РАЗРАБОТКА ЛЕКСИКО-СЕМАНТИЧЕСКОГО СЛОВАРЯ КИТАЙСКОГО ЯЗЫКА ДЛЯ МНОГОЯЗЫЧНОЙ СИСТЕМЫ АНАЛИЗА ТЕКСТОВ

Маничева Е. С. (Ekaterina_M@abbyy.com),
Дрейзис Ю. А. (Yulya_D@abbyy.com),
Селегей В. П. (Vladimir_S@abbyy.com)

ABBYY, Москва, Россия

В статье описывается продолжающийся проект создания лексико-семантического словаря ядерной лексики китайского языка для системы автоматического анализа естественного языка, разрабатываемой в компании ABBYY. Оцениваются особенности задачи описания китайского языка в многоязычном проекте, ориентированном на универсальные семантические представления. Приводятся некоторые методологические трудности, с которыми столкнулись разработчики и принятые решения. Созданный лексико-семантический ресурс может представлять ценность как для систем автоматического анализа текстов, так и как основа учебного словаря китайского языка, адресованного прежде всего обучаемым с повышенными требованиями к полноте и системности лексикографического описания.

Ключевые слова: лексико-семантический словарь, компьютерная лексикография, китайский язык, глубинная семантическая модель, лексико-грамматическая сочетаемость

DEVELOPMENT OF CHINESE LANGUAGE LEXICAL-SEMANTIC DICTIONARY FOR THE MULTI-LANGUAGE NLP SYSTEM

Manicheva E. S. (Ekaterina_M@abbyy.com),
Dreyzis Yu. A. (Yulya_D@abbyy.com),
V. P. Selegey (Vladimir_S@abbyy.com)

ABBYY Software House, Moscow, Russia

This paper deals with bilingual lexical-semantic dictionary of Mandarin Chinese designed for NLP purposes. This dictionary of Chinese core vocabulary has been compiled according to the principles of model-based universal multi-language linguistic technology Compreno, developed in ABBYY. Nowadays most of lexical data-bases of Mandarin Chinese are based on WordNet principles. Our work shows that Chinese language might as well successfully fit in an alternative universal lexico-semantic database. Here we present an overview of major methodological challenges and solutions to integrate Chinese language data into Compreno framework. At the moment lexical-semantic dictionary of Mandarin Chinese covers more than 8000 meanings with well-structured comprehensive information on deep semantic and syntactic model of a meaning, its lexical and grammatical co-occurrence restrictions, and further work on dictionary is still going on. This paper focuses on typological differences between Chinese and European languages in terms of basic unit for dictionary entry, grammar paradigm of a word and its meanings, differences in syntactic realizations of deep semantic model. The paper also gives reasons why certain theoretical approaches prevailing in Chinese linguistic tradition were revised to serve better application needs, what principles of meaning definition were taken into consideration to provide detailed and complete lexicographic descriptions, compare and contrast Chinese with Russian or any other language.

Key words: lexical-semantic dictionary, computational lexicography, Chinese language, deep semantic model, lexico-grammatical co-occurrence

Введение. Китайский язык — проверка универсальности моделей описания

Для многоязычных систем компьютерного анализа текстов, основанных на лингвистических моделях (model-based), особую важность имеет проверка адекватности используемых моделей языковых описаний для типологически сильно различающихся языков.

Проект ABBYY Compreno [1] ориентирован на создание различных технологий анализа текста на основе универсальных языковых моделей. По многим соображениям работы по созданию языковых описаний начались с русского и английского языков. В ходе этой масштабной работы модель данных претерпела существенные изменения. Расширение списка рабочих языков за счет французского и немецкого языков не позволяло говорить о реальной типологической корректности модели описания.

В этом смысле начатая в 2008 г. работа над китайским описанием явилась настоящей проверкой проектных принципов. В данном докладе речь пойдет о тех задачах и проблемах, которые связаны прежде всего с лексико-семантическим описанием. Как кажется авторам, и проблемы, с которыми мы столкнулись, и созданный лексикографический ресурс имеют значение, выходящее за рамки отдельного, пусть даже и весьма амбициозного, проекта.

