Е.В. Ягунова

ИССЛЕДОВАНИЕ КОНТЕКСТНОЙ ПРЕДСКАЗУЕМОСТИ ЕДИНИЦ ТЕКСТА С ПОМОЩЬЮ КОРПУСНЫХ РЕСУРСОВ[1]

Во время коммуникативного акта человек непрерывно планирует (программирует) свою речевую деятельность, осуществляя необходимые регулировки, переключения и т.д. С этой точки зрения, каждое следующее слово должно быть каким-то образом «сверено» и согласовано с тем, что уже воспринято или произнесено к текущему моменту. В большей степени нас будут интересовать процедуры контекстной предсказуемости в рамках восприятия текста (речи), в меньшей степени мы обращаемся к данным порождения текста. По-видимому, в дополнение к другим структурным характеристикам каждая структурная составляющая текста может быть описана через распределение вероятностей (силы влияния) разных позиций, которые способствуют (или не способствуют) адекватному восприятию соответствующих единиц текста. Природа такого влияния позиций может быть различного происхождения: некоторые факторы будут непосредственно связаны с лексической и/или семантической сочетаемостью/несочетаемостью, в то время как другие будут определяться правилами синтаксической организации клаузы, фразы или даже всего текста. В каждом случае вероятности не равны и зависят от множества разных факторов, которым приписываются разные веса.

По-видимому, минимальное «окно сверки» равно одной единице (в типичном случае одному слову), максимальное «окно сверки» полагаем равным тексту. Необходимо оговорить условность соположения в рассматриваемом смысле минимального и

[1] Работа выполнена при частичной финансовой поддержке гранта РГНФ (проект номер 07-04-00161а).

максимального окон сверки с точки зрения механизмов восприятия. Сравнительно небольшие окна сверки (минимальное или равное клаузе) являются «нормальными», так как не противоречат ограничениям на оперативную память человека. Текст, как правило, превышает объем оперативной памяти человека. Можно предположить функционирование иных механизмов контекстной предсказуемости в рамках текста целиком: подстройки к его особенностям и его понимания[1].

В данном докладе обозначим контуры парадигмы исследования механизмов контекстной предсказуемости с использованием корпусных ресурсов, доступных любому исследователю. Основной акцент делается на сосуществование процедур контекстной предсказуемости на разных «окнах сверки».

Сравнительно небольшие «окна сверки» – от контактного сочетания слов до фразы, по-видимому, в наибольшей степени соответствуют традиционным корпусным методам анализа сочетаемости единиц текста[2]. Прежде всего, эта проблема выводит нас на решение вопросов о коллокациях. К сожалению, большин-

[1] См. подробнее: *Ягунова Е.В.* Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь, 2008.

[2] Ср. «В естественном языке существуют особые механизмы, которые регулируют комбинаторику лексических единиц текста на формальном и содержательном уровнях. Данные механизмы необходимо учитывать при моделировании понимания текста, в связи с этим функционирование сочетаний слов, тяготеющих к совместному употреблению, является объектом пристального внимания учёных в аспекте их статистической устойчивости, формальной и семантической связанности» (*Митрофанова О.А., Белик В.В., Кадина В.В.* Корпусное исследование сочетаемостных предпочтений частотных лексем русского языка // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Периодическое издание. Выпуск 7 (14). 2008. С. 361.).

ство сервисов анализа биграмм (контактно встречающихся в тексте коллокатов) ориентировано исключительно на такие единицы как лексемы[1]. В результате для флективного русского языка часто не учитываются явления минимального синтаксиса.

Множество коллокаций очевидным образом неоднородно с фонетической, синтаксической и семантической точек зрения. По-видимому, наибольшей силой контекстной предсказуемости характеризуются компоненты коллокаций, характеризующихся максимальной целостностью и являющихся единицами (ментального) словаря[2] (инвентарными единицами языка по В.Б. Касевичу[3]). Можно предложить выделение таких подклассов, например, как предложно-падежные конструкции, составные слова (например, *друг друга*, *может быть*), сложные номинации и, как правило, устойчивые словосочетания. Процедуры контекстной предсказуемости в рассматриваемых «окнах сверки» (от словосочетания до фразы) зависят от типа рассматриваемых единиц (коллокаций)[4].

Кроме «полноценных» сложных номинаций в ряде случаев возможно выделять промежуточные элементы, например, представляющие собой воспроизводимые атрибутивные сочетания: например, *старый друг*, *богатый человек*, *старый человек*,

---

[1] См., например: *Аверин А.Н.* Разработка сервиса поиска биграмм // Труды Международной конференции «Корпусная лингвистика–2006». СПб., 2006. С. 5–15.

