

## Лексические функции и машинный перевод

### 1. Вводные замечания

Главное направление научных работ лаборатории компьютерной лингвистики ИПШ РАН - разработка действующей компьютерной модели естественного языка на основе известной лингвистической теории "Смысл  $\Leftrightarrow$  Текст" (см., например, *Мельчук 1974*). Основным практическим приложением этой модели является система ЭТАП-3 высококачественного машинного перевода (МП), в которой воплощены многие принципы указанной лингвистической теории (см. *Апресян и др. 1989, Апресян и др. 1992, Apresjan et al. 1992, Apresjan et al. 1993*). Рабочие языки системы - русский и английский, перевод между которыми осуществляется в обе стороны, однако система располагает также небольшими экспериментальными компонентами русско-французского, русско-немецкого и русско-корейского перевода.

### 2. Лексические функции в модели "Смысл $\Leftrightarrow$ Текст"

В семантическом компоненте этой модели самым существенным образом используется большое количество сведений о каждом слове моделируемого языка, которые помещаются в словарных статьях так называемого толково-комбинаторного словаря (ТКС). Одним из важнейших типов лексикографической информации, содержащихся в словарной статье некоторого слова X, являются максимально широкие сведения о тех словах, которые регулярным образом связаны с ключевым словом X по смыслу, а именно: а) его **парадигматические варианты**, или **замены**, - слова или словосочетания, которые могут заменять X в тех или иных контекстах; б) его **синтагматические партнеры**, или **параметры**, - лексические единицы, которыми некоторые общие смыслы идиоматично выражаются при данном слове X. Для обозначения таких связанных с X слов в модели "Смысл  $\Leftrightarrow$  Текст" используется термин **лексические корреляты**, а для описания зависимости между X-ом и его лексическими коррелятами применяется аппарат **лексических функций (ЛФ)**. В соответствии со сказанным выше используются два типа ЛФ - **ЛФ-замены** и **ЛФ-параметры**.

Идеология лексических функций и разработка перечня основных ЛФ принадлежат И.А.Мельчуку, А.К.Жолковскому и Ю.Д.Апресяну (см., в частности, *Жолковский и Мельчук 1965, Апресян и др. 1968*), а представительные фрагменты ТКС для русского и других языков (в первую очередь французского) составлены группой ученых московской лингвистической школы и их последователей (см. *Мельчук и Жолковский 1984, Mel'cuk et al. 1984, 1988, 1992*). Однако описание словарных статей в этих публикациях выполнено в виде, далеком от формализации и тем самым недоступном для компьютерной обработки.

### 3. Введение информации о лексических функциях в систему ЭТАП-3.

В 1994-1996 гг. в нашей лаборатории в рамках международного проекта INTAS был создан компьютерный учебник лексики, назначение которого - в игровой форме способствовать освоению лексики и сочетаемости как родного, так и неродного языков. Этот учебник разработан для русского, английского и немецкого языков и содержит свыше 1000 словарных статей, насыщенных информацией о ЛФ, для каждого из трех языков. Русский и английский словари этого учебника составил Ю.Д. Апресян, а немецкий словарь был по этим образцам подготовлен коллегами из Гейдельберга и Клагенфурта. Список ЛФ для учебника, отобранный и упорядоченный Ю.Д. Апресяном, содержит 96 функций. Описание каждой функции сопровождается ее толкованием. Все толкования также составлены Ю.Д. Апресяном.

Важно, что весь лексический материал учебника представлен в виде, допускающем практически полную формализацию. Поэтому появилась возможность перенести этот материал в русский и английский словари ЭТАПа-3 и включить в несколько систем лингвообработки, а именно:

- системы англо-русского и русско-английского МП;
- системы перифразирования в рамках одного языка (русского или английского);
- комбинированные системы "перевод+перифразирование" (перифразирование может осуществляться как во входном, так и в выходном языках).

Темой данной работы является использование ЛФ-параметров в МП системы ЭТАП-3. Что касается системы перифразирования, которая оперирует в первую очередь информацией о ЛФ-заменах и является важной составляющей семантического компонента модели "Смысл  $\Leftrightarrow$  Текст", то она реализована в ЭТАПе-3 в настоящий момент лишь в эскизной форме на небольшом количестве правил.

