ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р ИСО 24615— 2013

МЕНЕДЖМЕНТ ЯЗЫКОВЫХ РЕСУРСОВ

Система синтаксического аннотирования (SynAF)

ISO 24615:2010
Language resources management – Sintactic annotation framework (SynAF)
(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 8 ноября 2013 г. № 1387-ст

4. Настоящий стандарт идентичен международному стандарту ИСО 24615:2010 «Менеджмент языковых ресурсов. Система синтаксического аннотирования (SynAF)» (ISO 24615:2010 «Language resources management – Sintactic annotation framework (SynAF)»).

При применении настоящего стандарта рекомендуется использовать вместо ссылочных международных стандартов соответствующие им национальные стандарты Российской Федерации, сведения о которых приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в ГОСТ Р 1.0—2012 (раздел 8). Информация об изменениях к настоящему стандарту публикуется в ежегодно издаваемом информационном указателе «Национальные стандарты», а текст изменений и поправок – в ежемесячно издаваемых информационных указателях «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ежемесячно издаваемом информационном указателе «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

Введение

Настоящий стандарт основан на многочисленных проектах и рабочих материалах, предшествовавших этапу стандартизации, которые разрабатывались в течение 1990-х годов [9] и касались создания эталонных моделей и форматов представления синтаксической информации, являющейся результатом работы синтаксического анализатора или аннотациями языковых ресурсов (в банках древовидных структур). На протяжении ряда лет стандартом де-факто для построения банков древовидных структур служил проект инициативной группы Пенсильванского университета Penn Treebank; однако более поздние работы, например, инициативные проекты Negra/Tiger в Германии (см: gost.e-ecolog.ru /projekte/TIGER /TIGERCorpus/) и ISST в Италии [18], продемонстрировали практическую восстребованность более однородной базовой системы, которая способна охватывать в равной степени как отношения иерархической соподчинённости компонентов, так и явление зависимости в синтаксическом аннотировании.

"Затравкой" для стандартизации стал проект "LIRICS" группы eContent, объединивший усилия множества экспертов, которые инициировали разработку проекта стандарта ИСО 24615 (по системе SynAF). На подготовительном этапе эта группа подтвердила, что в существующих инициативных проектах в действительности используется общая модель данных, которая обеспечивает добротную основу для построения метамодели SynAF [см. результаты проведённых исследований в информационном бюллетене Deliverable D.3.1 "Evaluation of initiatives for morpho-syntactic and syntactic annotation" ("Оценка инициативных проектов в области морфосинтаксического и синтаксического аннотирования")] в рамках проекта Евросоюза LIRICS, информацию о котором можно получить по адресу gost.e-ecolog.ru).

Настоящим стандартом предлагается метамодель для синтаксического аннотирования со списком релевантных категорий данных, которые она охватывает. Эти категории данных доступны на сервере ISOCat (gost.e-ecolog.ru) в синтаксическом профиле (определённом в соответствии со стандартом ИСО 12620:2009).

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Менеджмент языковых ресурсов. Система синтаксического аннотирования (SynAF)

Language resources management – Sintactic annotation framework (SynAF)

Дата введения — 2015— 01— 01

1 Область применения

В настоящем стандарте описывается система синтаксического аннотирования SynAF, являющаяся высокоуровневой моделью для представления синтаксической аннотации лингвистических данных с целью обеспечения возможности работы со всеми языковыми ресурсами или компонентами обработки языковых данных. Настоящий стандарт является дополнением стандарта ИСО 24611, тесно связан с ним в части схемы морфосинтаксического аннотирования MAF (morpho-syntactic annotation framework) и предоставляет метамодель для синтаксических представлений, равно как и эталонные категории данных для представления информации по составляющим элементам и отношениям зависимости в сложных предложениях или других сопоставимых высказываниях и сегментах.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие международные стандарты (для датированных ссылок следует использовать только указанное издание, для недатированных ссылок следует использовать последнее издание указанного документа, включая любые поправки и изменения к нему):

ИСО 1087-1:2000 Терминологическая работа. Словарь. Часть 1. Теория и применение (ISO 1087-1:2000, Terminology work. Vocabulary. Part 1. Theory and application)

ИСО 1087-2:2000[1)] Терминологическая работа. Словарь. Часть 2. Применение вычислительной техники (ISO 1087-2:2000, Terminology work. Vocabulary. Part 2. Computer application)

[1] ИСО 12620:2009 Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов (ISO 12620:2009, Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources)

ИСО 24611 *Управление языковыми ресурсами* (ISO 24611:2012, Language resource management – Morpho-syntactic annotation framework (MAF))

Примечание – При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

[1)] Отменен.

3 Термины и определения

3.1 обстоятельственное слово, обстоятельство, адъюнкт (adjunct): Второстепенный элемент, ассоциируемый с глаголом в отличие от синтаксических аргументов (3.19)

ПРИМЕЧАНИЕ В качестве обстоятельственных слов в предложении могут выступать наречия.

