восьмью условиями, в результате которой было получено 2608 стимулов: 326 экспериментальных блоков по восемь предложений в каждом. Таким образом, при создании нужного морфосинтаксического шаблона для базовых предложений и функций для необходимых трансформаций предложенный метод автоматического порождения стимульных предложений может быть использован в дальнейшем в других синтаксических экспериментах.

[2] http://docs.deeppavlov.ai/en/master/features/models/morphotagger.html#advanced-models-bert-and-lemmatized-models

Также было продемонстрировано, что полученные с помощью трансформаций предложения, обладающие нетривиальной синтаксической структурой, могут быть использованы для дообучения моделей обработки естественного языка и исправления возможных ошибок морфологического и синтаксического анализа.

Библиография.

Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г., Фролова Т.И., Цинман Л.Л. 2015. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Сборник «Национальный корпус русского языка: 10 лет проекту». Труды Института русского языка им. В.В. Виноградова. М.. Вып. 6. С. 272-299.

Федорова, О. В. (2008). Основы экспериментальной психолингвистики: принципы организации эксперимента. М.: Спутник, 23.

Arnon, I. (2006, October). Cross-linguistic Variation in a Processing Account: The Case of Multiple Wh-questions. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 32, No. 1, pp. 23-35).

Bock, K., & Levelt, W. (1994). Grammatical encoding.

Burtsev M., Seliverstov A., Airapetyan R., Arkhipov M., Baymurzina D., Bushkov N., Gureenkova O., Khakhulin T., Kuratov Y., Kuznetsov D., Litinsky A., Logacheva V., Lymar A., Malykh V., Petrov M., Polulyakh V., Pugachev L., Sorokin A., Vikhreva M., Zaynutdinov M. 2018. Deeppavlov: Open-source library for dialogue systems. In Proceedings of ACL 2018, System Demonstrations (pp. 122-127).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Goodall, G. (2015). The D-linking effect on extraction from islands and non-islands. *Frontiers in psychology*, *5*, 1493.

Chomsky, N. (1973). Conditions on transformations. A festschrift for Morris Halle, ed. by Stephen R. Anderson and Paul Kiparsky, 232-86. *New York: Holt*.

Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2014). Processing effects in linguistic judgment data:(super-) additivity and reading span scores. Language and Cognition, 6(1), 111-145.

Korobov M. 2015. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp. 320-332.

Meyer, R. (2004). Superiority effects in Russian, Polish and Czech: Judgments and grammar. *University of Leipzig/University of Regensburg*.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Tsarfaty, R. (2016, May). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659-1666).

Pesetsky, D. (1987). Wh-in-situ: Movement and unselective binding. *The representation of (in) definiteness*, *98*, 98-129.

Popel, M., Žabokrtský, Z., & Vojtek, M. (2017, May). Udapi: Universal API for universal dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)* (pp. 96-101).

Stepanov, A. (1998). On wh-fronting in Russian.