По нашему мнению, организация информации о лексических функциях и работа с ними в системе МП должна удовлетворять следующим основным принципам:

- Информация о ЛФ в словарной статье комбинаторного словаря (КС)<sup>1</sup>, точно так же, как и сведения о синтаксических и семантических признаках слова и его модели управления, носит **внутриязыковой** характер. Она касается самого ключевого слова статьи и не должна быть ориентирована на перевод его лексических коррелятов на конкретный выходной язык и может быть использована, в равной степени, при перифразировании или при переводе на какой-либо третий язык.
- Информация о ЛФ в статье КС должна записываться в форме, привычной лингвистам, и быть максимально лаконичной.
- Информацию о ЛФ следует использовать в точно определенных местах алгоритма МП (а не "размазывать" ее по всем этапам).

---

<sup>1</sup> Словари системы ЭТАП-3 мы называем комбинаторными, а не толково-комбинаторными, поскольку они пока не содержат лексикографических толкований.

- Включение в МП работы с ЛФ ни при каких обстоятельствах не должно ухудшить качества МП.

#### 4. Примеры ЛФ в статьях КС и образцы переводов

Ниже приводятся две статьи русского КС и соответствующие им статьи английского КС, в которых хорошо представлена информация о ЛФ.

##### 15417 АПЛОДИСМЕНТЫ

POR:S  
 SYNT:мужск,мн!  
 DES:'факт', 'действие', 'абстракт'  
 D1.1:род  
 D2.1:дат  
 \_SYN: *рукоплескания/оvation*  
 \_V0: *аплодировать*  
 \_MULT: *взрыв/буря/гром*  
 \_MAGN: *бурный/продолжительный/восторженный/громкий*  
 \_ANTIMAGN: *сдержанный/скупой/редкий/жидкий*  
 \_VER: *заслуженный*  
 \_LABOR1-2: *встречать/награждать*  
 \_OPER2: *вызывать2*  
 \_INCEPFUNC0: *раздаваться/грянуть/вспыхивать*  
 \_FINFUNC0: *смолкать/стихать/прекращаться*  
 \*\*\*\*\*

ZONE:A

TRANS:APPLAUSE

##### 18751 APPLAUSE

POR:S  
 SYNT:count,voc  
 DES:'факт', 'действие', 'абстракт'  
 D1.1:*of*  
 D2.1:*to1*  
 \_SYN:*ovation*  
 \_V0:*applaud*  
 \_MAGN:*heavy/lengthy/prolonged/loud1/thunderous/ringing/terrific*  
 \_ANTIMAGN:*light2/weak/thin*  
 \_VER:*well-merited*  
 \_MULT:*storm/round1/thunder1*  
 \_OPER1:*give*  
 \_OPER2:*draw1/win/get*  
 \_LABOR1-2:*meet<with>/greet<with>/hail2<with>/rise1<in1>*  
 \_INCEPFUNC0:*break1<out>*  
 \_FINFUNC0:*subside/die<out>*  
 \*\*\*\*\*

ZONE:R

TRANS:АПЛОДИСМЕНТЫ

##### 02293 УВАЖЕНИЕ

POR:S  
SYNT:средн,ед!  
DES:'факт', 'чувство', 'свойство', 'абстракт'  
D1.1:род, 'лицо'  
D1.2:co\_стороны  
D2.1:к  
D2.2:род  
\_SYN: почтение  
\_ANTI: презрение  
\_GENER: отношение  
\_MAGN: большой/глубокий/огромный  
\_V0: уважать  
\_VER: искренний/заслуженный  
\_OPER1: испытывать/питать  
\_OPER2: пользоваться/внушать  
\_INCEPOPER2: снискивать/завоевывать  
\_FINOPER2: терять  
\_LABOR1-2: относиться<с3>  
\_CAUSFUNC0: вызывать2  
\_MANIF: оказывать  
TRAF:1-КОМПЛ.20  
\*\*\*\*\*

ZONE:A  
TRANS:RESPECT1

05593 RESPECT1  
POR:S  
DES:'факт', 'чувство', 'свойство', 'абстракт'  
D1.1: of  
D2.1: to1  
D2.2: of  
D2.2: for1  
D2.3: that1  
\_OPER1: have  
\_OPER2: command2  
\_INCEPOPER2: win  
\_LABOR1-2: hold1<in1>  
\_MANIF: show1  
\_VER: genuine  
\_MAGN: high1/profound/immense/solid1  
\_ANTIMAGN: scant11  
\*\*\*\*\*

ZONE:R  
TRANS:УВАЖЕНИЕ

Приведем теперь образцы переводов, полученных с использованием содержащейся в указанных статьях информации о ЛФ и без ее использования.