3.2 фрагмент (chunk): Нерекурсивная составляющая (3.4)

3.3 предложение (clause): Группа **фраз** (3.14), обычно содержащая некоторое высказывание

ПРИМЕЧАНИЕ Предложение может быть главным (3.10) или придаточным (3.17). В тех языках, где существует понятие законченности действия, глагол в предложении с глагольным сказуемым может быть совершенного или несовершенного вида – в зависимости от его конкретной формы. Главное предложение само по себе может представлять сложное высказывание (3.15). В модели SynAF предложение является особой формой конституенты (3.4).

3.4 конституента (constituent): Синтаксическая группировка слов [во **фразах** (3.14)], фраз [в **предложениях** (3.3) либо в других фразах] или элементарных предложений [в сложном предложении (3.15)], основанная на их структурных (или иерархических) свойствах

3.5 зависимость, отношение зависимости (dependency, dependency relation): Синтаксическая связь между **словоформами** (3.24) или **конституентами** (3.4), устанавливаемая на основе грамматических функций (3.7), которые конституенты выполняют по отношению друг к другу

3.6 (синтаксическая) дуга (syntactic edge, edge): Триплет, образуемый исходным узлом (3.12), целевым узлом и необязательными аннотациями (3.9)

ПРИМЕЧАНИЕ Нетерминальные узлы (3.13) имеют исходящую дугу синтаксической конституентности.

3.7 грамматическая функция (grammatical function): Грамматическая роль **словоформы** (3.24) или **конституенты** (3.4) в синтаксической среде, в которую они погружены

ПРИМЕЧАНИЕ Например, именная группа (NP) или имя существительное внутри сложного предложения может действовать как подлежащее (3.15) – соответственно положению глагола в графе отношения подчинения. Между именной группой как подлежащим и основным глаголом предложения существует грамматическая связь. Все грамматические отношения (подлежащее – сказуемое, вершина – модификатор и т.п.) категоризируются в соответствии с концептом отношения зависимости (3.5) между терминальными и нетерминальными узлами.

3.8 синтаксическая вершина, вершина, главное слово (syntactic head, head): Часть **конституенты** (3.4), определяющая её дистрибуцию (синтаксическое окружение, в котором может появляться конституента) и грамматические характеристики (например, если грамматический род главного слова – женский, то род конституенты в целом тоже будет женским)

ПРИМЕЧАНИЕ Опущение главного слова конституенты, как правило, не допускается.

3.9 (лингвистическая) аннотация (linguistic annotation, annotation): Пара "элемент - значение", представляющая лингвистическое свойство лингвистического сегмента

3.10 главное предложение (main clause): Предложение (3.3), которое само по себе может выступать в качестве законченного высказывания (3.15)

ПРИМЕЧАНИЕ В языках, предусматривающих различение завершённости и незавершённости действия, главное предложение обычно является законченным высказыванием; например: *Поезд опаздывает.*

3.11 модификатор, определение (modifier): Часть **конституенты** (3.4), описывающая свойство её вершины (3.8)

ПРИМЕЧАНИЕ Модификатор может помещаться до или после вершины фразы (3.14) (пре-модификатор или пост-модификатор). Модификаторы в конституенте не обязательны.

3.12 синтаксический узел (node, syntactic node): **Словоформа** (3.24) или **конституента** (3.4), рассматриваемая как элементарный синтаксический компонент синтаксического анализа

3.13 **нетерминальный узел** (non-terminal node): Синтаксический узел (3.12), не являющийся словоформой (3.24)

ПРИМЕЧАНИЕ Нетерминальный узел имеет исходящую дугу конституентности (3.6).

3.14 фраза, синтаксическая конструкция (phrase): Группа **словоформ** (3.24) (обычно состоящая из одного или нескольких слов), которая может выполнять определённую грамматическую функцию (3.7), например, в элементарном **предложении** (3.3)

ПРИМЕЧАНИЕ Допускается присутствие пустых фраз (представленных неопределённо-личными местоимениями); такие группы словоформ в английском языке иногда снабжаются пометой "pro" и в простых предложениях играют роль подлежащего). Группы словоформ, как правило, именуются по их главному слову, или вершине (3.8): например, могут быть, именные группы, глагольные группы, группы прилагательного, наречные группы и предложные группы. В просторечии фразы характеризуются как "раздутые слова", в том смысле, что части фразы, добавляемые к главному слову (вершине), усложняют и конкретизируют его референцию. В нашей модели фраза представляет собой специальный случай **составляющей** (3.4).

3.15 сложное предложение, высказывание (sentence): Связанная группа **словоформ** (3.24), содержащая предикацию, которая обычно выражает законченную мысль и образует базовую единицу структуры дискурса.

ПРИМЕЧАНИЕ Сложное предложение состоит из одного или нескольких простых предложений (3.3). При описании речевого общения обычно говорят о "высказываниях", а не о предложениях.

