ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р ИСО 24614-1— 2013

Менеджмент языковых ресурсов Пословная сегментация письменных текстов

Часть 1 Основные концепции и общие принципы

ISO 24614-1:2010
Language resource management – Word segmentation of written texts – Part 1: Basic concepts and general principles
(IDT)

Издание официальное

Москва
Стандартинформ
2014

gost.e-ecolog.ru

Предисловие

1 ПОДГОТОВЛЕН ЗАО «Проспект» на основе собственного аутентичного перевода на русский язык международного стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 08 ноября 2013г. № 1386-ст

4 Настоящий стандарт идентичен международному стандарту ИСО 24614-1:2010 «Менеджмент языковых ресурсов. Пословная сегментация письменных текстов. Часть 1. Основные концепции и общие принципы» (ISO 24614-1:2010 «Language resource management – Word segmentation of written texts – Part 1: Basic concepts and general principles»).

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в ГОСТ Р 1.0—2012 (раздел 8). Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок – в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования – на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (gost.ru)

© Стандартинформ, 2014

Настоящий стандарт не может быть воспроизведен, тиражирован и распространен в качестве официального издания без разрешения национального органа Российской Федерации по стандартизации

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Менеджмент языковых ресурсов. Пословная сегментация письменных текстов. Часть 1. Основные концепции и общие принципы

Language resource management – Word segmentation of written texts – Part 1: Basic concepts and general principles

Дата введения — 2015—01—01

1 Область применения

В настоящем стандарте представляются основные понятия и общие принципы пословной сегментации и даются не зависящие от языка руководящие указания по сегментации письменных текстов надежным и воспроизводимым способом на единицы пословной сегментации (WSU).

ПРИМЕЧАНИЕ: В связанной с языком научно-исследовательской и практической работе слово является фундаментальным и необходимым понятием. Поэтому для целей сегментации текста на слова важно иметь универсальное определение того, что включает слово. Нельзя просто использовать для разграничения слов правила, основанные на идентификации пробелов и знаков пунктуации. Такие правила не учитывают случаи сложных слов, которые пишутся через дефис, сокращений, идиом или словоподобных выражений, содержащих символы или цифры. Пословная сегментация еще более проблематична в языках, которые не содержат пробелов для разделения слов, например, для китайского и японского языков, а также в агглютинативных языках, где некоторые классы функциональных слов реализуются как аффиксы, например, в корейском языке.

Некоторые применения и сферы, которые требуют сегментировать тексты на слова и к которым, следовательно, применима данная часть ИСО 24614, представлены ниже

Перевод

Подсчет слов является главным методом оценки стоимости перевода. Пословная сегментация - это стандартная функция в системах переводческой памяти и в инструментальных средствах автоматизированного перевода (CAT). Пословная сегментация выполняется средствами извлечения терминов, которые иногда предоставляются в системах управления терминологией и в средствах CAT.

Управление контентом

Большинство систем и баз данных для управления информационным содержанием (контентом) предусматривают поиск по отдельным словам. Содержание, по которому производится поиск, должно быть сегментировано, чтобы была возможность сравнения со словом поиска. Кроме того, поисковые функции требуют знания границ слов.

Технологии распознавания речи

Системы речевого воспроизведения текста синтезируют речь на базе слов и поэтому требуют пословной сегментации для обеспечения возможности словарного поиска, расстановки ударений, установления просодического образца и др.

Прикладная лингвистика

Различные системы обработки текстов на естественных языках (NLP) должны сегментировать текст на слова для того, чтобы выполнить свои функции. Системы NLP включают:

- морфосинтаксические программы обработки,
- синтаксические анализаторы,
- программы проверки правописания,
- системы классификации текстов, и
- лингвистическое аннотирование корпуса текстов.

Лексикография

Лексические ресурсы часто оцениваются по их объёму - обычно на основе подсчёта числа слов.

ПРИМЕЧАНИЕ: Объём языковых ресурсов - весьма важный показатель для управления ими. Количественное определение объёма языковых ресурсов, как правило, основывается на подсчёте количества слов. Однако поскольку в приложениях NLP используются разные методы сегментации, каждый из них подсчитывает число слов по-разному и даёт в итоге разные суммы для одного и того же текста. Наличие надёжной воспроизводимой стандартной меры могло бы обеспечить получение сопоставимых результатов. Однако это не значит, что приложения не могут использовать свои специфические методы сегментации; например, в системе синтеза речи текст может сегментироваться на меньшие или большие единицы по сравнению с другими приложениями.

Издание официальное

2 Термины и определения

В данном документе используются следующие термины и определения:

2.1 сокращение (abbreviation): Вербальное обозначение, образованное путем исключения слов или отдельных букв из более длинной формы и идентифицирующее то же самое понятие.

[ISO 1087-1:2000]

2.2 аффикс (affix): Связанная морфема, которая может добавляться к основе или лексеме.

Примечание - Аффиксы можно классифицировать на несколько подтипов, например, префикс, суффикс, инфикс и циркумфикс. Аффиксы могут быть деривационными, инфлективными или агглютинативными.