1. Подготовительный этап: ресурсы и методики

Перед началом работ по составлению словаря был осуществлен мониторинг центров разработки, технологий и ресурсов по автоматической обработке китайского языка, китайской системной лексикографии, компьютерной и корпусной лингвистике, а также проектов по разработке моноязычных и многоязычных лексико-семантических баз.

На начальном этапе проекта нам была очень важна заинтересованная помощь коллег из РГГУ, ИССА, СПбГУ. В ходе совместных встреч и семинаров были обсуждены многие вопросы, легшие в основу методики описаний[1]. Поскольку имеющиеся лексические ресурсы для китайского языка оказались более скромными с точки зрения полноты лексикографического описания, чем ресурсы для других языков проекта, было принято решение, что лексико-семантическая часть китайского проекта должна быть относительно независимой, ориентированной не только на задачи автоматического анализа в Compreno, но и на создание учебного словаря китайского языка. Это обусловило интерес к проекту китаистов, не являющихся участниками проекта Compreno.

---

[1] Особенно ценным был вклад Тараса Ивченко, Юрия Музыченко, профессора Пекинского Университета Юй Шивэня, Веры Барас, Ирины Горбуновой, которым мы приносим искреннюю благодарность за участие в обсуждении методик на разных этапах проекта.

1.1. Ситуация с ресурсами по автоматической обработке китайского языка

Анализ имеющихся ресурсов по автоматической обработке текста (далее АОТ) показал, что лексико-семантической основой для решения различных задач по автоматической обработке китайского языка в Китае и на Тайване используются семантические сети, построенные по принципу WorldNet [3, 6, 7, 10,11]. При этом все они, за исключением разрабатываемого под руководством Центра компьютерной лингвистики Пекинского Университета (КНР) словаря понятий китайского языка Chinese Concept Dictionary (CCD), идут от структуры английского WordNet'a, а не китайского языка [7, 10].

Переводной принцип неизбежно приводит к существенной неполноте и искажениям в системе лексических значений языка. Кроме того, существующие лексические ресурсы оказались чрезвычайно бедными с точки зрения полноты лингвистической информации о содержащихся в них лексических единицах.

Таким образом, стало очевидным следующее:

- необходимо выбрать базовое лексическое описание китайского языка, хорошо покрывающее его частотное ядро;
- необходимо верифицировать систему лексических значений с точки зрения модели и дистрибуции и дополнить описание всеми необходимыми типами описаний;
- необходимо провести работу по отображению описываемых значений на структуру семантических классов универсальной лексико-семантической иерархии, в рамках которой происходит описание языков в Compreno.

1.2. Разработка методики лексикографического портретирования

В качестве базового словаря ядерной лексики было решено использовать китайский словарь HSK 汉语8000词词典A Dictionary of Chinese Usage [12]. Данный словарь был разработан в КНР в качестве лексического минимума для иностранцев, сдающих экзамен на определение уровня владения китайским языком. Словарь HSK8000 включает в себя 8822 словарных входа, упорядоченных по 4 приоритетам частотности. К первому приоритету относятся 1033 словарные статьи, 2018 — ко второму, 2202 — третьему и 3569 словарных статей — к четвертому приоритету. По мнению составителей словаря, этот лексический запас покрывает 95 % современной лексики.

Разработка лексико-семантического словаря конкретного языка основывается на универсальной модели, разработанной в рамках проекта Compreno [1]. По технологии работы собственно лингвистическому этапу описаний предшествует этап подготовки лексикографического описания, результатом которого становится словарь, который можно рассматривать как относительно независимый промежуточный результат описания.

В основных положениях лексикографическое описание строится на принципах, близких к принципам системной лексикографии, сформулированным Ю. Д. Апресяном в [2]. Описание должно в явном виде содержать все типы лексикографической информации, которые необходимы для процедуры включения лексического значения в лексико-семантическую иерархию, используемую в технологиях Compreno.