3.16 интервал (span): Пара точек (p1, p2), где p1 ≤ p2, идентифицирующая сегмент документа, к которому применима аннотация (3.9)

ПРИМЕЧАНИЕ Многократный интервал – это цепочка интервалов, в которой координаты конечной точки каждого предшествующего интервала меньше или равны координатам начальной точки последующего интервала.

3.17 придаточное предложение (subordinate clause): Элементарное предложение, которое выполняет некоторую **грамматическую функцию** (3.7) в синтаксическом обороте (3.14) [например, функцию определительного **предложения** (3.3) для имени существительного, образующего вершину (3.8) именного словосочетания] или в другом предложении

ПРИМЕЧАНИЕ Придаточное предложение обычно не самостоятельно, а является частью более длинного сложного предложения.

3.18 фрейм субкатегоризации (subcategorization frame): Набор ограничений, показывающих свойства синтаксических аргументов (3.19), которые могут или должны связываться с глаголом

ПРИМЕР – Альфред (/syntacticArgument/) читает книгу (/syntacticArgument/) сегодня (/adjunct/).

ПРИМЕЧАНИЕ Подлежащее, косвенное дополнение и прямое дополнение – это субкатегоризированные **грамматические функции** (3.7) внутри предложения; они подчиняются глаголу (то есть могут появляться во фреймах субкатегоризации).

3.19 синтаксический аргумент (syntactic argument): Важный функциональный элемент, запрашиваемый и интерпретируемый вершиной его **синтаксической конструкции** (3.14) или **узлом** (3.12), от которого он зависит (примером может служить именной аргумент предложной группы или глагол)

ПРИМЕЧАНИЕ Для глаголов и глагольных конструкций аргументы идентифицируют стороны процесса, на который указывает глагол. В некоторых объектных структурах синтаксические аргументы называются дополнениями.

3.20 (синтаксический) граф (syntactic graph, graph): Связанное множество **синтаксических узлов** (3.12) и **дуг** (3.6)

3.21 синтаксическое дерево (syntactic tree): Синтаксический **граф** (3.20), в котором каждый из узлов имеет единственный родительский узел

3.22 синтаксис, синтаксические правила (syntax): Способ соединения и/или группирования **словоформ** (3.24) в синтагмы для сбора информации о существующих отношениях между группируемыми единицами

3.23 терминальный узел (terminal node): **Синтаксический узел** (3.12), являющийся одиночной **словоформой** (3.24) или пустым элементом синтаксического отношения

3.24 словоформа (word form): Непрерывный или сегментированный объект речевого или текстового оборота, идентифицируемый как автономная лексема

4 Метамодель SynAF

4.1 Вводные замечания

В когнитивной обработке языковых данных синтаксические аннотации выполняют как минимум две функции:

1. представление лингвистической конституентности [подобно именным группам (NP)], описывающей структурированную последовательность морфосинтаксически аннотированных лексем (включая пустые элементы или следы, порождённые передвижениями на уровне составляющих), а также построение составляющих из сегментированных элементов;

2. представление отношений зависимости: например, отношения "главное слово - модификатор" и отношения между категориями одного вида (подобные связям между главными словами в именных аппозициях или именным соподчинениям в некоторых формализмах). Внутри синтаксической группы может существовать информация о зависимости между элементами, прошедшими этап морфосинтаксического аннотирования (например, прилагательное – это модификатор главного существительного внутри именной группы) или описываться конкретное отношение между синтаксическими составляющими на клаузальном и пропозициональном уровнях (то есть там, где именная группа выступает как "субъект" основного глагола элементарного или сложного предложения). Отношение зависимости может устанавливаться также для пустых элементов (например, для элемента pro в романских языках, где этот элемент выполняет грамматическую функцию).

Как следствие, синтаксические аннотации должны соответствовать многоуровневой стратегии аннотирования, обеспечивающей взаимосвязь синтаксического аннотирования по составляющим элементам и по отношениям зависимости, как это установлено в метамодели SynAF.

4.2 О метамодели SynAF

4.2.1 Общий обзор

Метамодель SynAF представляется как совокупность классов универсального языка моделирования UML, дополненная UML-парами "атрибут - значение", которые представляют соответствующие категории синтаксических данных. Текстовые описания SynAF определяют более полную информацию о классах SynAF, отношениях и расширениях, которые могут быть включены в диаграмму UML. Разработчики должны определить выбор категории данных (DCS) в соответствии с процедурами выбора категорий данных, установленными для SynAF (см. Рисунок 1). Для представления синтаксических аннотаций должны использоваться категории данных, указанные в приложении А.

4.2.2 Класс SyntacticNode

SyntacticNode – это параметризованный класс, категоризирующий как класс терминальных узлов, так и класс нетерминальных узлов. Синтаксические узлы могут быть задействованы в любом необходимом числе синтаксических отношений (см. п. 3.6, **синтаксические дуги**).