2.3 агглютинация (agglutination): Процесс присоединения одного или большего числа аффиксов к основе.

[ISO 24613:2008]

2.4 заимствование (borrowing): Процесс образования слова, в котором лингвистическое выражение заимствуется из другого языка, как правило, когда не существует термина для нового объекта или понятия.

2.5 связанная морфема (bound morpheme): Морфема (2.18), которая появляется только вместе с одной или несколькими другими морфемами.

Пример 1 – Китайский: иероглиф 伟 означает «великий», но он не может помещаться отдельно как слово в тексте. Вместо этого он может использоваться как составляющий элемент многих слов, например, 伟大 «великое», 伟人 «гигант», and 雄伟 «величие».

Пример 2 – Корейский: суффикс «-е», который эквивалентен английскому предлогу «to» — как в «hakkyo-e» (в школе), – это связанная морфема.

[ISO 24613:2008]

2.6 сложное слово (compound): Слово, построенное из двух или большего числа лексем.

Примечание 1 - Адаптированное определение 3.10 из ISO 24613:2008.

Примечание 2 - Сложное слово может быть эндоцентрическим, если оно имеет ведущее слово (т.е. основную часть, которая содержит основной смысл всего сложного слова) и модификаторы (которые ограничивают это смысловое значение) или экзоцентрическим, если оно не имеет ведущего слова. Сложное слово может быть длинным. Существуют два главных подтипа сложных слов в соответствии со степенью их лексикализации: составное слово и фразовое образование.

2.7 словосложение (compounding): Способ образования слов, при котором новое слово составляется путем соединения по крайней мере двух лексем в их исходных формах или с небольшими изменениями.

[ISO 24613:2008]

2.8 словообразование (derivation): Изменение в форме слова для создания нового слова, обычно путем модификации основы или аффиксации.

[ISO 24613:2008]

2.9 свободная морфема (free morpheme): Морфема, которая может самостоятельно использоваться как слово.

Пример – В английском слове «goodness» (доброта) основа «good» является свободной морфемой, тогда как часть, «-ness» таковой не является и представляет собой связанную морфему.

2.10 омограф (homograph): Каждая из двух или большего числа форм слова или слов с идентичным правописанием, но представляющие различные понятия (семантическая омонимия) или синтаксические функции (синтаксическая омонимия).

2.11 флексия (inflection): Процесс, в котором форма слова составляется путем добавления аффикса к основе.

Примечание - Флексия – это скорее грамматический, чем лексический процесс.

2.12 лемма (lemma): Обычная форма, выбранная для представления лексемы.

Пример – Для английских словоформ «find» (находить), «finds» (находит) «found» (найденный) и «finding» (отыскание) в качестве леммы для представления группы всех этих форм слова выбирается форма «find».

[ISO 24613:2008]

2.13 лемматизация (lemmatization): Процесс определения леммы для заданной формы слова в контексте.

Пример – В английском языке для слова «found» лемматизация даёт в результате в качестве леммы слово «find».

Примечание - Адаптированное определение 2.19 из ISO 1087-2:2000 и определение 3.14 из ISO 30042:2008.

2.14 лексема (lexeme): Абстрактная единица, как правило, связанная с набором форм, имеющих общее смысловое значение.

Примечание 1 - Лексема может быть частью другой лексемы – как результат словообразования и словосложения.

Примечание 2 - "Форма" определяется в ISO 24613 как "последовательность морфов".

2.15 лексикализация (lexicalization): Процесс создания функции лингвистических единиц, таких как слово.

Примечание - Такой лингвистической единицей может быть отдельный морф, например, «laugh» (смех), последовательность морфов, например, «apple pie» (яблочный пирог), или даже фраза, такая как «kick the bucket» (протянуть ноги), которая является идиоматическим выражением.

2.16 морфемный словарь (lexicon): Список статей, в основном озаглавленных леммами, с ассоциированной информацией.

2.17 морф (morph): Поверхностная форма, представленная уникальной морфемой.

Пример – В английском языке морфы морфемы множественного числа «-s» включают «-s», «-en», и «-NULL» (как в «boys», «oxen» и «sheep»), где «–NULL» не имеет уникальной поверхностной формы. Таким образом, слово «boys» состоит из двух морфов: «boy» и «-s», тогда как морфемами, соответствующими морфам «ox» и «-en» являются «ox» и «-s», соответственно.

2.18 морфема (morpheme): Наименьшая смысловая единица, выраженная последовательностью фонем или последовательностью графем.

Примечание - Существуют два подтипа морфем: свободные морфемы и связанные морфемы.

[ISO 24613:2008]

2.19 многословное выражение (multiword expression, MWE): Лексема, образованная последовательностью других лексем и имеющая свойства, не вытекающие из свойств отдельных лексем или их комбинации в нормальной форме

Примечание - Многословное выражение может быть сложным словом [составным словом или фразовым образованием, идиомой, фрагментом предложения или высказыванием (например, пословицей или привычным выражением)]. Не всегда можно определить часть речи всего многословного выражения (MWE).

