ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

А. А. Баркович

СПЕЦИФИКА ИНДЕКСАЦИИ СЕМАНТИКИ ТЕКСТА

Семантика, пронизывая всю языковую систему, позволяет получить объективное представление о приоритетах использования отдельных языковых единиц в том или ином *контексте*. В свою очередь, единицы, обладающие самостоятельным значением, формируют общую семантическую насыщенность текста. Степень содержательности текста зависит от дифференцированных данных, характерных речи и прямо представленных реализациями отдельной леммы и совокупности лемм. Впрочем, не менее существенна семантика контекста, ввиду сложности изучения которой задачи ее метаописания пока чаще формулируются, чем решаются на практике.

Одним из показателей, которые могут повысить лингвистическую убедительность результатов, полученных в речевой среде средствами прикладного инструментария, является цифровое представление насыщенности текстов кластеризованной и стратифицированной семантикой. Такую информацию относительно лексического уровня семантики КОД в корпусном формате можно получать, например, с помощью референтных статистических моделей.

Индекс семантичности (англ. *Semanticity Index*), или I_S (количество словоупотребления текста/корпуса текстов, разделенное на количество словоупотребления отдельной леммы) позволяет устанавливать степень насыщенности текстов отдельными леммами и их группами, семантическими категориями. Например, среди двадцати самых частотных слов английского языка – только три глагола: *be* – ‘быть’, *have* – ‘иметь’ и *do* – ‘делать’ (см. British National Corpus). Соответственно, *индекс семантичности* в первой двадцатке наиболее частотных слов British National Corpus (для 110 691 482 словоупотребления) будет колебаться от $I_S \approx 17{,}89$ для *the* (частотность – 6 187 267) – до $I_S \approx 214{,}03$ для *by* (частотность, соответственно – 517 171).

Выполнение подобных подсчетов вполне релевантно и для других языков – после создания репрезентативных текстовых массивов в электронном формате и доступа к исходным кодам разметки, если соответствующая спецификация будет отсутствовать по умолчанию.

Относительно употребления белорусских лексем статистические данные подтверждают общую тенденцию по межчастеречным отношениям в тексте. Согласно данным ресурса *Беларускі N-корпус* здесь с убедительной стабильностью преобладают служебные части речи. Так, четыре первые строки, представленные *у*, *і*, *на*, *з*, принадлежат предлогам и союзам, а три первые, кстати, соответствуют и русскоязычным вариантам (см. Беларускі N-корпус; МГУ–ЛОКЛЛ). Своеобразные маркеры устной речи местоимения *я*, *мы*, *ён* (рус. *он*) в содержащем существенный объем близкой к устной речи публи-

цистики (30 852 670 из 44 014 121 словоупотреблений) – здесь присутствуют в первой двадцатке: *я* – 16 позиция 44 014 121 единиц, *мы* – 17 позиция и 328 822 единиц, *ён* – 20 позиция и 292 791 единиц (см. Беларускі N-корпус).

Доступные для анализа данные относительно частеречного состава ресурса *Беларускі N-корпус* требуют дальнейшей обработки по причине «неснятой омонимии» в корпусе: например, определить пропорциональное соотношение союза *што* и местоимения *што* (рус. *что*) сегодня можно только приблизительно: контексты корпусного конкорданса свидетельствуют об относительно небольшом количественном преобладании союза *што* над местоимением *што*:

Ён пабачыў, што я хачу падняцца, а не падымуся;

Мы пасталі: думалі спачатку – немцы, зірнулі – пазналі, што гэта ідуць партызаны;

Тое, што адбывалася з Вялікімі Прусамі, дзве гэтыя жанчыны бачылі, перажылі кожная па-свойму;

Пытанне: – А пра што яны ўвечары гаварылі з людзьмі? (Алесь Адамовіч, Янка Брыль, Уладзімір Калеснік. Я з вогненнай вёскі, 1975).

Таким образом, приведенные примеры корпусный менеджер ресурса *Беларускі N-корпус* сопроводил, к сожалению, недостаточно дифференцированным в отношении частеречной принадлежности единиц метаописанием. Аналогично, зачастую, организовано метаязыковое сопровождение текстов в большинстве лингвистических корпусов, при этом, например, в широко известном НКРЯ омонимия снята только для сравнительно незначительного объема (в сумме около 6 млн словоупотреблений) текстов – работа выполнена сотрудниками проекта «вручную» (см. Национальный корпус русского языка). Таким образом, качество исследования КОД пока в значительной степени обусловлено задействованностью в пред- и постобработке текстов человека.

Тем не менее, *индекс семантичности*, представляя базовую информацию о конкретном параметре семантической идентичности текста, позволяет проводить достаточно широкие обобщения. Практика свидетельствует, что программными средствами можно обеспечить выполнение целого ряда алгоритмов лингвистической обработки текстов. Разработка узкоспециальных изначально задач в прикладном ключе часто характеризуется перспективностью, форматностью и возможностями включения в комплексные исследования.

Н. А. Богданова

ФОРМАЛИЗОВАННОЕ ПРЕДСТАВЛЕНИЕ ВНУТРИЖАНРОВОГО ТЕКСТОВОГО СХОДСТВА НА ЛОГИКО-СЕМАНТИЧЕСКОМ УРОВНЕ

Формализация логико-семантической организации текста с последующим сопоставлением полученных моделей опирается на следующие исходные предпосылки.