Для китайского язык пришлось существенным образом уточнить методику лексикографического описания. В начале работ были выделены 3 основных лексикографических типа статей (для глагола, прилагательного и существительного) и определены структуры описания лексики каждого из типов.

Лексикографическая статья должна была содержать максимум лексико-семантической и синтаксической информации, релевантной для будущих задач автоматической обработки текста. Для всех лексикографических типов указывалась следующая информация: транскрипция, написание в полной форме, орфографические варианты, перевод на русский и английский языки, толкование, семантическая и синтаксическая модель, сочетаемость с грамматическими модификаторами, возможность употребления в различных синтаксических позициях, стандартная сочетаемость, идиоматика, синонимы и антонимы. При описании глаголов и прилагательных необходимо было проиллюстрировать примерами глубинную и поверхностную модель, указать возможные залоговые трансформации, возможность образования редуплицированных форм и случаи именного употребления.

В процессе подготовки был разработан пакет методических документов для лексикографов: методика подготовки лексикографического портрета, инструкция по работе с лингвистическими ресурсами, указания по глоссированию примеров, расширенный список семантических ролей (называемых в проектной терминологии глубинными позициями, или ГП) с примерами их выражения в китайском языке. В качестве базовых источников по синтаксическому употреблению были выбраны 3 словаря грамматической сочетаемости: 1) HSK GTU: HSK词语用法详解. A Guide to the Usage of HSK Vocabulary [15]; 2) DC-CC-YF: 汉语动词用法词典 (Словарь употребления глаголов) [13]; 3) XRC-CC-YF: 汉语形容词用法词典 (Словарь употребления прилагательных) [14]. Недостающую или недостоверную информацию лексикографы должны были проверять по другим печатным и электронным словарям, корпусам, конкордансам и иным надежным источникам в литературе и интернете.

2. Интеграция китайского языка в проектную идеологию

Поскольку основной задачей доклада является прежде всего описание наиболее модельно-независимой компоненты — собственно лексикографического описания, мы не будем касаться здесь деталей устройства иерархии, в которую, в конечном счете, оно отображается. Остановимся только на тех элементах структуры иерархии, которые явно присутствуют в исходном

лексикографическом описании и могут рассматриваться как объективные свойства языка, а не артефакты реализации.

Семантическая иерархия представляет собой лексико-семантический словарь, в котором содержится вся лексика языка, необходимая для анализа и синтеза текста в процессе перевода. СИ организована как дерево родо-видовых отношений, в узлах которого находятся **семантические классы** (СК) — универсальные (единые для всех языков) объекты, отражающие некоторое понятийное содержание, и **лексические классы** (ЛК) — конкретноязыковые объекты. СК характеризуются единичным наследованием и моделью семантических отношений (глубинной моделью), которая отражает потенциально возможные отношения с другими классами в любом естественном языке.

Глубинная модель описывается через глубинные позиции (ГП) — семантические отношения, в которые вступает управляющий элемент со своими зависимыми. Лексические потомки СК в конкретных языках описываются в иерархии через взаимосвязь глубинной модели управления и поверхностных форм выражения этих отношений, или диатезные соответствия.

В одном СК могут объединяться синонимы, антонимы и регулярные дериваты, образующие общее семантическое поле. Например, в СК 'TEMPERATURE' в русском языке в качестве потомков расположены лексические классы *«холодный», «горячий», «тепло», «мороз», «греть», «разогревать»* и ряд других. Все ЛК отмечены смыслоразличительными атрибутами — **семантемами,** являющимися элементарными семантическими признаками. Например, лексические потомки одного СК могут различаться по семантической категории **полярности** (*холодный/горячий, тепло/холод*), **локализации в пространстве** (*сзади/спереди/вокруг; выбежать/вбежать*), **размеру** (*дом/домище/домишко*), **стилистике** (*видеть/зреть*) и многим другим.

Особенности синтаксического употребления и грамматического значения маркируются в СИ при помощи **граммем**. Например, глагол «зреть» отличается от глагола «видеть» не только стилевой окраской, но и грамматическими свойствами. Глагол «видеть» может подключать клоз при помощи союза «что», а его синоним «зреть» — нет. Данные особенности словоупотребления будут помечены в DPS соответствующей граммемой.