[2] См. подробнее: *Ягунова Е.В.* Вариативность стратегий восприятия…; *Ягунова Е.В.* Неоднословные целостности в словаре и корпусе // Труды Международной конференции «Корпусная лингвистика–2006». СПб., 2006. С. 395–412.

[3] *Касевич В.Б.* Семантика. Синтаксис. Морфология. М., 1988.

[4] См., например: *Ягунова Е.В.* Вариативность стратегий восприятия…; *Ягунова Е.В.* Неоднословные целостности в словаре и корпусе ….

здоровый человек, *хорошо одетый мужчина* и многие другие[1]. Эти воспроизводимые и высокочастотные по данным Национального корпуса русского языка (НКРЯ, **www.ruscorpora.ru**) атрибутивные сочетания скорее всего будут рассматриваться как целостные единицы. Для статистической характеристики коллокаций-инвентарных единиц в большинстве случаев, вероятно, достаточно частоты встречаемости этих коллокаций (т.е., например, вполне достаточно сервиса НКРЯ).

Наряду с инвентарными единицами в процессе речевой деятельности активно используются такие единицы, которые конструируются по заданным правилам в процессе речевой деятельности (конструктивные единицы языка по В.Б. Касевичу)[2]. Статистические характеристики такого рода единиц текста (конструкций и/или коллокаций) в существенной степени определяют работу процедур контекстной предсказуемости. Для исследования коллокаций – конструктивных единиц, вероятно, не менее важным показателем является коэффициент взаимной информации *MI*[3] как мера ассоциативной связи между коллокатами.

---

[1] См. *Воейкова М.Д.* Выражение квалитативной характеристики человека в русском и немецком языке // Проблемы функциональной грамматики. Полевые структуры. СПб., Наука, 2005.

[2] См. *Касевич В.Б.* Семантика. Синтаксис. Морфология. М., 1988. Естественно, невозможно провести четкое разграничение между классами коллокаций, относящимися к инвентарным и конструктивным единицам. Так, в зависимости от предметной области и функционального стиля текста одно и то же словосочетание может, вероятно, рассматриваться или как конструктивная, или как инвентарная единица (высокочастотное или даже терминологическое словосочетание).

[3] При извлечении биграмм из корпуса текстов *MI* позволяет выявить наибольшее число сочетаний, зарегистрированных в лексикографических источниках; доля биграмм со знаками пунктуации в экспериментах с *MI* оказывается существенно ниже, чем при использовании других мер, в частности, *T* и *Log-Likelihood* (*Khokhlova M.* Collocations in

Коэффициент *MI* позволяет оценивать силу ассоциативной связи внутри сочетания слов на основе соотношения частоты встречаемости биграммы и независимых употреблений коллокатов, с учётом объема корпуса.

Чем выше эта характеристика, тем выше должна быть сила контекстной предсказуемости синтагматического соседа в речевой деятельности.

Однако понятие «синтагматический сосед» требует уточнения; прежде всего, с точки зрения того, какая единица – словоформа или лемма – рассматривается в качестве коллоката. По-видимому, это уточнение касается основополагающих функциональных вопросов:

- ✓ **вида речевой деятельности**: восприятие / порождение,
- ✓ **формы речи или модальности восприятия**: устная / письменная,
- ✓ **этапа** или **уровня** восприятия или порождения речи,
- ✓ **статуса единицы**: единица словаря или «только» оперативная единица.

В случае восприятия текста и при опоре на более поверхностный уровень наибольший вес, вероятно, имеет максимально грамматически оформленная синтагматичность и, соответственно, такой коллокат как словоформа. При более глубинном уровне обработки вероятностный прогноз – в случае как восприятия, так и порождения речи – может осуществляться с опорой на лексем-

---

Russian: Analysis of Association Measures // Computer Treatment of Slavic and East European Languages: 4th International Seminar. Bratislava, 2007. P. 96–103). Ср., например, то, что в сервисе поиска биграмм на АОТ (**www.aot.ru**) высчитывается (и используется для сортировки) только этот коэффициент (*Аверин А.Н*. Разработка сервиса поиска биграмм…). Однако в ряде случаев приходится использовать другие критерии. Так, например, при выделении биграмм с высокочастотными коллокатами (например, *может быть*) может потребоваться использование логарифма правдоподобия (*Log-Likelihood*).

ные варианты коллокатов. Таким образом, при решении разных задач контекстной предсказуемости оказывается важным сопоставлять данные по сочетаниям как словоформ, так и лексем. Более того, в большом числе случаев необходимо сопоставлять эти данные. При решении такого рода задач может помочь, например, открытый ресурс, созданный в университете Лидса (Великобритания) под руководством С.А. Шарова, с помощью корпус-менеджера CQP[1]. Этот ресурс позволяет получать данные как по словоформам, так и лексемам.