[ISO 24613:2008]

2.20 фразовое образование (phrasal compound): Слово, состоящее из двух или большего числа лексем, смысл которого вытекает из составляющих его элементов

Пример – В английском языке словосочетание «apple pie» – это фразовое образование, состоящее из двух лексем , «apple» (яблоко) и «pie» (пирог), чьи значения сохраняются в смысле сложного слова .

Примечание 1 - В идиомах используются два или большее число лексических единиц, тем не менее они не являются фразовым образованием.

Примечание 2 - Фразовое образование может рассматриваться некоторыми лингвистами как фраза. Однако на практике не всегда существует чёткое различие между составным словом и фразовым образованием, или между фразовым образованием и фразой вследствие размытости семантической предсказуемости и степени лексикализации. Лексическая статистика — в особенности частота слов — может играть в этой связи важную роль.

2.21 редупликация (reduplication): Явление повторения слова целиком или частично.

2.22 основа (stem): Лингвистическая единица, чья форма меньше или равна форме единственной лексемы и которая может подвергаться инфлективному, агглютинативному, композиционному или словообразовательному процессу.

[ISO 24613:2008]

2.23 слово (word): Лексема, которая, как минимум, характеризуется частью речи.

[ISO 24613:2008]

2.24 форма слова (word form): Морфосинтаксический вариант данного слова.

Пример – В английском языке цепочка слов «find», «finds», «found», и «finding» представляет различные формы слова «find».

2.25 пословная сегментация (word segmentation): Процесс разделения текста на последовательность единиц пословной сегментации.

2.26 единица пословной сегментации (word segmentation unit WSU): Форма слова или символьная строка некоторого другого типа, которая трактуется как единица текста.

Примечание - Символьная строка, которая не является формой слова, может состоять из цифровых символов, иностранных символов, знаков пунктуации или некоторых других разнообразных символов, таких как китайские иероглифы, химические знаки, например, H_2O, или сочетание латинских и цифровых символов, например, F16.

2.27 структура слова (word structure): Внутренняя структура слова, выявляемая при морфологическом анализе.

Примечание - В агглютинативных языках, например, корейском, японском и турецком, слово может состоять из последовательности морфем со сравнительно высоким отношением морфем на слово, где каждый включенный аффикс (как словообразовательный, так и инфлективный) обычно однозначно выражает конкретное грамматическое значение. Структура слова в этих языках может быть очень сложной, со свободными морфемами и отдельными аффиксами как составляющими его элементами.

2.28 составное слово (word compound): Сложное слово, общее значение которого совершенно не выводимо из составляющих его частей.

Пример – «Hotdog» (бутерброд с сосиской), «ice-cream» (мороженое), «blackboard» (информационная доска).

3 Базовая структура для пословной сегментации

3.1 Основные понятия, относящиеся к пословной сегментации

Понятия, описанные в данном разделе, важны для понимания принципов пословной сегментации.

На рисунке 1 показано взаимоотношение между абстрактными сущностями «морфемой» и «лексемой» и конкретными сущностями: «морфом», «формами слова» и «словарём». Конкретной формой морфемы является морф. Конкретной формой лексемы является форма слова. Словарь составляется в основном из лемм, которые выводятся из форм слова с помощью процесса лемматизации.

ПРИМЕЧАНИЕ 1: Термины, такие как «морфема» и «слово» имеют различные значения в областях лингвистики и терминологии. Эти и другие термины используются (как описано в разделе 2) в соответствии с их лингвистической интерпретацией.

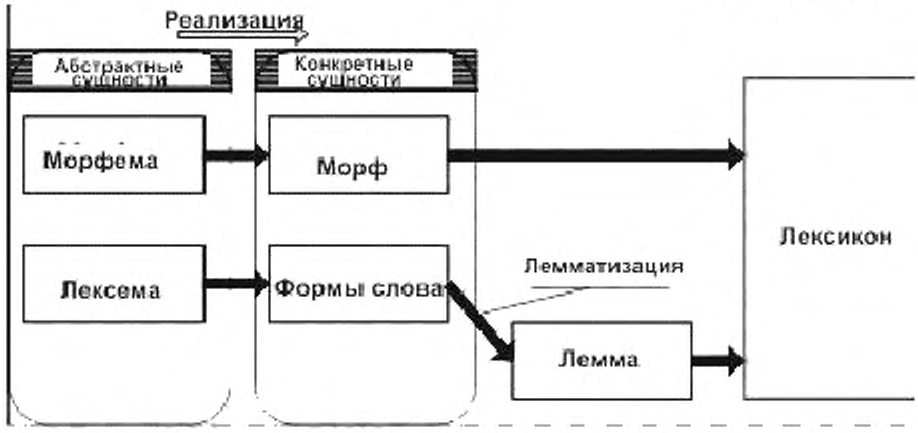

Рисунок 1. Связь между абстрактными и конкретными сущностями при построении словаря

Морфология изучает смысловые единицы языка, а также то, как они сочетаются при образовании слов. Морфологию можно разделить на лексическую морфологию, которая касается в основном словообразования на базе лексем, и на инфлективную либо агглютинативную морфологию (в зависимости от типа языка), которая рассматривает, главным образом, словообразование на основе морфем. Лексическая морфология включает в себя процессы словообразования, словосложения, сокращения, заимствования и редупликации (геминации).