4.2.3 Класс T_Node

Класс *T_Node* представляет терминальные узлы синтаксического дерева, состоящего из словоформ, прошедших этап морфосинтаксического аннотирования, а также из пустых элементов, когда они необходимы. Узлы этого класса определяются на одном *интервале* или на множестве интервалов (множественные интервалы обеспечивают учет нарушений непрерывности составляющих частей текста). Для аннотирования узлов T_Nodes используются средства автоматической синтаксической категоризации, действующие на уровне отдельных слов.

4.2.4 Класс NT_Node

Класс *NT_Node* представляет нетерминальные узлы синтаксического дерева. Синтаксические деревья состоят в основном из узлов T_Nodes и NT_Nodes, а также пустых элементов, когда они необходимы. Узлы T_Nodes make reference to a span. Так с помощью древовидного синтаксического представления могут быть получены интервалы и для NT_Nodes. Для аннотирования узлов NT_Nodes используются средства автоматической синтаксической категоризации, действующие на уровне фраз и на более высоких уровнях (клаузальном и сентенциальном).

4.2.5 Класс SyntacticEdge

Класс *SynacticEdge* представляет отношение между синтаксическими узлами (как терминальными, так и нетерминальными). Например, отношение зависимости – это бинарное отношение, образуемое парой узлов – исходным и целевым, с одной или большим числом аннотаций.

В частности, синтаксическая дуга может аннотироваться по типу /syntacticEdgeType/ (см. приложение А), концептуальной областью которого может быть одна из двух дуг: /primarySyntacticEdge/ либо /secondarySyntacticEdge/, но не только эти дуги.

4.2.6 Класс Annotation

Класс *Annotation* представляет результат применения синтаксической информации к аннотированным данным SynAF, а также (см. Рисунок 1) применение морфосинтаксической информации к данным, прошедшим этап морфосинтаксического аннотирования (MAF).

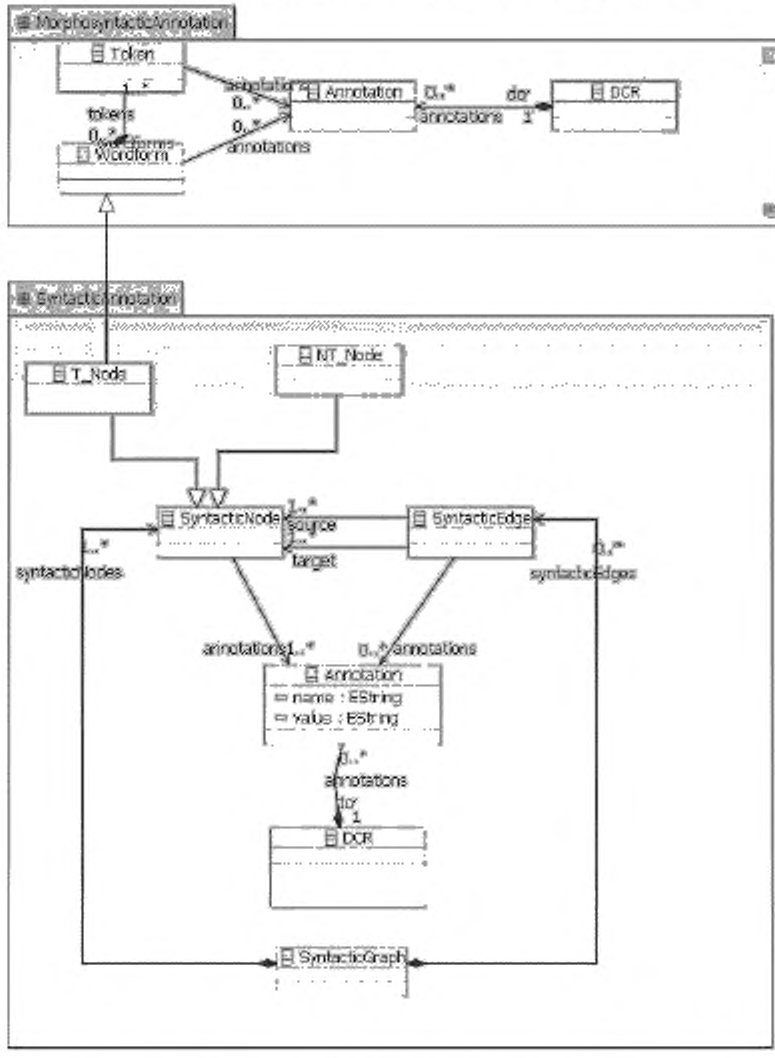

Рисунок 1 — Метамодель SynAF (скомпонованная средствами MAF)

Приложение
(обязательное)

Категории данных для метамодели SynAF

А.1 Общие положения

Приведённые ниже категории данных должны использоваться для представления синтаксических аннотаций в сочетании с метамоделью SynAF. При необходимости в конкретных приложениях могут определяться дополнительные категории данных, которые должны описываться в соответствии с требованиями стандарта ИСО 12620 и регистрироваться в реестре категорий данных ISOCat.