В соответствии с логикой СИ были проанализированы и критически пересмотрены принципы выделения значений исходного словаря HSK8000. Кроме того, в процессе заведения в СИ китайских классов стала формироваться система грамматических категорий, опирающаяся не столько на существующие грамматические теории, сколько на эмпирический материал, и вместе с тем согласующаяся с проектными принципами описания синтаксиса.

2.1. Критический взгляд на базовый словарь HSK8000

Словарь HSK8000 достаточно полно покрывает лексическое ядро и дает базовую информацию о грамматической сочетаемости описываемых

единиц, именно поэтому он был выбран в качестве словарной базы и основы описания. Однако в ходе системных работ над лексико-семантическим словарем для системы АОТ стала наглядно видна его специфика как учебного словаря, ориентированного в первую очередь на иностранного студента. Главная задача словаря HSK8000 состоит не в том, чтобы системно описать лексику, а в том, чтобы помочь иностранцу овладеть базовым словарем и грамматикой. К несистемности описания лексики можно отнести следующие случаи:

1. Спорные приоритеты частотности.

Например, в 1 первый приоритет частотности попало слово 留学生 (liú xué shēng) *студент, проходящий стажировку за рубежом*, а на самом деле это слово частотно только в ограниченной языковой среде.

2. Непоследовательность в выделении единиц для словарного входа.

Словарными входами в словаре бывают единицы разных уровней:

- морфемы (化 (huà) в значении суффикса *-изировать; -изация*, 主义 (zhǔ yì) в значении суффикса *-изм*);
- слова (发展 (fā zhǎn) *развивать, развитие*, 父母 (fù mǔ) *родители*);
- словоформы с грамматическим показателем (养成 (yǎng chéng) *воспитывать* + показатель результатива);
- идиоматические единицы (一帆顺风 (yì fān shùn fēng) *идти как по маслу; без сучка, без задоринки*, 十全十美 (shí quán shí měi) *верх совершенства*, букв. 10 совершенств, 10 достоинств);
- свободные словосочетания (红旗 (hóng qí) *красный флаг*, 按劳分配 (àn láo fēn pèi) *распределять, распределение по труду)*;
- синтаксические показатели (着 (zhe) *показатель прогрессива*, 勿 (wù) *не, отрицание при императиве*);
- синтаксические конструкции (从。。。看来 (cóng kàn lái) *отсюда видно, можно сделать вывод*, 东跑西走 (dōng pǎo xī zǒu) *носиться туда-сюда*).

Для удобства коллективной работы словарь HSK8000 был поставлен под специальное программное приложение, в котором можно проводить полный мониторинг обработки значения от этапа назначения исполнителя до этапа проверки приписывания диатезных соответствий и делать SQL запросы.

По результатам анализа исходной словарной базы на первом этапе работ было решено отложить описание грамматических значений, синтаксических конструкций (670 значений) и идиоматических единиц (269 значений). Кроме того, из словаря HSK 8000 были исключены 3000 значений, которые не относились к базовым единицам словаря, а именно:

1. Свободные словосочетания, которые можно разбирать и переводить пословно.

Например:

Словосочетание 红旗 hóng qí *красный флаг*, предложная группа 按劳分配 àn láo fēn pèi *распределять, распределение по труду*.

В случае отсутствия в словаре вокабулы и соответствующего значения, лексикограф добавлял и описывал данное значение, а не значение всего словосочетания.

2. Словоформы с грамматическим показателем.

Например:

Ссловарный вход 养成 yǎng chéng в значении «воспитывать» был исключен в пользу аналогичного значения глагола 养 yǎng. Результативный показатель 成 chéng не является обязательным для выражения данного значения, оно может быть выражено как при помощи других результативных показателей, так и без них.

3. Случаи повторного выделения значения для слов в нехарактерной синтаксической позиции

Например: 发展 fā zhǎn *развивать* и *развитие*.

