Секция 1
ГРАММАТИКА И ПРИКЛАДНАЯ ЛИНГВИСТИКА

О.И. Бабина
г. Челябинск

АВТОМАТИЧЕСКАЯ ПАРАДИГМАТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ СЛОВОФОРМ НА ОСНОВЕ АНАЛИЗА ИХ МОРФОЛОГИЧЕСКОЙ КВАЗИСТРУКТУРЫ

В современных условиях необходимости обработки огромных массивов текстовой информации остро стоят проблемы разработки приложений инженерной лингвистике. Актуальным сегодня является корпусный подход, при котором для построения лингвистической базы знаний используются данные большого массива текстов. Корпус текстов используется для построения лексической базы знаний – лексикона. В лексиконе лексические единицы, как правило, организованы в лексико-грамматические парадигмы. Определение парадигм вручную является задачей трудоемкой и ресурсоемкой. В связи с этим представляется необходимым автоматизировать парадигматическую идентификацию словоформ, то есть автоматизировать определение грамматических (а также лексико-семантических) характеристик словоформ, встречающихся в корпусе текстов.

В современной инженерной морфологии превалирует аддитивный подход квазиагглютинативной аппроксимации. Суть его заключается в том, что строение словоформы представляется через последовательность сегментов, практически не подверженных варьированию. Морфологическая модель, построенная согласно этому подходу, использует при синтезе словоформ списки квазиморфем (квазиоснов и квазисуффиксов), которые, будучи связаны отношением порядка, присоединяются друг другу по заданным лингвистом правилам, разрешающим или запрещающим соединение в цепочку классов квазиморфем.

Так, например, некоторая морфологическая модель такого типа может содержать квазиосновы A = [*bas-, cris-, analys-*] (квазиморфемы нулевого порядка), нулевой квазисуффикс (квазиморфема первого порядка) и квазисуффиксы *-is, -es* (квазиморфемы второго порядка), последним приписываются грамматические характеристики s = [ед.ч.] и p = [мн.ч.] соответственно. Причем имеется формообразующее правило, согласно которому квазиосновы из списка A сочетаются с нулевой квазиморфемой первого порядка (с грамматической пометкой N=[сущ.]) и указанными квазиморфемами второго порядка, приобретая соответствующие им грамматические характеристики. В результате применения этого правила в данной морфологической модели возможен синтез английских слов *basis, crisis, analysis* и соответствующих им форм множественного числа:

bas- + Ø [сущ.] + *-es* [мн.ч.] → *bases*~Np

Таким образом, словоформа представляет собой последовательность присоединяемых друг к другу сегментов, каждый из которых неизменен и словоформа в целом, принимая грамматические характеристики, приписываемые составляющим ее сегментам, грамматически оформлена.

Такой подход может быть применен в инженерной лингвистике также при решении задачи парадигматической идентификации словоформ. Рассматривая парадигматическую идентификацию в контексте проблемы составления лексиконов, лежащих в основе приложений инженерной лингвистики, мы предлагаем использовать аддитивный подход на частично неизвестных данных. Идея метода заключается в том, чтобы, имея на входе список словоформ из корпуса текста, создать разбиение, определяя по квазисуффиксам лексико-грамматическую принадлежность для слов, образованных в результате деривации и подвергающихся флективным изменениям при словоизменении. Определение лексико-грамматической принадлежности словоформы осуществляется на основе частичной информации о ее структуре (наличии квазисуффиксов в плане выражения словоформы).