Пример 1.

**Исходная фраза:**

(1) *Московские студенты наградили нового декана жидкими аплодисментами.*

**Переводы:**

(1.1) *Moscow students met the new dean with a light applause.*

(1.2) \**Moscow students rewarded the new dean with a liquid applause* (что-то вроде '... наградили аплодисментами в жидком виде').

В (1.1) при переводе использованы сведения о значениях ЛФ LABOR1-2 и ANTIMAGN, а именно:

- LABOR1-2 (*аплодисменты*) = *награждать* ⇒ LABOR1-2 (*applause*) = *meet<with>*;
- ANTIMAGN (*аплодисменты*) = *жидкий* ⇒ ANTIMAGN (*applause*) = *light*.

В (1.2) все сведения о ЛФ игнорируются, и фраза получается неграмматичной.

Пример 2.

**Исходная фраза:**

(2) *The actor won terrific applause.*

**Переводы:**

(2.1) *Актер вызвал бурные аплодисменты.*

(2.2) \**Актер выиграл страшные аплодисменты.*

В (2.1) использованы сведения о значениях ЛФ OPER2 и MAGN:

- OPER2 (*applause*) = *win* ⇒ OPER2 (*аплодисменты*) = *вызывать*
- MAGN (*applause*) = *terrific* ⇒ MAGN (*аплодисменты*) = *бурный*

В (2.2) все эти сведения игнорируются.

Пример 3.

**Исходная фраза:**

(3) *Наш руководитель пользуется глубоким уважением своих коллег.*

**Переводы:**

(3.1) *Our leader commands high respect of his colleagues.*

(3.2) \**Our leader uses deep respect of his colleagues.* (что-то вроде '... использует глубоко лежащее уважение...').

В (3.1), в отличие от (3.2), учитывается значение ЛФ OPER2 и MAGN:

- OPER2 (*уважение*) = *пользоваться* ⇒ OPER2 (*respect*) = *command*;
- MAGN (*уважение*) = *глубокий* ⇒ MAGN (*respect*) = *high*.

Пример 4.

**Исходная фраза:**

(4) *Telling the plain truth, Mary won everybody's respect.*

**Переводы:**

(4.1) *Говоря чистую правду, Мэри снискала уважение всех.*

(4.2) *\*Говоря простую правду, Мэри выиграла уважение всех.*

В (4.1), в отличие от (4.2), учитывается значение ЛФ MAGN и INCEROPER2 (для разных слов):

- INCEROPER2 (*respect*) = *win*  $\Rightarrow$  OPER2 (*respect*) = *снискивать*;
- MAGN (*truth*) = *plain*  $\Rightarrow$  MAGN (*правда*) = *чистый*.

## 5. Информация о ЛФ и алгоритм МП<sup>2</sup>

Как видно из приведенных примеров, в качестве значений ЛФ в статьях КС может выступать либо лексема данного языка в основной форме, либо лексема с предлогом или (для английского языка) фразовым наречием, которые указываются в угловых скобках. В то же время нельзя, например, использовать в качестве значений ЛФ целые словосочетания, которые нередко встречаются в неформализованных ТКС, такие как OPER1 (*гнев*) = *быть во власти* или EQUIP (*университет*) = *профессорско-преподавательский состав*. Это естественное ограничение, наложенное на допустимые значения ЛФ, позволило перенести в КС системы ЭТАП-3 не менее 95% материала, собранного в компьютерном учебнике лексики.

Следует особо подчеркнуть, что сама форма представления ЛФ в КС системы ЭТАП-3 в точности повторяет ту, которая принята в этом учебнике и тем самым является вполне привычной для составителя словаря.

Помимо словарной информации о ЛФ, алгоритму МП доступны правила двух типов, обслуживающие эту информацию. Для каждого языка введено около 20 таких правил.

Первое обращение к этой информации происходит после получения синтаксической структуры переводимой фразы. Правила **первого типа** отождествляют некоторые узлы этой структуры как значения ЛФ-параметров от других узлов - аргументов этих ЛФ. При этом проверяется тип зависимости между этими словами (которая может быть и опосредованной), а также наличие требуемых предлогов или наречий, если они указаны в записи значения ЛФ.