2.2. Критерии выделения значений в лексико-семантическом словаре

Типологической особенностью китайского языка является бедная морфология. Поэтому основным критерием для частеречной классификации в китайской лингвистической традиции становится критерий наиболее частотного синтаксического употребления слова в данном значении [8]. Большинство китайских двусложных предикатов может употребляться в именной функции как отглагольное имя, имя-характеристика, имя, обозначающее абстрактное понятие и имя, обозначающее конкретный объект. Разные случаи грамматической конверсии могут трактоваться как слова одной или разных частей речи в одном или нескольких значениях. В словаре HSK8000 критерии выделения значений и определения части речи не были эксплицитно сформулированы, и логика составителей не всегда представлялась последовательной.

Согласно проектным принципам описания языков, вместо понятия часть речи было решено оперировать понятием тип синтаксической парадигмы (ТСП), который описывает всё множество возможных составляющих с определенным типом элементов в качестве ядра. Следуя этим принципам, а также традициям отечественного [5] и западного [4, 9] китаеведения мы выделили следующие типы синтаксических парадигм для знаменательных слов: имя, предикат, наречие, числительное и местоимение. Для обозначения лексикализованных грамматических свойств слова в каждом конкретном значении использовались граммемы соответствующих грамматических категорий.

Проиллюстрируем разницу подходов к выделению значений и их грамматического типа в словаре HSK8000 и китайском словаре, построенном на базе СИ, на примере трех вокабул:

1. 发展 (fā zhǎn) : 发展经济 (fā zhǎn jīng jì) *развивать экономику,* 经济发展 (jīng jì fā zhǎn) — *развитие экономики.*

HSK8000: два значения для глагола *развивать* и существительного *развитие*

Китайский словарь в СИ (далее СИКит): одно значение, предикативный ТСП, возможность именного употребления обозначена граммемой.

2. 稳定 (wěn dìng) : 稳定关系 (wěn dìng guān xì) *стабилизировать отношения* 稳定的关系 (wěn dìng de guān xì) *стабильные отношения* 关系的稳定 (guān xì de wěn dìng) — *стабилизирование (стабильность) отношений*

HSK8000: одно значение для прилагательного *стабильный*

СИКит: одно значение, предикативный тип СП, возможность именного употребления, атрибутивного и именного употребления со значением характеристики промаркированы соответствующими граммемами и семантемой.

3. 翻译 (fān yì) : 翻译文章 (fān yì wén zhāng) — *переводить статью,* 文章的翻译 (wén zhāng de fān yì) — *перевод статьи,* 一篇翻译登在网站 (yī piān fān yì dēng zài wǎng zhàn) *опубликовать статью на сайте,* 一位翻译 (yī wèi fān yì) *один переводчик.*

HSK8000: два значения для глагола (*переводить*) и существительного (*переводчик*).

СИКит: три значения. Первое значение — *переводить* (предикат), именное употребление отмечено граммемой. Второе значение — *переводчик* (имя). Третье значение *перевод* как материальный носитель, тип текста (имя).

Таким образом, в соответствии с задачами АОТ главным критерием выделения значения является различие семантической и синтаксической модели, а тонкие различия в семантике и переводе до определенной степени вторичны. В том случае, если словарь будет использоваться для задач машинного перевода, точность перевода слова в том или ином языковом контексте можно будет обеспечить механизмами лексикализованного перевода при помощи переводных правил и словосочетаний. На данном этапе работ лексикограф только приводит примеры, не вписывающиеся в предлагаемый им переводной эквивалент, и по возможности указывает контексты, в которых предпочтителен иной перевод.