Общая схема метода представлена на рис. 1. Лингвистическая база знаний, необходимая для осуществления процедуры парадигматической идентификации, включает:

1) набор квазисуффиксов первого порядка, которые представляют собой сегменты, следующие непосредственно за квазиосновой и условно соответствующие формантной части в словообразовательном процессе. Каждому квазисуффиксу приписывается следующая лексико-грамматическая информация: а) часть речи и определяемые частью речи грамматические категории; б) номер парадигм (списков квазисуффиксов второго порядка, сочетающихся с данным квазисуффиксом); в) номер списка правил соединения с квазисуффиксами второго порядка; г) семантический класс; д) (факультативно) номер правил синтеза словоформы, мотивирующей образование деривата с данным квазисуффиксом. Некоторые квазисуффиксы могут не предполагать синтеза мотивирующей словоформы. Например, русский квазисуффикс *–м,* которому приписана следующая информация: а) [сущ., ср.р.]; б) ссылка на список квазисуффиксов [*-я* [Им.п., ед.ч.], *-ени* [Р.п., ед.ч.], *-ени* [Д.п., ед.ч.], и т.д.] и т.д. отыскивает формы таких слов как *время, племя, знамя* и т.д., которые в современном состоянии языка вряд ли мыслятся как дериваты от каких-либо других слов;

2) набор списков квазисуффиксов второго порядка, приблизительно соответствуют флексиям. Каждому квазисуффиксу второго порядка приписывается информация о морфологической форме, соответствующей данному квазисуффиксу.

3) Набор правил соединения квазисуффиксов первого и второго порядка. Так, например, для итальянского квазисуффикса *-c* с грамматиче-

ской информацией: а) [глаг.]; б) ссылка на список квазисуффиксов второго порядка [*-are,-o,-i,-a,-iamo,-ate,-ano,* и т.д.], правило для образования формы настоящего времени, 1 л., мн. ч. имеет вид (пятый квазисуффикс в списке):

$$5: Suf^1 + Suf^2 \rightarrow Suf^1 + \text{‘h’} + Suf^2$$

Согласно этому правилу, при присоединении к основе, оканчивающейся квазисуффиксом *–c,* квазисуффикса *–iamo* окончание искомой словоформы примет вид *–chiamo*.

4) Набор правил синтеза начальной формы слов, мотивирующих образование дериватов с квазисуффиксами 1 порядка. Например, для английского квазисуффикса *–or*, правила синтеза начальной формы слова

$$[X + S]^{сущ} \rightarrow [X]^{глаг}$$
$$[X + S]^{сущ} \rightarrow [X + \text{‘e’}]^{глаг},$$

где X – последовательность символов, формирующих псевдооснову, S – место квазисуффикса первого порядка. При исполнении алгоритма согласно данным правилам для словоформы *renovator* сначала синтезируется форма *renovat*, к которой присоединяются возможные для данной части речи квазисуффиксы второго порядка, репрезентирующие начальную форму этой части речи (для английского языка – нулевой квазисуффикс), и осуществляется ее поиск по исходному списку. После того, как форма не будет найдена в списке, по второму правилу из набора осуществляется синтез формы *renovate*. Если форма имеется в списке, она помечается соответствующей частью речи (глагол).

Правила синтеза мотивирующих словоформ имеет целью расширить влияние правил алгоритма и определить лексико-грамматическую принадлежность некоторых недеривационных словоформ. Это особенно актуально для слабо-флективных аналитических языков.

Результатом прохождения всех этапов алгоритма является конфигурация, в которой словоформы, предположительно принадлежащие одной морфологической парадигме, объединены в одну строку таблицы. Столбцы этой таблицы репрезентируют абстрактную конфигурацию парадигмы соответствующей части речи. Фрагмент такого описания для английских существительных с псевдосуффиксом *–er* представлен ниже («пустоты» в ячейках объясняются отсутствием в выборке соответствующих форм) (см. табл. 1).

Таблица 1

Common Case, singular	*Common Case, plural*	*Genitive Case, singular*	*Genitive Case, plural*
Cleaner	cleaners	cleaner’s	
Clipper			clippers’

Преимуществом использования данной модели парадигматической идентификации является в том, что в лингвистической базе знаний описывается лишь относительно немногочисленная замкнутая группа квазисуф-

фиксов и правил их присоединения, при этом она не включает в свой состав теоретически бесконечный открытый класс квазиоснов. Это сокращает временные затраты на разработку базы знаний. Вместе с тем, алгоритм позволяет автоматизировать определение лексико-грамматического класса большой части словоформ из корпуса текстов.

