**Массив параметрических данных для русского языка RuParam**

П. В. Гращенков (МГУ), Л. И. Паско (МГУ), К. А. Студеникина (МГУ)

Одной из важнейших идей современной лингвистической теории является концепция параметров. Впервые она была предложена в работах Н. Хомского и в дальнейшем развита его последователями, см. [Chomsky 1981, 2005; Chomsky & Lasnik 1993] и др. Параметры универсальны и принимают одной из двух возможных значений: ветвление направо или налево от вершины, наличие или отсутствие морфологического падежа и т.д. Базовая грамматика конкретного языка является результатом настройки этих параметров. При этом, к сожалению, не существует определенного инвентаря параметров, с которым были бы согласны все лингвисты. Обычно лишь называется некоторое количество наиболее бесспорных параметров: например, направление ветвления, передвижение вершины, наличие морфологического падежа, pro-drop, согласование. В результате дискуссии о природе параметров появилась также идея более элементарных микропараметров, см. [Baker 2008; Li & Wei 2019] и др.

Параметризация языковых данных необходима не только в теории языка, но и в практических областях, в том числе, в лингводидактике. В частности, при изучении иностранного языка обучающиеся должны усвоить, по каким грамматическим параметрам изучаемый язык отличается от их родного языка и в чем именно заключаются эти отличия. Программа преподавания иностранного языка, как и содержание экзаменов на уровень владения языком, должны покрывать все релевантные для него грамматические параметры, чтобы приблизить языковую компетенцию учащихся к компетенции носителей.

Другая прикладная область, в которой стали востребованы языковые данные в параметрическом формате, — компьютерная лингвистика, где такие данные создаются для оценки близости генераций нейросетевых языковых моделей к языку носителей. Впервые набор данных по лингвистической приемлемости (Corpus of Linguistic Acceptability, CoLA) был предложен для английского языка [Warstadt, Singh, Bowman 2019]. Он включал в себя более 10 тысяч предложений, взятых из литературы по теоретическому синтаксису. Исследователи провели сравнение трансформерных моделей и выяснили, что наиболее высокое качество наблюдается для простых предложений, а также установили ряд некоторых других особенностей языка трансформерных моделей.

Аналогичные корпуса лингвистической приемлемости стали появляться и для других языков: ItaCoLA для итальянского [Trotta et al 2021], RuCoLA для русского [Mikhailov et al. 2022]. Развитием работы над CoLa стал корпус BLiMP (Benchmark of Linguistic Minimal Pairs), [Warstadt et al. 2020]. Он состоит из 67 отдельных наборов данных, каждый из которых содержит 1000 пар минимально отличающихся предложений: одно из предложений в паре грамматично, второе — нет. Предложения были сгенерированы нейросетевой моделью по специальным шаблонам.

Существующие на данный момент параметрические корпуса имеют ряд проблемных мест. Перечислим наиболее критичные: i) в некоторых из них отсутствует классификация по грамматическим параметрам; ii) часть примеров синтезирована искусственно при помощи нейросетей; iii) такие корпуса часто включают в себя скорее вариативные, а не однозначные параметры; iv) иногда — например, в RuCola — процент неграмматичных вариантов достаточно незначителен.

Таким образом, как минимум для русского языка (а, возможно, и для других) актуальна задача создания корпуса лингвистической приемлемости, который включал бы в себя базовые систематизированные параметры и ограничения в виде пар грамматичных и неграмматичных примеров. В нашем докладе мы представим новый

параметрический датасет для русского языка, RuParam. Исходно датасет создавался для оценки больших языковых моделей, но может применяться и за пределами компьютерной лингвистики. Возможные варианты применения нашего корпуса включают также исследования в области усвоения русского как второго языка и экспериментальные исследования по разным аспектам русской грамматики. RuParam можно рассматривать и как общее параметрическое описание грамматики русского языка.

Массив RuParam устроен как корпус минимальных пар, которых на данный момент насчитывается около 8 тысяч. Каждому грамматичному предложению сопоставлен его неграмматичный аналог, а источник неграмматичности снабжен тегом в результате экспертной разметки, (1). Значительную часть датасета составляют структурированные данные, полученные на основе заданий теста по русскому языку как иностранному (ТРКИ). Мы использовали материалы лексико-грамматического теста, где необходимо выбрать верный способ заполнения пропуска в предложении из нескольких возможных вариантов.

(1) Грамматичное предложение: А кому ты написал это письмо?
Неграмматичное предложение: А кого ты написал это письмо?
Тег: глагольное управление

Кроме этого RuParam содержит данные, полученные из реальных текстов и размеченные по грамматическим параметрам с точки зрения лингвистической теории. В эту часть корпуса вошли случаи, которые важны для теории языка, но при этом редко встречаются в материалах по РКИ, — например, непроективность, связывание анафоров, лицензирование выражений с отрицательной полярностью и др.

RuParam обладает несколькими преимуществами по сравнению с другими параметрическими базами данных. Так, в отличие от BLiMP, он основывается на независимо полученных, а не искусственно сгенерированных для конкретной задачи данных. По сравнению с RuCoLa корпус RuParam не ограничивается нетривиальными и зачастую вариативными феноменами, а включает в себя в основном базовые грамматические параметры. Также наш корпус решает проблему недостатка отрицательного материала — грамматичные и неграмматичные предложения представлены в нем в равном количестве.

В докладе будут представлены состав и структура RuParam и приведены отдельные примеры. Мы также обсудим работы по расширению и улучшению корпуса.

**Литература**

Baker M. 2008. The macroparameter in a microparametric world. In T. Biberauer (ed.), *The limits of syntactic variation*, pp. 351–374. Amsterdam: John Benjamins.
Chomsky N. 1981. *Lectures on government and binding*. Dordrecht: Foris.
Chomsky N. 2005. Three factors in language design, *Linguistic Inquiry*, 36(1), pp. 1–22.
Chomsky N. & Lasnik H. 1993. The theory of principles and parameters. In J. Jacobs, A. von Stechow, W. Sternefeld, & T. Vennemann (eds.) *Syntax: an international handbook of contemporary research*, De Gruyter, Berlin.
Li Y. & Wei W. 2019. Microparameters and language variation, *Glossa: a journal of general linguistics*, 4(1), p. 106.
Mikhailov V., Shamardina T., Ryabinin M., Pestova A., Smurov I., & Artemova E. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5207–5227.

Trotta D., Guarasci R., Leonardelli E., Tonelli S. 2021. Monolingual and Cross-Lingual Acceptability Judgments with the Italian CoLA corpus. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2929–2940.
Warstadt A., Singh A., & Bowman S. R. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641.
Warstadt A., Parrish A., Liu H., Mohananey A., Peng W., Wang Sh., & Bowman S. R. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 377–392.