ПРИМЕЧАНИЕ 2: Термин «лексическая морфология» используется чаще, чем «словообразовательная морфология», поскольку редупликация — это всего лишь один из способов образования слов.

Инфлективная и агглютинативная морфологии содержат два разных типа аффиксации и редупликации. Редупликация может давать в результате новые формы слова, поэтому она рассматривается также как процесс в лексической морфологии. Например, в языке африкаанс редупликация используется для подчеркивания значения повторяемого слова; например, слово "krap", означает "царапать", в то время как "krap-krap-krap" означат "сильно царапать". Для агглютинативных языков, где аффиксы присоединяются к основам, для выполнения пословной сегментации требуется особый набор морфологических правил.

ПРИМЕЧАНИЕ 3: Эти правила приводятся в стандарте ИСО 24614-2.

См. рисунок 2.

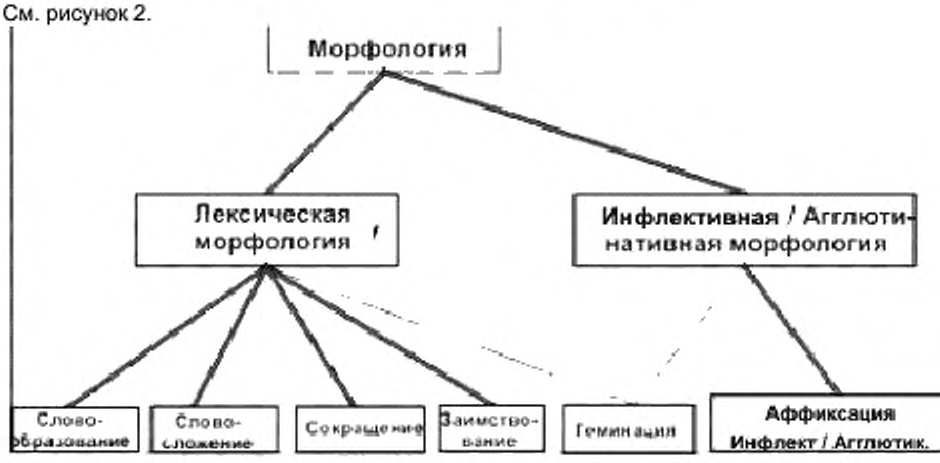

Рисунок 2. Система морфологии в языках

Многословные выражения (MWE) включают в себя сложные слова, идиомы, пословицы или разговорные выражения (см. рисунок 3). К сложным словам относятся составные слова и фразовые образования. Значение сложного слова не может быть выведено из значения его отдельных частей. Например, словосочетание «White House» (Белый Дом), обозначающее резиденцию президента США, относится к уникальному понятию, а не только к дому, который является белым. Однако значение фразового образования может быть выведено из значений его отдельных частей. Например, словосочетание «apple pie» (яблочный пирог) обозначает пирог (pie), сделанный из яблок (apples); по аналогии, «blueberry pie» (черничный пирог) — это пирог, сделанный из черники (blueberries). Тем не менее предыдущее словосочетание рассматривается как фразовое образование (см. определение термина и пример в п. 2.20) и содержит всего одну единицу пословной сегментации, поскольку сочетание слов «apple pie» встречается часто и даже используется в идиоматическом выражении, «American as apple pie» (традиционный для американцев), тогда как словосочетание «blueberry pie»" не обладает подобным свойством.

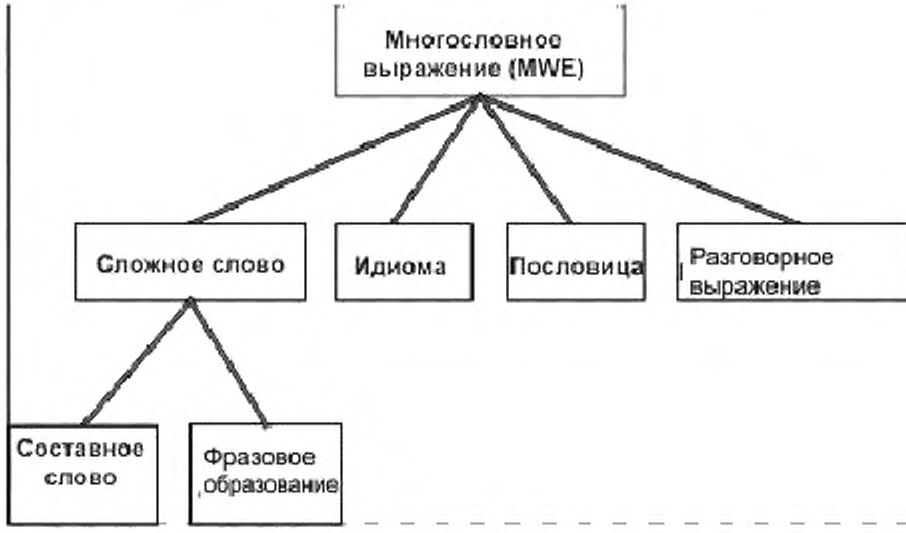

Рисунок 3. Типы MWE

WSU состоит из форм слова и других символьных строк. Символьные строки включают в себя либо содержат вперемешку цифровые или иностранные символы, знаки пунктуации или некоторые другие символьные строки: например, иероглифы в китайском тексте или символы согласных и гласных звуков в корейском тексте. Например, «Bravo!» содержит в слове восклицательный знак.