Второе обращение к информации о ЛФ происходит на этапе перевода лексем на выходной язык. Сначала, как обычно, по зоне TRANS (т.е. по зоне стандартного перевода) переводятся те лексеммы, которые не являются значениями ЛФ-параметров. Затем правилами **второго типа** ЛФ-параметры переводятся по зоне ЛФ статьи **уже переведенного** ключевого слова. При этом в выходную фразу вводятся необходимые предлоги или наречия, если они указаны в значении ЛФ. Если же в словарной статье переведенного ключевого слова значения рассматриваемой ЛФ отсутствуют, то такие ЛФ-параметры переводятся по собственной зоне TRANS.

---

<sup>2</sup> В программной реализации модернизированного алгоритма системы ЭТАП-3 приняли активное участие В.Г. Сизов и И.Л. Сагалова, которым авторы выражают искреннюю признательность.

Алгоритм МП настроен и на перевод фраз с "вложенными" ЛФ (ср. *упорно добиваться цели*, где OPER1 (*цель*) = *добиваться* и MAGN (*добиваться*) = *упорно*).

Важной особенностью модернизированного алгоритма МП является возможность **множественного синтеза** выходного предложения. В частности, если в словарной статье лексемы выходного языка указано несколько значений требуемой ЛФ, то сначала для перевода берется первое значение, а затем, если поступит заявка на альтернативный перевод, будут поочередно использованы и остальные значения.

Например, в словарной статье английского существительного *veto* приводятся три значения ЛФ OPER1: *put*, *place* и *set*. Соответственно, предложение *Франция наложит вето на любую резолюцию ООН, осуждающую эту страну* (где *налагать* = OPER1 (*вето*)) получит несколько переводных эквивалентов: *France will put a veto on any resolution of the UN, condemning this country; France will place a veto...; France will set a veto...*

Отметим, наконец, что все правила обработки лексических функций носят **факультативный** характер. Это значит, что при порождении альтернативных переводных эквивалентов алгоритм может отменить применение этих правил и прибегнуть к буквальному, неидиоматичному переводу значений ЛФ, т.е. по существу вернуться к старому "дофункциональному" состоянию ЭТАПа-3. Тем самым обеспечивается последний из приведенных в п.3 принципов модернизации системы ("прежде всего - не навреди")

#### б. Лексические функции и модель управления слова

Одна из важнейших проблем всякой системы МП - представление в словаре многозначных лексем. В системе ЭТАП-3 обычно поступают следующим образом: для существенно разных значений лексем в словарь вводятся самостоятельные статьи, а для таких ее значений, которые не слишком отличаются друг от друга, используется одна обобщенная статья. Такое решение позволяет избежать чрезмерной омонимии на этапе анализа. В обобщенной статье, естественно, оказывается обобщенной и ее модель управления (МУ). При этом возможна ситуация, когда данная МУ отличается от "идеальной" МУ, которая была бы построена для некоторого конкретного значения лексемы (если бы не пришлось прибегать к ее обобщению), не только способами выражения валентностей, но и числом этих валентностей (т.е. количеством мест в модели управления). Например, в МУ статьи русского глагола *поступать*, как и в других глаголах перемещения, для описания этого перемещения фиксируется два актантных места: *откуда* и *куда* (ср. *Из осведомленных источников в газету поступила интересная информация*). В то же время в некоторых значениях этот глагол имеет по существу лишь одну валентность: *куда* (ср.

(5) *Юноша поступил в институт,*

при невозможности или крайней сомнительности *\*Юноша поступил из армии в институт*). Поэтому на этапе анализа между глаголом и его дополнением

может возникнуть синтаксическое отношение, не вполне адекватное использованному значению глагола (2-компл вместо 1-компл), что в ряде случаев может сказаться на качестве перевода. При использовании информации о ЛФ это несоответствие может быть устранено. Покажем это на примере перевода предложения (5). В статье слова *институт* есть запись INCEPRREAL1: *поступать<v1>*. Из определения INCEPREAL1 известно, что для глагола, являющегося значением этой ЛФ, ключевое слово представляет собой первое дополнение. Это знание позволяет нам подправить (“реинтерпретировать”) синтаксическую структуру, заменив 2-компл на 1-компл, и в результате получить с помощью аппарата ЛФ безупречный перевод *The young man entered the institute*.