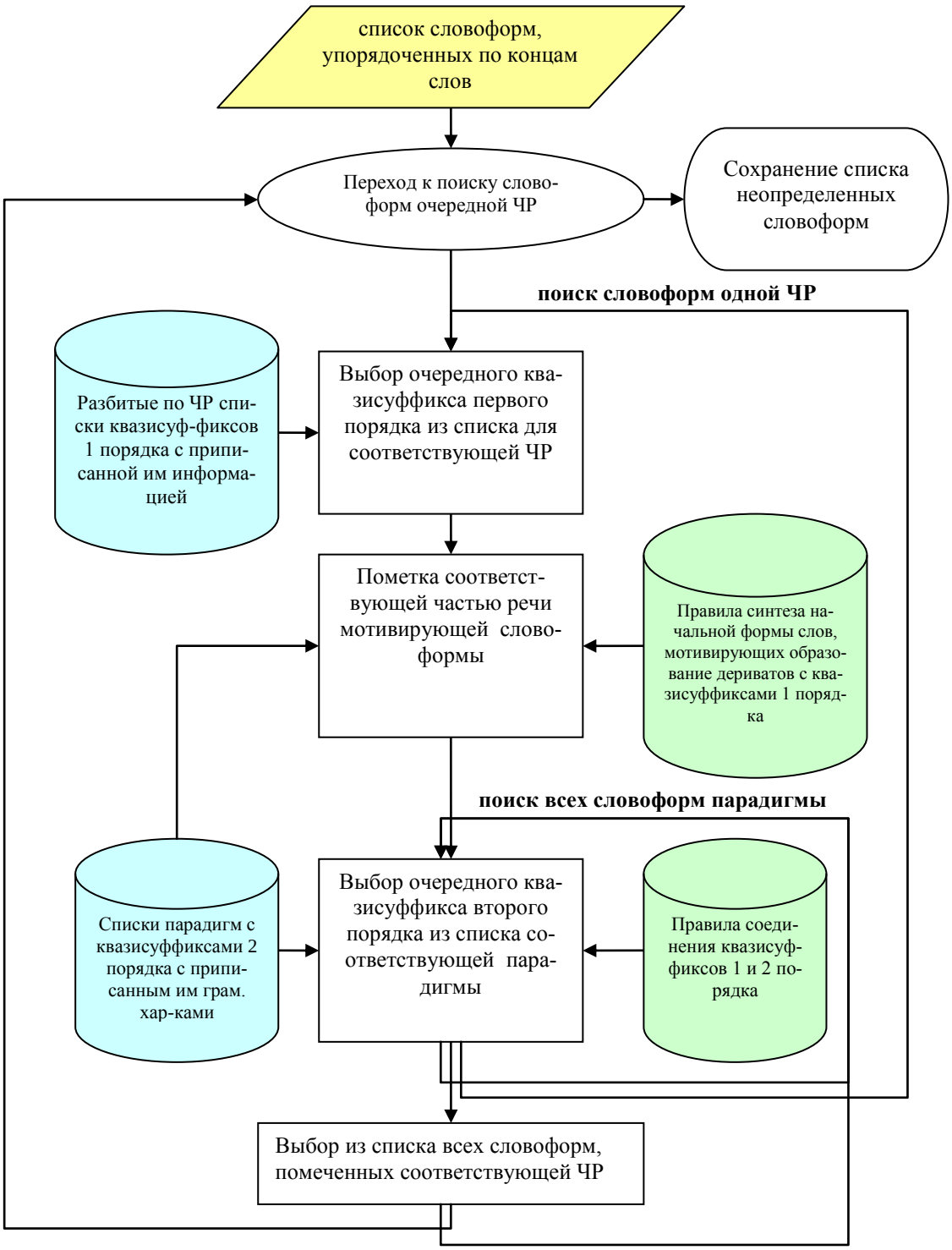

Рис. 1. Общая схема информационных потоков при выполнении первичной парадигматической идентификации