Лексический аспект снятия омонимии в морфологически аннотированном корпусе (на примере пулар)[1]

Язык, обладающий электронным корпусом текстов, становится более доступным для исследования специалистами разных областей. Согласно определению [Sinclair 2004: 24], «Корпус – это собрание электронных текстов, отобранных по определённым критериям с целью отобразить, насколько это возможно, язык и его вариативность и послужить источником информации для лингвистических исследований». В случае если корпус морфологически аннотирован хотя бы частично, его ценность возрастает в разы, поскольку языковые данные тем самым становятся доступны не только специалистам по языку, но и типологам, не владеющим языком, но желающим использовать материал. Однако для создания морфологически аннотированного корпуса требуется, естественно, гораздо больше ресурсов, чем для неаннотирвоанного, особенно в части адаптации морфофонологии и грамматики под общепринятые конвенции [Косогорова 2013] (самыми популярными являются общеизвестные Лейпцигские правила глоссирования), а также снятия омонимии.

Носители естественного языка постоянно сталкиваются со случаями языковой омонимии, как грамматической, так и лексической. В случае грамматической омонимии парсер (программа, создающая автоматическую морфологическую разметку) может быть запрограммирован на жёсткий алгоритм снятия омонимии, связанный с соседними морфемами или лексемами.

Так, например, в языке пулар (фула < атлантическая семья < макросемья Нигер-Конго) морфема -ii имеет несколько значений (см. пример 1). В случае если она присоединяется к существительному или согласуемому адъективу, она обозначает минимальную ступень показателя класса NGII. В случае если она встречается в глагольной словоформе, она может обозначать либо показатель активного залога сильного перфективного вида, либо показатель медиального залога слабого перфективного вида. В этом случае различение происходит путём а) проверки лексемы на переходность, если лексема изначально двухвалентная, но имеет один актант, вместо ожидаемых двух, то она стоит в медиальном залоге и б) в случае если лексема изначально одновалентная, проверки её на «исконную медиальность», то есть, на принадлежность к списку исходно медиальных лексем, которые не могут принадлежать ни к какому другому залогу вне зависимости от контекста и актантной ситуации. Обе этих проверки вполне выполнимы автоматически, хотя и не являются, строго говоря, показателем наличия синтаксической разметки [McEnery, Wilson 1997: 6]. И наконец, в случае, если морфема -ii встречается не в финальной позиции, она обозначает медиальный перфектив в причастии. Особая элегантность системы заключается в том, что причастия требуют полной ступени показателя класса, что сводит омонимию классного и видо-залогового аффикса внутри одного лексического разряда к нулю. Подробнее кросспарадигмальной омонимии в предикате пулар и методиках её разрешения см. [Косогорова 2015].

[1] Работа подготовлена при поддержке проекта РНФ #17-78-20071

(1)[2]

a) Показатель именного класса NGII

baal-**ii**	goot-**ii**
овца-sgNGII	один-sgNGII

‘одна овца’

b) Предикат: Активный залог сильного перфектива

jooni	baal-i	lann-**ii**
теперь	овца-plDI	заканчиваться-ACT.PFV.s

‘Теперь овцы кончились’.

c) Предикат: Медиальный залог слабого перфектива (переходная форма с актантной деривацией)

sari	suuɗ-**ii**
заяц	прятать-MD.PFV.w

‘Заяц спрятался’.

d) Причастие: медиально-перфективная форма

debb-o	hedd-**ii**-ɗo	on	yiɗ-aa	mo
женщина-sgO	оставаться-MD.PFV-sgO	DEF.sgO	любить-ACT.ST.NEG	DO.3.sgO

‘Оставшаяся женщина не любила его’.

Пример (2) демонстрирует другую технику различения грамматических омонимов: морфема -(i)d- может встречаться только в глаголах и причастиях в позиции словообразовательного аффикса, и при наличии зависимого от лексемы дополнения с предлогом e она выражает значение социатива (а также комитатива, см. об этом подробнее [Kosogorova 2019]). При отсутствии такого дополнения морфема выражает значение терминатива.

(2)

a) Социатив

maw-ɗo	won-**d**-i	e	pay-koy
быть.старым-sgO	быть-SOC-ACT.PFV.w	Prep	ребёнок-plKOY

‘Старик жил с детьми.’

b) Терминатив

gell-al	ngal	no	ƴoƴ-**id**-i
куропатка-sgNGAL	DEF.sgNGAL	COP	быть.хитрым-TERM-ACT.PFV

‘Куропатка была хитрой’.

[2] Здесь и далее текстовые примеры взяты из Корпуса пулар (http://corpuspulaar.somee.com/Search.aspx).

Снятие лексической омонимии – это намного более кропотливый процесс, в частности потому, что он не всегда поддаётся автоматизации из-за отсутствия чёткого контекстного решения. Справедливости ради, отметим, что случаи, когда лексическая омонимия снимается с помощью контекста, существуют, но они крайне редки: см. пример (3), в котором глагол *accude* может иметь первым значением ‘покидать, оставлять’ и быть при этом двухвалентным, а вторым - ‘позволять’, и быть при этом трёхвалентным.