### Заключение

Итак, аппарат ЛФ является удобным средством для описания лексических коррелятов слов. Если даже не принимать во внимание полифункциональность информации о ЛФ, которая может использоваться в различных системах лингвообработки, таких как системы перифразирования, программы, обучающие лексике и др., а ограничиться лишь машинным переводом, то нетрудно убедиться в том, что этот аппарат и здесь может принести несомненную пользу.

Разумеется, любая система МП должна иметь средства для обработки несвободных (прежде всего, терминологических) словосочетаний. Естественно, есть такие средства и в системе ЭТАП-3. Для перевода словосочетания здесь обычно достаточно ввести в словарную статью ключевого слова ссылку на некоторое правило, задав при ней в качестве параметров слова входного и выходного словосочетаний. Очевидно, что таким же образом можно организовать перевод и такой специфической группы словосочетаний, каковыми являются сочетания, включающие аргумент ЛФ и ее значение. Однако, аппарат ЛФ имеет перед переводом такого рода следующие преимущества:

- Составитель словаря получает удобный инструмент для представления "лексического мира" ключевого слова, а пользователь - наглядное средство для знакомства с этим миром.
- Информация о ЛФ не ориентирована на конкретный выходной язык, поэтому при подключении к системе МП нового языка не нужно производить никаких изменений в зоне ЛФ словарных статей, написанных ранее. В то же время традиционные правила перевода словосочетаний носят принципиально двуязычный характер и тем самым при добавлении нового языка в словарных статьях всех рабочих языков переводы словосочетаний потребуются обеспечивать заново. (Впрочем, для повышения эффективности перевода определенное согласование в описании зон ЛФ соответствующих слов разных языков - скажем, согласование списков имен ЛФ и порядка подачи в словарях разных значений одной ЛФ для этих слов - может оказаться желательным и даже необходимым).
- Информация о ЛФ в статьях КС, как было показано выше, может рассматриваться в качестве дополнения к МУ, позволяя в некоторых случаях



(а именно, когда многозначное слово употреблено в лексико-функциональном значении), подправлять синтаксическую структуру.

#### Литература

- Апресян и др. 1968:* Ю.Д. Апресян, А.К. Жолковский, И.А. Мельчук. О системе семантического синтеза. Образцы словарных статей. НТИ, серия 2, 1968, № 11, с. 8-21.
- Апресян и др. 1989:* Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин, А.В. Лазурский, Н.В. Перцов, В.З. Санников, Л.Л. Цинман. Лингвистическое обеспечение системы ЭТАП-2. М., Наука, 1989, 295 с.
- Апресян и др. 1992:* Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин, А.В. Лазурский, Л.Г. Митюшин, В.З. Санников, Л.Л. Цинман. Лингвистический процессор для сложных информационных систем., М., Наука, 1992. 256 с.
- Жолковский и Мельчук 1965:* А.К. Жолковский, И.А. Мельчук. О возможном методе и инструментах семантического синтеза. НТИ, 1965, № 6, с. 23-28
- Мельчук 1974:* И.А. Мельчук. Опыт теории лингвистических моделей "Смысл ↔ Текст". М., Наука, 1974, 314 с.
- Мельчук и Жолковский 1984:* И.А. Мельчук, А.К. Жолковский. Толково-комбинаторный словарь современного русского языка. Wiener Slawistischer Almanach, Sonderband 14, 1984, 992 с.
- Aprésjan et al. 1992.* Ju. D. Aprésjan, I.M. Boguslavsky, L.L. Iomdin et al. ETAP-2: The Linguistics of a Machine Translation System. META, 1992, Vol. 37, No 1, pp. 97-112.
- Aprésjan et al. 1993.* Ju. D. Aprésjan, I.M. Boguslavsky, L.L. Iomdin et al. Le système de traduction automatique "ETAP". In: La Traductique. P.Bouillon et A.Clas, ed. Montréal, Les Presses de l'Université de Montréal - AUPELF/UREF, 1993, p. 377-391.
- Mel'čuk et al. 1984.:* Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Adèle Lessard. Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques I. Les Presses de l'Université de Montréal, 1984.
- Mel'čuk et al. 1988:* Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Louise Dagenais, Лйо Elnitsky, Lidija Iordanskaja, Marie-Nolle Lefebvre, Suzanne Mantha. Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II. Les Presses de l'Université de Montréal, 1988.
- Mel'čuk et al. 1992:* Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha. Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III. Les Presses de l'Université de Montréal, 1992.