ПРИМЕЧАНИЕ 4: В некоторых случаях WSU содержат связанные морфемы, такие, как в корейском языке субстантивные суффиксы «-e» в «hakkyo-e»" (в школе) и «-ga» в «hakkyo-ga» (школа - именительный падеж), которые трактуются как принадлежащие к уникальной части речи, называемой «josa» (вспомогательная часть речи).

Структура единиц пословной сегментации представлена на рисунке 4.

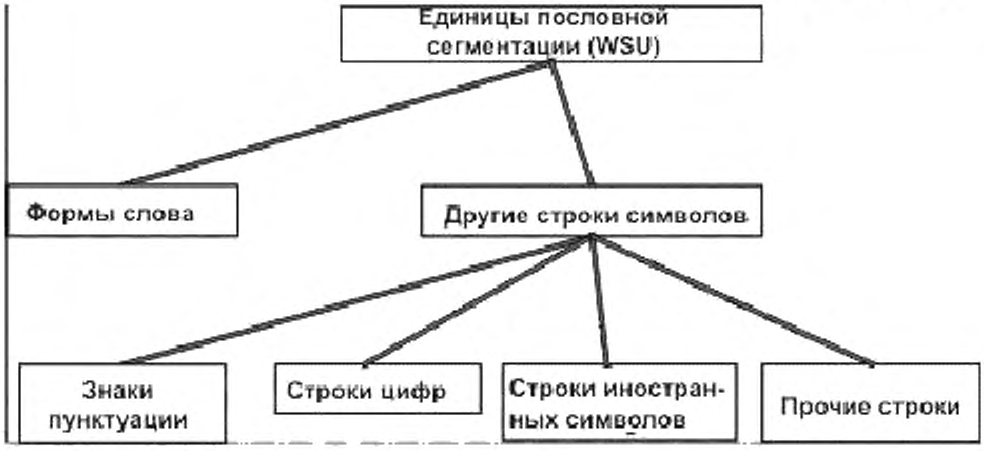

Рисунок 4. Типы WSU.

3.2 Ресурсы, которые могут облегчить пословную сегментацию

Проведению пословной сегментации в отдельной языковой области могут помочь следующие компоненты и ресурсы:

1. подходящий словарь;
2. список аффиксов, включая префиксы, суффиксы и инфиксы, если таковые имеются;
3. список связанных морфем, отличных от аффиксов;
4. спецификация для морфологии языка — для установления выхода пословной сегментации на базе зависимых от языка явлений по принципам, описанным в разделе 4;
5. представительный корпус текстов языка.

Чтобы обеспечить совместимость пословной сегментации разных текстов (или одного текста разными средствами) и гарантировать, что сегментация даст сравнимые количества, когда она применяется для подсчёта числа маркеров (см. п. 3.3) в текстовом документе, ресурсы, указанные выше в п.п. от а) до е), должны быть детально описаны в части их содержания.

3.3 Процесс пословной сегментации

Процесс пословной сегментации отображён в общих чертах на рисунке 5.