А.2 Базовые категории синтаксических данных

/annotation/

Определение– аннотация – информация, добавляемая к слову, фразе, элементарному предложению, сложному предложению, тексту или к связывающему их отношению

/annotationDepth/

Концептуальная область– /deepParsing/, /shallowParsing/, /tagging/

Определение– глубина аннотирования – уровень информационной содержательности, описываемый аннотацией

/annotationStyle/

Концептуальная область– /embeddedNotation/, /mixedNotation/, /standoffNotation/

Определение– стиль аннотации

/annotationType/

Концептуальная область– /constituency/ /constituencyAndDependency/ /dependency/

Определение– тип аннотации

/clitic/

Определение– клитика – безударное слово, которое не может само по себе служить обычным высказыванием, и произношение которого находится в фонологической зависимости от соседнего слова

ПРИМЕЧАНИЕ– Существует большое разнообразие клитик. Иногда в английском языке клитизируемые формы ограничиваются усеченными формами вспомогательных конструкций, таких как I'm, she'll и т.п. Однако в некоторых случаях к клитикам относятся также артикли.

/constituency/

Определение– конституентность – механизм, позволяющий соединять слова во фразы, фразы – в более сложные синтаксические конструкции либо в предложения и предложения – в высказывания

ПРИМЕЧАНИЕ– Построение текста из высказываний обычно не считается конституентностью.

/constituencyAndDependency/

Определение– соединение свойств конституентности и зависимости

/contiguous/

Определение– соседство – способность грамматической единицы разделять границу с другой единицей

/deepParsing/

Определение– глубокий синтаксический анализ – процесс полного декодирования предложений и отношений, присутствующих в высказывании

/dependency/

Определение– зависимость – механизм, позволяющий связывать слова или (в некоторых формализмах) фразы и предложения на основе использования бинарного разделения, зависящего от главного слова, и возможного аннотирования грамматической функции

/doubleNegation/

Определение– двойное отрицание – конструкция, содержащая две отрицательные формы в одном предложении

ПРИМЕЧАНИЕ– Пример для английского языка: "I'm not unhappy" (Я не несчастлив).

/embeddedNotation/

Определение– вложенная аннотация – аннотация, добавленная в текст

ПРИМЕЧАНИЕ– Исходная организация текста изменяется.

/enclitic/ - ВС: /clitic/

Определение– клитика – элемент, зависящий от предшествующего слова

/first/

Определение– первый – метка, которая ставится перед любой упорядоченной структурой

/mixedNotation/

Определение– смешанная нотация – аннотация гибридного стиля, при котором перемежаются вольный стиль и вложенные структуры

/morphosyntacticAnnotation/ - ВС: /annotation/

Определение– морфосинтаксическая аннотация – аннотация, связанная с морфологией слов и их частей речи

/negation/

Определение– отрицание – конструкция, выражающая ограничение, накладываемое на некоторые или на все значения высказываний, слов или фраз

ПРИМЕЧАНИЕ– Отрицание может представляться отрицательными частицами (например, "not") или префиксами (например, "un" или "non"). Пример для английского языка: "I'm not happy".

/next/

Определение– непосредственно следующий

/primarySyntacticEdge/

Определение– первичная синтаксическая дуга – синтаксическая дуга, выбираемая по умолчанию и выражающая отношение конституентности; исходит из конституенты и заканчивается в компоненте этой конституенты

/predicate/

Определение– сказуемое – фраза или слово в предложении, которые порождают высказывание, относящееся к подлежащему этого предложения. Большинство предложений может таким образом разбиваться на подлежащее и сказуемое, и в этом разбиении сказуемое является функцией, которая распространяется на подлежащее.

ПРИМЕЧАНИЕ– Пример: высказывание "Kevin kicks the ball" рассматривается как подлежащее ("Kevin"), ассоциируемое с группой сказуемого ("kicks the ball").

/previous/

Определение– непосредственно предшествующий

/proclitic/ - ВС: /clitic/

Определение– проклитика – клитика, зависящая от последующего слова

ПРИМЕЧАНИЕ– Пример: артикль "the" в словосочетании "the boy".

/propagation/

Определение– акт распространения лингвистического свойства одной грамматической единицы а другую

/secondarySyntacticEdge/

Определение– вторичная синтаксическая дуга – неориентированная дуга, выражающая синтаксическую конституентность. Такие дуги могут использоваться для выражения отношений между вершиной древовидной синтаксической структуры и механизмом перекрестных ссылок на ее опущенный подчинённый узел.

ПРИМЕЧАНИЕ– Пример для английского языка: в высказывании "I saw Bill, but went straight back home afterwards" ("Я увидел с Биллом, но потом сразу пошёл домой"), слово "I" может служить явно выраженным подлежащим для первого элементарного предложения, главенство которого представляется первичной синтаксической дугой; но во втором элементарном предложении следующая, вторичная синтаксическая дуга, ведущая к узлу "I", может сделать совершенно ясным, что "I" – это подлежащее также и для второго элементарного предложения, хотя оно в явном виде не присутствует ни в какой из его частей, доминируемых первичными дугами. Такой механизм используется в некоторых формальных представлениях для того, чтобы избежать введения пустых элементов, замещающих "опущенные" носители грамматической функции.