Например, у глагола 喂 (wèi) в словаре HSK 8000 выделено 2 значения: 1) кормить, давать еду животным и 2) давать еду, лекарство людям. Семантическая и синтаксическая модель обоих значений не имеет принципиальных различий. И в русском, и в английском языках перевод будет одинаковым независимо от того, дают пищу животным или людям,– «кормить кого-то чем-то, feed somebody with something». Однако если заполнителем данной глубинной позиции будет потомок СК, обозначающий лекарственные средства, предпочтительным представляется перевод «давать кому-то что-то». Если заполнителем глубинной позиции будет жидкий объект (имеющий семантему <<Liquid>>), то более точный перевод будет «поить кого-то чем-то». Предложенные переводные эквиваленты являются непосредственными потомками СК, в котором лежит лексический класс «давать», и наследуют его модель, поэтому выделение более чем одного значения глагола 喂 (wèi) представляется нецелесообразным. В большинстве контекстов подходит перевод «кормить», поэтому глагол 喂 (wèi) был

размещен лексикографом в СК ‘TO_FEED’, а случаи отладки перевода были прокомментированы.

Таким образом, по результатам обработки первых трех приоритетов словаря HSK8000 и заведения их в СИ в ходе перегруппировки, добавления и исключения значений на 01 апреля 2012 было описано 7852 лексических значения из 4590 вокабул.

2.3. Типологические особенности синтаксического выражения глубинной модели

Работа над словарной статьей в проектной среде включала следующие задачи: верификация, сортировка и ранжирование исходных словарных значений из словаря HSK8000, составление подробного лексикографического комментария к каждому значению согласно его типу синтаксической парадигмы, приписывание смыслоразличительных в рамках СК семантем и атрибутов грамматических свойств.

В мере обработки языкового материала лексикографический комментарий становился более формализованным, особенно в части описания диатезных схем и залогов. Поверхностная реализация глубиной модели стала описываться в терминах поверхностных позиций, выделенных исходя из представлений об оптимальном синтаксическом разборе. Грамматические свойства лексики маркировались граммемами. По мере обработки все большего объема лексики расширялась номенклатура грамматических категорий и поверхностных позиций. На данный момент для китайского языка определены 72 поверхностные позиции и 30 грамматических категорий (ГК). Используемые на данный момент ГК описывают следующие свойства: тип синтаксической парадигмы, тип объекта при предикате, модель редупликации, возможность залоговых трансформаций, тип имени и счетного слова, возможность слова выступать в той или иной синтаксической роли, линейный порядок в предложении, особые случаи подключения зависимых, модель деривации.

Типологическими особенностями китайского языка является относительно жесткий порядок слов и ограниченные возможности подключения зависимых в линейном порядке относительно ядра. В китайском языке при глаголе в препозиции и постпозиции к ядру, как правило, одновременно может выражаться не более двух актантных ГП. По этой причине семантическая модель китайского значения и его переводного эквивалента в европейском языке зачастую отличаются несимметричностью. При переводе предложения с многоактантной структурой на китайский язык приходиться прибегать к следующим переводческим трансформациям:

- редупликации глагола;
- выносу дочерней зависимой в прадочернюю;
- введению сочиненной или зависимой предикации;
- выносу зависимой в топик.

Пример 1

Русский глагол *«ударить»* допускает одновременное выражение при ядре четырех ГП: [Agent] (кто ударяет), [Object] (кого ударяет), [PlaceOfContact] (куда ударяет) и [Instrument] (чем ударяет).

(1) [Agent: Сяо Ли] ударил [Object: Сяо Вана] [Instrument: кулаком] [PlaceOfContact: по голове].

Китайский переводной эквивалент глагола 打 (dǎ) допускает только одну зависимую в постпозиции и две в препозиции. При необходимости описать аналогичную ситуацию в китайском языке нужно либо удвоить глагол, либо вынести зависимую [PlaceOfContact] в прадочернюю к [Object].