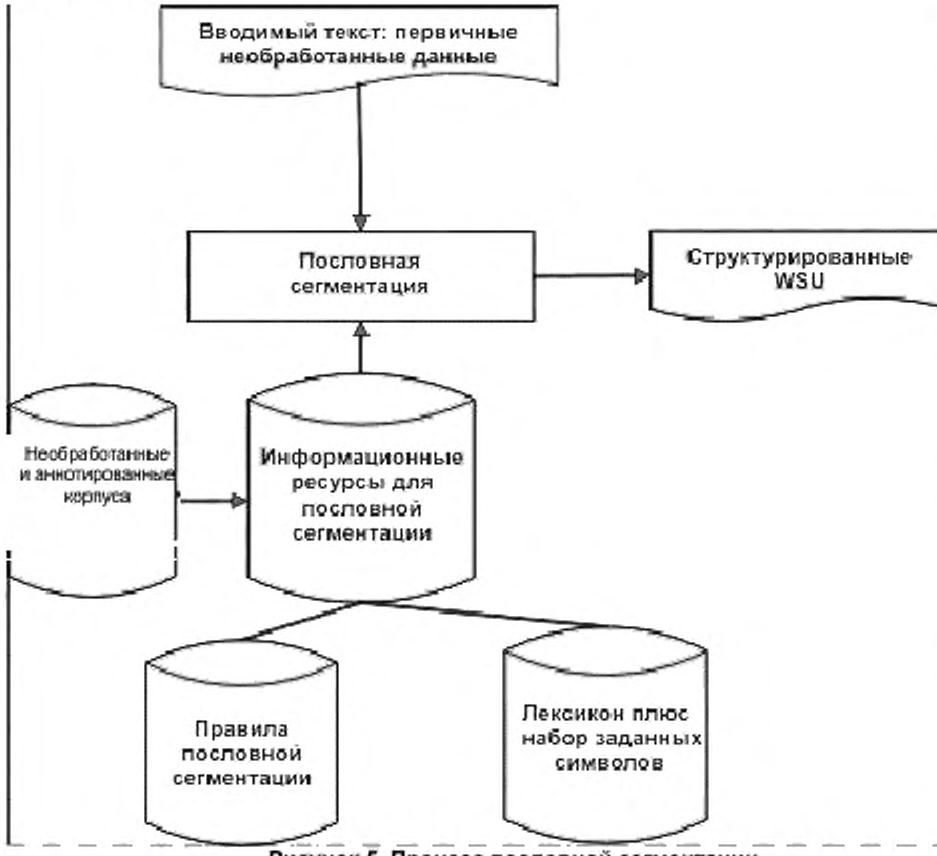

Рисунок 5. Процесс пословной сегментации

При заданных необработанных первичных данных текст сегментируется на символы и маркируется указателями местоположения, а затем сегментируется на подходящие базовые единицы в соответствии с требованиями стандарта ИСО 24612. Необработанные и аннотированные текстовые корпусы обеспечивают основу для построения словаря, который содержит словоформы и, возможно, список связанных морфем и символов. Также предоставляется набор правил пословной сегментации. Эти корпусы, правила пословной сегментации и словарь вместе составляют ресурсы, необходимые для преобразования первичной сегментации к сегментации, образуемой WSU.

Пример для китайского языка показан в форме графа на рисунке 6.



Рисунок 6. Иллюстрация первичной сегментации и пословной сегментации

На уровне первичной сегментации каждый символ помечается как интервал между двумя указателями местоположения (например, первый символ «白» на рисунке 6 помечается интервалом <0,1>). На уровне лингвистического аннотирования пословной сегментации первый результат «白菜» («white vegetable/ белое растение») определяется как слово, помеченное интервалом <0,2>, поскольку эти два символа не могут рассматриваться независимо. Второй единицей является слово из одного символа «和» («и»). Третьей единицей «猪肉» («pig meat/свиное мясо», «pork/свинина») является фразовое образование, помеченное интервалом <3,5> с внутренней структурой, которая состоит из двух WSU «猪» («pig/свиное») и «肉» («meat/мясо»), помеченных интервалами <3,4> и <4,5>, соответственно. В последнем случае существует два WSU, поскольку эти два символа могут существовать независимо, и каждый из них может вносить свой вклад в смысловое значение.

Пословная сегментация применяется к необработанному тексту и заканчивается разбиением заданного текста на последовательность WSU; в свою очередь, WSU может иметь внутреннюю структуру сегментации, когда разрешаются альтернативные сегментации. Во фрагменте текста задано предложение «Джон покинул Соединённые Штаты Америки»; это предложение сначала можно разделить на сегменты, называемые «маркерами», на основе некоторых правил сегментации – в данном случае просто на основе идентификации пробелов (для языков, в которых не используются пробелы, например, для китайского языка, для расстановки меток необходимо использовать другие правила). Затем, путём обращения к словарю цепочка сегментов, таких как «the United States of America», может трактоваться как одна лексическая единица, называемая «словом» или MWE и рассматриваемая как тип слова. Результаты второго этапа зависят от содержания словаря; некоторые словари могут включать не всю цепочку «the United States of America» как лемму, а только «United States of America» или даже только «United States».

4 Общие принципы пословной сегментации

4.1 Универсальный принцип морфологии

Универсальный принцип, положенный в основу стандарта ИСО 24614, состоит в том, что в каждом языке констатируется наличие слов и меньших лексических единиц, называемых «морфемами».

4.2 Принципы обоснования наличия WSU

4.2.1 Общие положения

Для контроля правильности выделения единиц пословной сегментации ниже приводятся две группы не зависящих от конкретного языка принципов: одна основана на лингвистической концепции, а другая выбрана с практической точки зрения. Специфичные для того или иного языка исключения описываются в других частях ИСО 24614, где рассматриваются конкретные языки. В разных ситуациях могут использоваться разные принципы, даже применительно к идентичным строкам текста.

4.2.2 Принципы, основанные на лингвистической концепции

а) Принцип связанной морфемы

Если к слову присоединяется связанная морфема, то результатом будет выделение одной WSU, как, например, в случае связанной морфемы «un» в слове «unhappy» (несчастливый).

б) Принцип лексической целостности

При применении синтаксических правил внутренняя структура слова не принимается во внимание. Если слово-кандидат удовлетворяет этому принципу, то оно, скорее всего, является отдельной WSU. К примеру, в словосочетании «the White House», определяющем резиденцию президента США, ничего невозможно вставить между вторым и третьим маркерами, т.е. нельзя сказать «White clean House» («Белый чистый Дом»), тогда как про обычный дом можно сказать «the white clean house».

