**Сравнительный анализ NLP-модели и ручной разметки в морфологической декомпозиции слов**

Ходжаметова Милана Илимдаровна

Студентка Национального исследовательского университета «Высшая школа экономики», Нижний Новгород, Россия

Стажер-исследователь Центра Языка и Мозга

В условиях интенсивного развития лингвистических и нейронаучных дисциплин особое внимание уделяется исследованию морфологических структур слов, поскольку морфемы, как минимальные семантико-грамматические единицы языка, играют определяющую роль в формировании лексико-семантических и синтаксических свойств языковых систем, а также в мозговой обработке текстов при чтении. Этот фокус приобретает особую актуальность в анализе флективных языков, таких как русский, отличающихся высокой степенью морфологической сложности благодаря обилию аффиксов, формирующих многообразие производных форм. Новизна обусловлена устранением существующего пробела в изучении нейрональных механизмов морфологической обработки в русском языке, тем самым обогащая теоретическую базу лингвистики и нейрокогнитивных наук. Данное исследование концентрируется на сравнении автоматизированного метода разметки с помощью NLP-модели и ручного подхода для определения морфемной структуры слов. Цель работы — выявить различия между результатами ручной и компьютерной разметки и сопоставить их с данными магнитоэнцефалографии (МЭГ). Предполагается, что морфологическая декомпозиция, учитывающая, корневую продуктивность слова, существенно воздействует на обработку слов в головном мозге, регистрируемую посредством магнитоэнцефалографии (МЭГ). Таким образом, данное исследование стремится внести вклад в понимание взаимосвязей между лингвистическими характеристиками слов и их нейронными коррелятами, что имеет значение для дальнейшего развития вычислительной лингвистики и нейронаук.

Первый этап исследования включал отбор высокочастотных и низкочастотных слов. Морфологическая структура слов определялась с помощью ручной разметки, фиксирующей количество морфем и производных слов, и автоматизированной разметки с использованием Морфессора, который представляет собой NLP-модель, разработанную для неуправляемого обнаружения морфем в текстовых данных. Она основана на принципе минимальной длины описания и использует алгоритм упаковки, который стремится создать оптимально компактный и описательный лексикон единиц, называемых морфами, для описания обучающего корпуса. [Hakala et al.: 2]. Нейрональные корреляты были зафиксированы во время прочтения высокочастотных и низкочастотных слов и проанализированы с помощью магнитоэнцефалографии (МЭГ), регистрирующей поля, связанные с активацией и ее временем. Материалом исследования послужили МЭГ-данные с параметрами амплитуды и времени для двух групп русских слов: высокочастотных и низкочастотных, каждая из которых включала по 65 слов, а также корпус текстов на русском языке для обучения и тестирования NLP-модели. Данный подход обеспечил комплексный анализ морфологических и нейронных характеристик.

В рамках проведённого исследования были получены следующие результаты. Сравнительный анализ ручной и автоматизированной разметки показал расхождения в усредненном количестве морфем: ручная разметка зафиксировала 2,08 морфем для низкочастотных и 2,22 для высокочастотных слов, тогда как Морфессор указал 2,63 и 1,65 соответственно. Статистический t-тест для ручной разметки показал незначимые различия по морфемам (t = -0,66, p = 0,51), но значимые по производным словам (5,55 для низкочастотных и 10,54 для высокочастотных; t = -5,64, p < 0,001), что подчёркивает ключевую роль корневой продуктивности в анализе морфологии русского языка. Анализ МЭГ-данных выявил, что высокочастотные слова демонстрируют более низкую активность и быстрее достигают временного пика (53 мс), чем низкочастотные (72 мс), со статистически значимыми различиями, подтверждающими влияние морфологии на обработку языка. Таким образом, указанные различия отражают иерархическую поморфемную декомпозицию, обеспечивающую более быструю обработку слов за счет частой встречаемости их корней, что не учитывается линейными моделями [Oseki, Marantz: 215].

На основе данных были установлены значительные различия между высокочастотными и низкочастотными словами по параметру производности, что подчёркивает их ключевую роль в морфологии русского языка. В отличие от зарубежных исследований европейских языков, для русского языка важной оказалась корневая продуктивность. Это знание о производных словах позволяет улучшить поморфемное деление и анализ нейронных процессов. В перспективе этот результат позволяет улучшить поморфемное деление в алгоритме NLP-моделей и проведению более точного анализа нейронных процессов.

**Литература**

1. Hakala T. et al. Information properties of morphologically complex words modulate brain activity during word reading // Human Brain Mapping. 2018. № 6. С.2.
2. Oseki Y., Marantz A. Modeling morphological processing in human magnetoencephalography // Society for Computation in Linguistics. 2020. № 1. С.215.