(2) [小李 (xiǎo lǐ)]/[一拳 (yī quán)]/打 (dǎ)/小王 (xiǎo wáng)/, /打 (dǎ)/在 (zài)/头 (tóu)/上 (shàng)

Сяо Ли/один/кулак/ударить/Сяо Ван/,/ударить/предлог/голова/

(3) 小李 (xiǎolǐ)//一 (yī)/拳 (quán)//打 (dǎ)/[PlaceOfContact:/在 (zài)[/Object:/小王 (xiǎo wáng)/]/头 (tóu)/上 (shàng)]

Сяо Ли/один/кулак/ударить/предлог/Сяо Ван/голова/послелог

Пример 2

В китайском языке ГП [Instrument] при глаголах движения не выражается непосредственно при ядре. Выражение этой зависимой возможно только посредством дополнительной предикации. При этом в входном и выходном языке мы будем иметь две разные глубинные структуры:

(4) [Agent: Он] вернулся [Locative_FinalPoint:домой] [Instrument: на самолете]

(5) [Agent: 他 (tā)/[MannerOfPositionAndMotion:坐 (zuò)/[Locative:飞机 (fēi jī)]]/回 (huí)/[Locative_FinalPoint:家 (jiā)]

Он/сидеть/самолет/возвращаться/домой

Таким образом, в переводе мы получаем — «он, сидя в самолете, вернулся домой».

Пример 3

Часто при невозможности выразить зависимую при ядре, она выносится в топик.

(6) [Sphere:说话 (shuōhuà)/]他 (tā)/都 (dōu)/很 (hěn)/仔细 (zǐ xì)

разговаривать/он/всегда/очень/внимательный

Он очень внимателен [Sphere: в том, что говорит].

Пример 4

При глаголе 播 bō «сеять» выражение ГП [Locative_IntialPoint] и [Locative_FinalPoint] невозможно, однако, оно допустимо при его двусложном синониме 播种 bō zhòng (сеять).

(7) [从 cóng/村子 cūn zi/南 nán/边 biān/][到 dào/北 běi/边 biān]都 dōu/播种 bō zhǒng/了 le/[小麦 xiǎo mài]

с/деревня/юг/край/до/север/край/все/PFV/пшеница
[С южного края деревни] [до северного края] все засеяно пшеницей.

Когда проблема несимметричности носит универсальный и регулярный характер, преобразования исходной структуры не комментируются, но фиксируются как актуальные проблемы семантических и синтаксических описаний китайского языка, которые будет решаться на дальнейших этапах работ. На данный момент лексикограф только указывает случаи несимметричности для основных (актантных и наиболее частотных) глубинных позиций для лк и СК, поясняет, с каким синонимом из СК она может быть выражена, а также предлагает возможные варианты переводческих трансформаций.

Заключение: перспективы разработок и использования словаря

На данный момент в СИ описаны порядка 8000 тысяч значений и обработана лексика первых трех приоритетов словаря HSK8000, что, по мнению составителей, покрывает 91% лексики современного китайского языка. Такой объем работ позволяет считать разработку лексико-семантического словаря ядерной лексики китайского состоявшейся.

В ходе 4-летних работ был создан уникальный продукт, не имеющих аналогов в мире, — это первый системный переводной лексико-семантический словарь ядерной лексики китайского языка. В словаре представлена подробная информация по семантической и синтаксической сочетаемости, словоупотреблению, идиоматике. Подход к описанию менялся и развивался в ходе разработок, поэтому часть комментариев имеет смысл расширить с учетом приобретенного в ходе работ опыта. Кроме того, формат и содержание словарной статьи может быть пересмотрен, сокращен или расширен с учетом возможного применения этого продукта. На данный момент перспективными представляются следующие сценарии использования словаря:

1. Самостоятельный продукт — переводной словарь грамматической сочетаемости

В зависимости от аудитории (изучающих китайский язык или профессиональных лингвистов), примеры в словаре могут быть снабжены транскрипцией, проглоссированы, переведены на любой иностранный язык,

а семантическая и синтаксическая модель упрощена или описана более подробно и формализовано.

2. Лексическая база для автоматической обработки китайского языка

Данный словарь изначально разрабатывался под задачи системы анализа текста, создаваемой в компании ABBYY. Специфика китайского языка и недостаточная проработанность грамматической теории стала причиной того, что работа над лексикографическими описаниями стоилась по принципу поиска наиболее оптимального синтаксического анализа и учёта всех грамматических явлений, что отражено в проектной методике. Данный словарь может быть использован как лексическая база для задач автоматической обработки китайского языка, разрабатываемых не только в компании ABBYY, но и за ее пределами.