/shallowParsing/

Определение– поверхностный синтаксический анализ – процесс идентификации частей предложения

/standoffNotation/

Определение– аннотация, которая записывается вне древовидной структуры грамматическими единицами и содержит ссылки на них

ПРИМЕЧАНИЕ– Исходная организация текста остаётся неизменной.

/syntacticAnnotation/ - ВС: /annotation/

Определение– синтаксическая аннотация – аннотация, описывающая отношения конституентности и/или зависимости

ПРИМЕЧАНИЕ– синтаксическая аннотация не работает непосредственно со смысловым значением высказывания

/syntacticFeature/

Определение– синтаксическое свойство – свойство, используемое в описании синтаксических правил языка

/syntacticEdgeType/

Концептуальная область– /primarySyntacticEdge/, /secondarySyntacticEdge/

Определение– тип синтаксической дуги – характеризует синтаксическую дугу в соответствии с её ролью в синтаксическом представлении

/syntacticRestriction/

Определение– синтаксическое ограничение – правило, которое ограничивает возможности синтаксической структуры по сравнению с теми, которые она предоставляет в конкретном языке

/tagging/

Определение– процесс аннотирования части речи для каждого слова

/whType/

Определение– свойство предложения, начинающегося с вопросительного слова

ПРИМЕЧАНИЕ– В английском языке вопросительное предложение "who is he?" относится к типу whType.

/yesNoType/

Определение– свойство высказывания, на которое возможен только положительный либо отрицательный ответ, или подтверждение либо отрицание

ПРИМЕЧАНИЕ– В английском языке вопросительное предложение "Are you coming?" относится к типу yesNoType.

А.3 Категории данных, относящиеся к свойству конституентности

/adjectiveChunk/ - ВС: /chunk/

Определение– адъективный фрагмент – фрагмент, начинающийся с имени прилагательного

/adjectivePhrase/ - ВС: /phrase/

Определение– адъективная группа – синтаксическая группа, начинающаяся с имени прилагательного

/adpositionChunk/ - ВС: /chunk/

Определение– аппозиционный фрагмент – фрагмент, содержащий один либо несколько предлогов или послелогов, которые не обязательно смежны, и не обязательно находятся на одной и той же стороне фрагмента

/adpositionPhrase/ - ВС: /phrase/

Определение– аппозиционная группа – синтаксическая группа, содержащая один либо несколько предлогов или послелогов и содержащая дополнение, например, в виде именной группы

ПРИМЕЧАНИЕ– Предлоги и послелоги не обязательно смежны и не обязательно находятся на одной и той же стороне фразы.

/adverbChunk/ - ВС: /chunk/

Определение– наречный фрагмент – фрагмент, начинающийся с наречия

/adverbPhrase/ - ВС: /phrase/

Определение– наречная группа – фраза, начинающаяся с наречия

/chunk/ - ВС: /grammaticalUnit/

Определение– фрагмент – горизонтальная последовательность слов, обычно содержащая больше одного слова

ПРИМЕЧАНИЕ– Фрагмент не может содержать никаких подструктур; часто он подобен фразе и в большинстве случаев неразрывен.

/clause/ - ВС: /grammaticalUnit/

Определение– клауза – единица грамматической организации, которая меньше или равна предложению, но больше чем фразы и слова, и обычно имеет собственное подлежащее

ПРИМЕЧАНИЕ– Предложения традиционно классифицируются как главные (независимые или соподчинённые) и придаточные (или зависимые). Пример в английском языке: the boy arrived (главное предложение) after the rain started (придаточное предложение). Предложение может быть законченным выражением, как, например, "they came". Выражения могут состоять из подвыражений.

/comparativePhrase/ - ВС: /phrase/

Определение– сравнительный оборот

ПРИМЕЧАНИЕ– В английском языке для выражения сравнения существуют конструкции финитной группы (например, larger) и сопоставительной группы (например, more beautiful).

/coordinatedPhrase/ - ВС: /phrase/

Определение– синтаксическая группа, выражающая отношение сочинения

/declarativeClause/ - BC: /clause/
Определение– повествовательное предложение – предложение, содержащее условно истинное высказывание

ПРИМЕЧАНИЕ– Обычно этот термин используется в противоположность определениям "вопросительный" и "повелительный".

/grammaticalUnit/
Определение– грамматическая единица – термин, относящийся к слову, фразе, предложению или высказыванию
/imperativeClause/ - BC: /clause/
Определение– повелительное предложение, выражающее побуждение к действию, указание, команду

ПРИМЕЧАНИЕ– Обычно этот термин используется в противоположность определениям "вопросительный" и "повествовательный".