Дополнительного уточнения требует рассмотрение противопоставлений (1) «правый vs. левый сосед» и (2) «контактный vs. дистантный сосед». Причем это уточнение необходимо не только при исследовании конструктивных единиц, но и коллокаций, соответствующих таким инвентарным единицам языка как составное слово и устойчивое словосочетание.

В случае с инвентарными единицами речь идет о таких вариантах как, например, *друг дру+га* : *дру+г__за__дру+га* (разрывные составные слова) и *мо+жет_быть* : *быть_мо+жет* (составные слова, различающиеся порядком следования компонентов); или фразеологизм, который может различаться (1) разрывностью, (2) порядком следования компонентов, (3) грамматической формой основного компонента, например, *дать дуба* : *дуба дать* : *дуба не дал* и т.д.

Рассмотрение этих противопоставлений невозможно вне выделения таких типов контекстной предсказуемости как «линейное vs. нелинейное предсказание». Например, восприятие письменного текста представляет почти в чистом виде нелинейный тип предсказания, в том смысле, что испытуемые одновременно видят слова, окружающие рассматриваемое, текущее слово, т.е. «окно сверки» может включать не только левый (в большей степени), но и правый контексты. В случае восприятия звучащего

[1] URL: **http://corpus1.leeds.ac.uk/ruscorpora.html**

текста левый и правый контексты становятся неравноценными: «окно сверки» включает главным образом левый контекст[1].

Таким образом, при исследовании контекстной предсказуемости синтагматического соседа в составе свободного словосочетания необходимо учитывать все названные параметры и противопоставления. К сожалению, пока не существует единого открытого корпусного ресурса, в рамках которого возможно получение такого рода «всесторонних» данных. Однако, сопоставляя и грамотно анализируя данные, полученные с помощью разных ресурсов, можно решить многие задачи такого рода исследований.

Отдельную проблему составляет исследование контекстной предсказуемости в максимальном «окне сверки», равном тексту, и в минимальном – равном слову. Для решения этой задачи необходимо учитывать такие параметры как функциональный стиль (или жанр) и предметная область текста, т.е. рассматривать распределение частот не только (не столько) в «генеральном» корпусе текстов, но в корпусе текстов одного функционального стиля (подкорпусе), так как человек **заранее** выбирает «подсловарь», заданный условиями коммуникации (например, подкорпусе и, соответственно, подсловаре публицистики или бытового диалога). Отличия между подсловарями могут состоять как в несовпадении состава единиц, так и в разных индексах частотности для одних и тех же единиц. Кроме того – человек выбирает «текущий словарь» по мере подстройки в ходе коммуникации под особенности

[1] Подробнее см., например: *Ягунова Е.В.* Роль ключевых слов при восприятии звучащего и письменного текста (на материале русского языка) // Человек пишущий и читающий: проблемы и наблюдения: Материалы и наблюдения: Материалы международной конференции 14–16 марта 2002 г., Санкт-Петербург. СПб., Издательство СПбГУ, 2004. С. 197–204; *Касевич В.Б., Ягунова Е.В.* Контекстная предсказуе¬мость слов в тексте (на материале русского и французского языков) // Вестник Пермского университета. Вып. 3. 2006.

конкретного текста (что особенно характерно для восприятия текста). Состав единиц и индексы частотности «текущего словаря» соответствуют тому, что можно назвать частотностью по тексту[1]. Таким образом, для проведения такого рода исследований необходимо соотносить как минимум три вида частотных характеристик: общеязыковая частота встречаемости (по «генеральному» корпусу), частота встречаемости по подкорпусу, задаваемому функциональным стилем и предметной областью, и частотность единиц в пределах определенного текста[2]. До некоторой степени решение такого рода задач возможно с помощью ресурсов на основе НКРЯ (объем и возможность задавать жанровые подкорпуса на **www.ruscorpora.ru**), а главное – ведущимися работами по созданию прототипов частотных списков по жанрам[3], а в дальнейшем и частотного словаря русского языка[4].

В заключение подчеркнем, что для проведения комплексного исследования процедур контекстной предсказуемости необходимо использование **разных** методов анализа корпусных ресурсов. И в настоящее время появляется все большее число подобных возможностей.

---

[1] Подробнее см.: *Ягунова Е.В.* Вариативность стратегий восприятия…

[2] Ср. методы статистического определения ключевых слов, используемые, например, при решении задач автоматического поиска документов.

[3] На **http://corpus.leeds.ac.uk/serge/frqlist/**. Так, «Частотный словарь административных текстов, значимая лексика» был использован при отнесении лексики к деловому функциональному стилю (подробнее см. *Ягунова Е.В.* Вариативность стратегий восприятия…).

[4] См. *Ляшевская О.Н., Шаров С.А.* Частотный словарь Национального корпуса русского языка: концепция и технология создания // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Периодическое издание. Выпуск 7 (14). 2008.