в) Принцип невозможности вывода смыслового значения слова из составляющих его частей

Если слово-кандидат обладает свойством семантической непредсказуемости, то оно является отдельной единицей пословной сегментации. Например, слово «blackboard» не обязательно обозначает чёрную классную доску; так как во многих случаях такие доски бывают зелёными. Поэтому данное слово образует единственную WSU.

г) Принцип идиоматичности

Если какая-то последовательность словоформ используется в качестве идиоматического выражения, то она рассматривается как отдельная WSU (например, словосочетание «kick the bucket» используется как идиоматическое выражение «протянуть ноги»).

д) Принцип непродуктивности

Если слово-кандидат непродуктивно для образования слов, то оно, скорее всего, должно быть отдельной WSU. Например, в китайском языке иероглифы «白菜» дословно означают «белый овощ» и являются непродуктивным китайским словом, так как символ со значением «белый», не может быть заменён на символ, обозначающий любой другой цвет, поскольку получаемое словосочетание в китайском языке отсутствует.

4.2.3 Принципы практичности

а) Частотный принцип

Частота употребления является основным критерием при количественном определении степени лексикализации слова-кандидата. Слово или последовательность слов с высокой частотностью, скорее всего, должны быть отдельной WSU.

б) Принцип целостности словоформы (термин из науки о процессах познания)

Всякие сущности, по всей вероятности, должны восприниматься как нечто целое. Поэтому данный принцип создаёт основу для включения некоторых словосочетаний в качестве лемм в словарь, даже если они внешне выглядят как отдельные смысловые единицы.

в) Принцип категоризированных элементов-прототипов (из когнитивной лингвистики)

Согласно теории прототипов в отношении ментального лексикона, категоризированные элементы-прототипы более ярки, чем элементы, не являющиеся прототипами. Они более точно запоминаются в кратковременной памяти, легче удерживаются в долговременной памяти и легче вспоминаются людьми. Этот принцип даёт обоснование для включения некоторых фразовых образований, могущих служить прототипами, в шаблоны продуктивного словообразования, подобно «apple pie» в английском языке с лексическим шаблоном «фрукт + пирог» и «猪肉» ("свинина") в китайском языке с шаблоном «животное + мясо».

г) Принцип языковой экономии

Если включение слова-кандидата в словарь призвано облегчить его лингвистический анализ, то, скорее всего, это должно быть слово. Например, в китайском языке цепочка иероглифов «大中小学» («университет, средняя школа и начальная школа») является сокращением с довольно сложной структурой «большая, средняя или малая школа», где «большая школа» означает «университет», а «малая школа» – это «начальная школа». Цепочку «大中小学» нелегко идентифицировать как отдельную WSU, если она не зафиксирована в словаре.

4.3 Принцип полноты словарной статьи

В принципе все часто употребляемые WSU включаются в словарь, который должен быть также динамичным и адаптивным, чтобы можно было вводить новые WSU.

4.4 Принципы, касающиеся результатов пословной сегментации

а) Принцип разбиения по крупности

Результатом процесса пословной сегментации могут быть внутренние структуры получаемых WSU для представления возможных альтернативных сегментаций для различных сфер применения.

в) Принцип максимизации области действия аффиксов

Основа слова управляет всеми присоединёнными к ней аффиксами, а аффиксы формируют ту или иную часть одной и той же WSU. Например, при наличии в тексте слова «unexpectedness» (неожиданность): цепочка «un-expect-ed-ness» в целом является отдельной WSU, хотя слова

«expected» (ожидаемый) или «unexpected2" (неожиданный), если они встречаются в тексте, также представляют собой отдельные WSU.

г) Принцип максимизации области действия словосложения

Если сложное слово, найденное в тексте, включает в себя в соответствии со справочным словарём другое сложное слово, то более длинное сложное слово рассматривается как WSU, а более короткое сложное слово может быть помечено как альтернативная внутренняя WSU. Например, если в английском тексте присутствует строка «network operating system» (сетевая операционная система), то вся эта строка в целом трактуется как WSU, где более короткое словосочетание «operating system» (операционная система) может быть альтернативной WSU.

д) Принцип сегментации для других строк

Любые символьные строки, включая цифровые строки, строки иностранных символов или знаки пунктуации и любые их комбинации могут образовывать WSU, если они воспринимаются в тексте как несущие некоторую синтаксическую функцию. Например, в английском тексте «The Second World War ended in 1945» (Вторая мировая война закончилась в 1945 году) цепочка цифр «1945» является WSU, а в предложении «ㄱ is the first consonant character in Korean» (ㄱ является символом первого согласного звука в корейском" языке) символ «ㄱ» является в данном конкретном контексте единицей пословной сегментации, где этот знак функционирует как подлежащее этого предложения.

4.5 Принцип полного охвата и совместимости при использовании данной части ИСО 24614

Данная часть стандарта ИСО 24614 предназначена для использования применительно к любому тексту на любом языке. Однако для определённых языков необходимы некоторые изменения, описанные в других частях ИСО 24614. В последующих частях будет также дополнительно обсуждаться и иллюстрироваться важная роль контекста в определении крупности (гранулярности) разбиения и области действия пословной сегментации.