/interrogativeClause/ - BC: /clause/
Определение– вопросительное предложение

ПРИМЕЧАНИЕ– Обычно этот термин используется в противоположность определениям "повествовательный" и "повелительный". Пример в английском языке: "who are you?"

/nounChunk/ - BC: /chunk/
Определение– именной фрагмент – фрагмент с именем существительным в вершине дерева
/nounPhrase/ - BC: /phrase/
Определение– именная фраза – фраза с именем существительным в вершине дерева
/phrase/ - BC: /grammaticalUnit/
Определение– синтаксическая группа, фраза – структурный элемент, построенный на основе главного слова, определяющего грамматические свойства элемента, и состоящий из нуля, одного или большего числа слов и/или других синтаксических групп; не имеет характерной для предложения субъектно-предикатной структуры.

ПРИМЕЧАНИЕ– Фраза может содержать вложенные подструктуры. Традиционно рассматривается как часть иерархической структуры, занимающая промежуточное положение между предложением и словом. Обычно выделяются несколько типов фраз (синтаксических групп): наречная группа, группа прилагательного и др. – в зависимости от главного слова.

/postpositionChunk/ - BC: /chunk/
Определение– фрагмент с послелогом в вершине дерева
/postpositionPhrase/ - BC: /phrase/
Определение– фраза с послелогом в вершине дерева
/prepositionChunk/ - BC: /chunk/
Определение– фрагмент с предлогом в вершине дерева
/prepositionPhrase/ - BC: /phrase/
Определение– фраза с предлогом в вершине дерева
/prepositionVerbPhrase/ - BC: /phrase/
Определение– глагольно-предложная группа – глагольная группа, представленная предлогом
/relativeClause/ - BC: /clause/
Определение– определительное придаточное предложение – предложение выполняющее роль определения для именной группы, представленной относительным местоимением, и могущее быть эллиптическим.

ПРИМЕЧАНИЕ– В английском языке определительное придаточное предложение представляется относительным местоимением, например, таким как "who". Определительные придаточные предложения могут быть ограничительными, когда они определяют охватываемую именную подгруппу, или не ограничивающими, просто добавляющими определение. Примером в английском языке может служить следующая пара предложений: "the men who were fighting were brave" (а не участвовавшие в сражении не были храбрыми) и "the men, who were fighting, were brave" (все люди были храбрыми и в свое время участвовали в сражении)

/sentence/ - BC: /grammaticalUnit/
Определение– грамматическая организация, при которой имеется одно главное предложение и все относящиеся к нему придаточные предложения с рекурсивной последовательностью их придаточных предложений.

ПРИМЕЧАНИЕ– Предложения могут разделяться по типам на простые и сложные, то есть состоящие из одной субъектно-предикатной единицы или из нескольких таких единиц.

/superlativePhrase/ - ВС: /phrase/
Определение– фраза, выражающая значение превосходной степени

ПРИМЕЧАНИЕ–Для выражение превосходной степени а английском языке существуют конструкции финитной группы (например, largest) и сопоставительной фразовой группы (например, the most interesting).

/syntacticConstituent/
Определение– синтаксическая конституента – грамматическая единица, образующая часть более крупной грамматической единицы, является составляющей (конституентой) этой более крупной единицы. Если две грамматические единицы соединены напрямую синтаксической дугой конституентности, то имеет место непосредственная составляющая; в противном случае (при отсутствии такой дуги) речь идёт о косвенной составляющей.
/verbNucleus/ - ВС: /chunk/
Определение– глагольное ядро – фрагмент, образуемый обособленным глаголом и, возможно, ассоциируемый с его клитиками
/verbPhrase/ - ВС: /phrase/
Определение– глагольная группа – синтаксическая группа с глаголом в вершине древовидного представления

А.4 Категории данных, связанные с отношениями

/adjectiveModifier/ - ВС: /adjectiveModifier/
Определение– адъективный модификатор – отношение, в котором модификация осуществляется с помощью имени прилагательного
/adjunct/
Определение– адъюнкт, обстоятельственное слово – необязательная либо второстепенная грамматическая единица, которая может быть удалена без нарушения грамматических правил относительно остальных частей языковой конструкции

ПРИМЕЧАНИЕ–Обстоятельственными словами обычно бывают наречия, как, например, в предложении "Peter kicked the ball yesterday". После удаления обстоятельственного слова предложение "Peter kicked the ball" сохраняет свою грамматическую правильность.

/adverbModifier/ - ВС: /modifier/
Определение– обстоятельство – отношение, в котором модификации подвергается наречие
/apposed/
Определение– аппозиционный – свойство, порождаемое нахождением лингвистической элемента в окружении предлогов и послелогов
/apposition/
Определение– приложение – отношение между лингвистическими единицами, которые разделяют один и тот же (или похожий) объект ссылки и одну и ту же грамматическую функцию в одном и том же предложении таким образом, что одна единица расширяет смысл другой

ПРИМЕЧАНИЕ–Пример для английского языка: "Smith, the barber, came in" ("Пришёл Смит, парикмахер").