(3)

a)

na'-i	ko	neene	makko	acc-id-i	mo	kon
корова-plDI	REL	мать.sgO	IO.3.sgO	оставлять-SOC-ACT.PFV	DO.3.sgO	DEF

‘Коровы, что мать его оставила ему’.

b)

ɓe	inn-i	yo	o	acc-u	ɓe
3.plBE	говорить-ACT.PFV.w	PART	3.sgO	позволять-ACT.OPT	DO.3.plBE

waal-a		ka	makko
проводить.ночь-ACT.POT.w		PREP	IO.3.sgO

‘Они сказали, чтобы она позволила им переночевать у неё’.

Прочие же случаи, не поддающиеся алгортмизации, требуют ручного подхода с применением лексикостатистических методов.

Первое, наиболее базовое решение, свойственное небольшим корпусам на начальном этапе – это фиксация в словнике корпуса наиболее частотного или первого словарного значения лексемы. Так, при принятии такого решения, словарная статья

> **immoo** 1) подниматься, вставать 2) +inf. предпринимать что-л. 3) подниматься, начинаться (о ветре) 4) вспыхивать, начинаться (о ссоре) 5) уезжать из; ~**e wuro** уехать из деревни 6) поправляться, выздоравливать
>
> [Зубко 1980, 43]

будет сведена к соответствию ‘*immoo* – подниматься’, которое в данном случае вполне способно заместить статью, но в случае более обширного семантического поля такое решение вызовет затруднения. Достоинством такого метода является простая автоматизация перевода. При использовании первого словарного значения лексема отображается с помощью эквивалента, наиболее близкого к центру семантического поля, что в принципе позволяет достроить более периферическое значение по контексту самостоятельно (хотя и не всегда, к сожалению).

Такое решение, конечно, годится в качестве базового, но у него есть ряд очевидных недочётов. Во-первых, оно не вполне отражает разнообразие значений лексемы, а некоторые из них могут иметь достаточно широкий разброс и фразеологию, чтобы сборка значений в литературный перевод стала неочевидной, как в (4). Во-вторых, такое решение

никак не решает проблему истинных омонимов, которые не укладываются в единое лексическое гнездо в силу различного происхождения, случайного совпадения или значительной диахронической отдалённости, как продемонстрировано в (5). Тогда в любом случае должно применяться следующее решение.

(4)

> **itta** 1) убирать, устранять; ~ **bote** находить недостатки; критиковать; осуждать; ~ **e cikke** оставлять надежды; ~ **ɗonka** утолять жажду; ~ **giiɗe** ослабить бдительность; ~ **hersa** а) преодолевать стыд б) смывать позор; ~ **hoore mun** а) устраняться, удаляться б) освобождаться в) погибать; ~ **kufune** снимать головной убор; ~ **kooyɓe** завтракать; ~ **noone aaden** лишать индивидуальности; ~ **nulal** давать поручение, посылать с по-ручением; ~ **paɗe** разуваться, снимать обувь; ~ **palje** устранять недостатки; ~ **wonkii** убивать кого-л.; ~ **wudere** снимать пань; ~ **yimɓe** поставлять людей (напр. для ведения войны); ~ **yoomere** переставать грустить; ~ (**neɗɗo**) **e** забирать или освобождать кого-л. откуда-л.; ~ (**neɗɗo**) **e karhankaaku** освобождать кого-л. от рабства; ~ (**neɗɗo**) **e laawol** убирать кого-л. с дороги, заставлять уйти со своего пути; ~ (**neɗɗo**) **e limoore** вычеркивать из списка 2) собирать урожай (фруктов) 3) рвать (цветы) 4) увеличиваться в объеме или количестве 5) платить, выплачивать; ~ **kaalisi** платить деньги; ~ **sagalle**, **jonna neɗɗo** платить налог кому-л.; ~ **sadaka** приносить жертву 6) вновь привозить или приносить из

[Зубко 1980, 45]

(5) истинные омонимы

> **inna** I 1) говорить; *innu coggu maa!* назови свою цену! 2) называть, присваивать имя; произносить имя; *o inni mo* он назвал его, он дал ему имя; ~ **hoore mun** считать себя
>
> **inna** II O [-'en] f. courte от **inniraawo**, **inniraaɗo** [мать]

[Зубко 1980, 44]

Второй способ снятия лексической омонимии является наиболее трудозатратным и поэтому наименее подходящим для автоматизированного корпуса. Он заключается в том, чтобы ввести в словарь все значения лексемы и её возможных омонимов, а затем в каждом случае выбирать необходимое значение вручную. Так, парсер, настроенный на разбор и аннотирование текстов пулар, предлагает при таком решении следующие варианты для лексемы *itta*:

itta

> > убирать
> > собирать.урожай
> > рвать
> > увеличиваться
> > платить
> > вновь.привозить

Однако точность этого способа позволяет создать семантически адекватный перевод уже на стадии глоссирования, что позволяет исследователям, не знакомым с языком, значительно лучше ориентироваться в тексте и значительно точнее подбирать примеры.