3. Техническое задание для учебной и/или научной грамматики.

В результате работы над словарем и обработки большого лексического материала стали очевидны несовершенство и непоследовательность имеющихся на данный момент учебных и научных грамматик китайского языка. Кроме того, для решения задач автоматической обработки текста нужна специальная грамматика, словарь и база данных грамматических свойств. Пример грамматического словаря такого типа — словарь грамматической информации, созданный в Институте Компьютерной Лингвистики Пекинского университета под руководством профессора Юй Шивеня [16].

Разработка лексико-семантического словаря ядерной китайского языка заложила крепкую основу для будущих системных описаний грамматического строя китайского языка, что могло бы лечь в основу принципиально новой грамматики языка, ориентированной на решение разнообразных прикладных, педагогических и научных задач.

References

1. *Anisimovich K. V., Druzhkin K. Y., Minlos P. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012). Syntactic and Semantic Parser based on ABBYY Compreno Linguistic Technologies. Kompyuternaia Lingvistika I Intellektual'nye Technologii: Trudi Mezhdunarodnoj Konferentsii "Dialog 2012" T. 2 [Computational Linguistics and Itellectual Technologies: Proceedings of the International Conference "Dialog 2012" Vol. 2]. Bekasovo, pp. 91–103.
2. *Apres'an Yu. D.* (1995). Integral'noe opisanie yazika i sistemnaya leksikografiya [Integral description of language and systemic lexicography]. Moscow.
3. *Chu-Ren Huang, Ru-Yng Chang, Hsiang-Pin Lee.* Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO, Academia Sinica, available at: http://bow.sinica.edu.tw/file/BOW040528WN01.pdf

4. *Claudia Ross, Jing-heng Sheng Ma* (2006). Modern Chinese Grammar. Routledge Modern Grammar Series, London and New York.
5. *Dragunov A. A.* (1952). Issledovaniya po grammatike sovremennogo kitajskogo yazika [Studies on Modern Chinese Language Grammar], Moscow and Leningrad.
6. *Kam-Fai Wong, Wenji Li, Ruifeng Xu, Zheng-sheng Zhang* (2009). Introduction to Chinese Natural Language Processing, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.
7. *Liu Yang, Yu Shiwen, Yu Jiangsheng* (2005). CCD Yuyi zhishikude gouzao yanjiu [Study on the Construction of CCD]. Xiaoxing weixing jisuanji xitong [Mini-micro systems], Vol. 26, Ed. 8.
8. *Van Li* (1989) Parts of speech [Chasti rechi], Novoe v zarubezhnoj lingvistike [New Trends in Foreign Linguistics], Moscow, Vol. 22., pp. 37–53.
9. *Yip Po-Ching, and Don Rimmington* (2004). Chinese: A Comprehensive Grammar. Routledge Comprehensive Grammars Series, Routlenge.
10. *Yu Jiangsheng, Yu Shiwen* (2002). Zhongwen gainian cidiande jiegou. [The Structure of Chinese Concept Dictionary]. Zhongwen xinxi xuebao [Journal of Chinese Information Processing, 2002, Vol. 16, Ed.4.
11. *Zhendong Dong, Qiang Dong.* HowNet, available at: http://www.keenage.com/

Dictionaries

1. 汉语8000词词典. Hanyu 8000 ci cidian [A Dictionary of Chinese Usage: 8000 words], Beijing, 2007
2. 汉语动词用法词典. Hanyu dongci yongfa cidian [A Dictionary of Chinese verbs usage], Beijing, 2005
3. 汉语形容词用法词典. Hanyu xingrongci yongfa cidian [A Dictionary of Chinese adjectives usage], Beijing, 2003
4. HSK词语用法详解. HSK ciyu yongfa xiangjie. [A Guide to the Usage of HSK Vocabulary], Beijing, 2006
5. 现代汉语语法信息词典详. Xiangdai hanyu yufa xinxicidian xiangjie [The Grammatical Knowledge-base of Contemporary Chinese — A Complete Specification], Beijing, 2002.