Приложение А
(справочное)

Представление процесса пословной сегментации на языке XML

Данная иллюстрация представления единиц пословной сегментации на языке XML соответствует требованиям стандарта ISO 24611.

```
<?xml version="1.0" encoding="UTF-8" ?>
<maf addressing="xpointer">
  <seg xml:id="seg0">白菜和猪肉</seg>
  <token xml:id="tok1" target="#string-range(seg0,0,1)" />
  <token xml:id="tok2" target="#string-range(seg0,1,1)" />
  <token xml:id="tok3" target="#string-range(seg0,2,1)" />
  <token xml:id="tok4" target="#string-range(seg0,3,1)" />
  <token xml:id="tok5" target="#string-range(seg0,4,1)" />
  <wordForm lemma="白菜" tokens="#tok1 #tok2"
      entry="urn:lexicon:cn:white_vegetable" />
  <wordForm lemma="和" tokens="#tok3" entry="urn:lexicon:cn:and" />
  <wordForm lemma="猪肉" tokens="#tok4 #tok5"
      entry="urn:lexicon:cn:pork">
    <wordForm lemma="猪" tokens="#tok4" entry="urn:lexicon:cn:pig" />
    <wordForm lemma="肉" tokens="#tok5" entry="urn:lexicon:cn:meat" />
  </wordForm>
</maf>
```

Библиография

[1] ISO 639-1:2002, *Коды для представления названий языков. Часть 1. Код альфа-2*

[2] ISO 639-2:1998, *Коды для представления названий языков. Часть 2 Код альфа-3*

[3] ISO 639-3:2007, *Коды для представления названий языков. Часть 3. Код альфа-3 для всестороннего охвата языков*

[4] ISO 639-5:2008, *Коды для представления названий языков. Часть 6. Код альфа-3 для семейств и групп языков*

[5] ISO 704, *Терминологическая работа. Принципы и методы*

[6] ISO 860, *Терминологическая работа. Гармонизация понятий и терминов*

[7] ISO 1087-1:2000, *Терминологическая работа. Словарь. Часть 1: Теория и применение*

[8] ISO 1087-2:2000, *Терминологическая работа. Словарь. Часть 2: Компьютерные приложения*

[9] ISO 24611, *Управление языковыми ресурсами. Структура морфо-синтаксического аннотирования* [1)]

[10] ISO 24612, *Управление языковыми ресурсами. Структура лингвистического аннотирования (LAF)*[1)]

[11] ISO 24613:2008, *Управление языковыми ресурсами. Схема лексической разметки (LMF)*

[12] ISO 12620, *Компьютерные приложения в терминологии. Категории данных*

[13] ISO 16642:2003, *Применение компьютера в терминологических целях. Структура терминологической разметки.*

[14] ISO 30042:2008, *Системы управления терминологией, знаниями и содержимым. Схема TermBase eXchange (TBX)*

[15] *Britannica Online Encyclopedia*, gost.e-ecolog.ru

[16] Allen, J., *Natural Language Understanding*, (1994) Addison Wesley

[17] Aronoff, M. and Rees-Miller, J., *The Handbook of Linguistics*. 2001, Blackwell

[18] Biber, D. et al., *Corpus Linguistics*. 1998, Cambridge University Press

[19] Bussmann, H., *Routledge Dictionary of Language and Linguistics*. 1996, Routledge

[20] Crystal, D., *The Cambridge Encyclopedia of Language*. 1997, Cambridge University Press

[21] Johnson, K. and Johnson, H., *Encyclopedia Dictionary of Applied Linguistics: A Handbook for Language Teaching*. 1999, Blackwell

[22] Kennedy, G., *An Introduction to Corpus Linguistics*. 1998, Addison Wesley Longman

[23] Matthews, P.H., *Morphology*. 1991, Cambridge University Press

[24] Packard, J.L., *The Morphology of Chinese: A Linguistic and Cognitive Approach*. 2000, Cambridge University Press

[25] Poole, S.C., *An Introduction to Linguistics*, 1999, Macmillan

[26] Richards, J. et al., *Longman Dictionary of Applied Linguistics*. 1985, Longman

[27] Ungerer, F. and Schmidt, H-J., *An Introduction to Cognitive Linguistics*. 1996, Addison Wesley Longman

[28] Zhu, Dexi, *Lecture on Grammar*, 2003, Commercial Press (written in Chinese)

[1)] Готовится к публикации.

УДК 001.4:006.354 ОКС 01.020, 01.040.01

Ключевые слова: менеджмент языковых ресурсов, пословная сегментация письменных текстов, терминология.

Подписано в печать 01.10.2014. Формат 60x84 1/8.

Усл. печ. л. 1,86. Тираж 31 экз. Зак. 3817.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

ФГУП «СТАНДАРТИНФОРМ»

123995 Москва, Гранатный пер., 4.

gost.e-ecolog.ru gost.e-ecolog.ru

gost.e-ecolog.ru