Третий способ логически вытекает из второго: если имеется несколько вариантов значения лексемы, и при этом не имеется контекстной возможности их различения, то для снятия омонимии можно привлечь лексикостатистические методы. Здесь можно использовать две возможности.

Если в наличии имеется адекватный по качеству словарь, перечисляющий все значения конкретной лексемы в порядке их удаления от центра семантического поля, можно использовать этот порядок. Тогда при разборе программа будет автоматически присваивать наиболее семантически центральное значение лексемы, а затем оператор будет проверять полученное значение на соответствие, и при несовпадении выбирать следующее по списку. Отличие данного метода от второго способа снятия омонимии (ручного) заключается в том, что варианты, предлагаемые оператору, уже ранжированы по степени удалённости от центра семантического поля (от более вероятно подходящего к менее), что, по опыту, значительно сокращает усилия, и даже в случае полного отсутствия ручной адаптации позволяет получить «сырую» первую модель снятия омонимии. Так, для лексемы *itta* ранжированный список будет выглядеть следующим образом (цифры – процент выбора оператором конкретного значения – высчитываются автоматически):

itta

42.253> убирать
33.860> собирать.урожай
15.002> рвать
5.941> увеличиваться
1.612> платить
1.332> вновь.привозить

При отсутствии достоверного порядка, словарную статью можно использовать хотя бы как набор значений лексемы. Тогда программа-парсер должна быть «обучаемой», то есть, она должна записывать количество выборов, совершённых в пользу того или иного значения лексемы, и в соответствии с этой статистикой создавать ранжирование значений, которое затем будет использовано с помощью метода, описанного выше. Разумеется, на период обучения снятие омонимии сводится к ручному, однако для языков без заранее имеющихся словарей, а только с лексиконом, этот вариант является приоритетным.

В случае если язык совершенно не описан, программа-парсер может помочь с созданием лексикона, тогда к её задачам добавляется создание лексикона. При отсутствии адекватного значения лексемы, оператор может добавить его в список, а дальше сбор статистики будет осуществляться в обычном режиме.

Таким образом, при максимальной степени автоматизации для небольших корпусов можно получить статистически ранжированный словник, автоматически предлагающий первое словарное значение (и определяющий его при необходимости). Затем из него в полуавтоматическом режиме можно получить следующие значения для выбранной лексемы.

К числу достоинств такого словника можно, без сомнения, отнести возможность его простой трансформации в словарную статью, поскольку практически все данные, которые требуются для этого (частеречные пометы, фразеология, лексические значения в порядке убывания), уже записаны в словник при парсере. Их нужно только организовать и оформить в установленном порядке, с чем справится простейший скрипт.

Таким образом, процесс снятия лексической омонимии для некрупных корпусов с поморфемным аннотированием является, бесспорно, сложным в реализации. Но, хотя для небольших по объёму корпусов, для которых неприменимы методы комплексного статистического обучения (используемые во многих многомиллионных корпусах), полная автоматизация для снятия лексической омонимии невозможна, существуют решения, достаточно упрощающие работу оператора и имеющие к тому же побочный эффект структуризации словника, в частности, с целью создания печатного словаря.

Литература:

Зубко Г.В. Фула-русско-французский словарь. М.: Русский язык, 1980.

Коваль А.И., Косогорова М.А. К проблемным вопросам морфоглоссирования текстов пулар-фульфульде // Вопросы филологии №2 (41). М.:Gaudeamus 2012. Сс. 30-48.

Корпусные исследования по русской грамматике. Ред.: К.Л. Киселева, В.А. Плунгян, Е.В. Рахилина, С.Г. Татевосов. М.: Пробел-2000, 2009.

Косогорова М.А. Опыт подготовки текстов пулар к автоматической разметке: проблемы и перспективы // Исследования по языкам Африки, выпуск 4. М.: ИД Ключ-С, 2013. Сс.108-121.

Косогорова М.А. Кросспарадигмальная омонимия глагола пулар: типы методики разрешения // Исследования по языкам Африки. Т. 5. ИД Ключ-С Москва, 2015. Сс.155–164.

Biber D., Conrad S., Reppen R.. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998.

Kennedy G. D. An Introduction to Corpus Linguistics. London: Edison Wesley Longman Limited, 1998.

Kosogorova, M. Functions of prepositions in Pular: e and ka // Studia Linguistica 2019. In print.

Corpus-Based Perspectives in Linguistics. Kawaguchi, Yugi et al., ed. Tokyo: Joy Benja-mins Publishing Company, 2007.

McEnery T., Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1997.

Sinclair J. Corpus Creation // Corpus Linguistics: Readings in a Widening Discipline. Sampson G. & McCarthy D. (eds). New York: Continuum. 2004.