/attribute/
Определение– атрибут – отношение, связывающее имя прилагательное или имя существительное, когда они выступают в роли модификатора центрального слова именной группы
/auxiliary/
Определение– конструкция со вспомогательным глаголом – отношение, связывающее вспомогательный и основной глаголы
/comparativeRelation/ - ВС: /relation/
Определение– сопоставительное отношение – отношение, выражающее результат выполняемой процедуры сравнения значений
/complementizer/
Определение– комплементайзер – отношение между подчинительным союзом и глаголом, отмечающим вложенное дополнительное придаточное предложение

ПРИМЕЧАНИЕ–Например, в предложении "I said that he was leaving" ("Я сказал, что он уходит") вложенное предложение – это "he was leaving", а подчинительный союз – "that".

/coordination/
Определение– сочинительная связь – отношение, связывающее "равноправные" лингвистические единицы с эквивалентным синтаксическим статусом; примером может служить последовательность фраз или слов, возможно, соединённых союзом
/coordinator/
Определение– координатор – слово или последовательность слов, осуществляющие сочинительную

связь в сложном предложении

ПРИМЕЧАНИЕ – Как правило, координатором является сочинительный союз.

/directObject/

Определение – прямое дополнение – отношение между глаголом и его аргументом, в котором аргумент не отделён предлогом или послелогом и не отмечен косвенным падежом.

ПРИМЕЧАНИЕ – В предложении "the man gave the boy a book" словосочетание "a book" является прямым дополнением, тогда как "the boy" во многих формальных конструкциях трактуется как косвенное дополнение – по причине предполагаемого (и издавна реально наблюдаемого) косвенного падежа (датива).

/genitive/ - ВС: /relation/

Определение – родительный падеж, генитив – отношение, обычно выражающее принадлежность или иную приимённую связь модифицирующего существительного

/head/

Определение – вершина – центральное слово фразы, определяющее её дистрибуцию и грамматические свойства

/introducer/

Определение – маркер элемент – вводный лингвистический элемент (например, словоформа), отмечающий начало синтаксической группы

ПРИМЕЧАНИЕ – Например, предлог является маркером именной группы.

/juxtaposition/

Определение – непосредственное соседство – отношение, при котором две лингвистические единицы следуют совместно

/leftCoordinated/

Определение – элемент с левосторонней координацией – член предположения, стоящий перед координатором

/modifier/

Определение – модификатор – отношение между грамматическими единицами, при котором одна из них зависит от другой и расширяет ее.

ПРИМЕЧАНИЕ – В предложении "the big tree in the garden" обе грамматические единицы – "the big" и "in the garden" модифицируют единицу "tree".

/nounModifier/ - ВС: /modifier/

Определение – именной модификатор – отношение, при котором модифицируется имя существительное

/postnominalModifier/ - ВС: /modifier/

Определение – пост-именной модификатор – модификатор, стоящий после имени существительного

/prenominalModifier/ - ВС: /modifier/

Определение – предыменной модификатор – модификатор, стоящий перед именем существительным

/prepositionModifier/ - ВС: /modifier/

Определение – предложный модификатор – отношение, при котором модифицируемым элементом является предлог

/relation/

Определение – отношение – аннотированная связь между двумя и более грамматическими единицами

/relativeRelation/ - ВС: /relation/, /nounModifier/

Определение – отношение, связывающее определительное придаточное предложение с именной группой, на которую дается ссылка

/rightCoordinated/

Определение – элемент с правосторонней координацией – член предложения, стоящий справа от координатора

/structureHead/

Определение – идентификатор вершины структурного представления – двоичное значение, указывающее, является ли компонент главным синтаксическим звеном

/subject/

Определение – подлежащее – отношение между синтаксической группой и глаголом, представляющее одушевлённый или неодушевленный предмет, о котором говорится в предложении

/superlativeRelation/ - ВС: /relation/

Определение – отношение, выражающее значение превосходной степени

/syntacticArgument/

Определение – аргумент синтаксической конструкции

/syntacticFunction/

Определение – отношение между лингвистической формой и другими частями лингвистической

системы, в которой эта форма используется

ПРИМЕЧАНИЕ– Пример – подлежащее.

/syntacticHead/

Определение– вершина синтаксической структуры – элемент, который определяет дистрибуцию и грамматические свойства лексической единицы, состоящей из самого этого элемента и его зависимых узлов

/verbComplement/

Определение– глагольное дополнение – отношение между фразой и глаголом, где данная фраза не является для глагола центральным звеном

ПРИМЕЧАНИЕ– Противоположность прямому дополнению.

/verbModifier/ - ВС: /modifier/

Определение– отношение, представляющее модификацию глагола

Приложение Б
(справочное)

Связь с системой лингвистического аннотирования

Для реализации системы SynAF необходимо руководствоваться требованиями стандарта ИСО 24612 к системе лингвистического аннотирования (LAF – Linguistic Annotation Framework). LAF обеспечивает общую основу для представления аннотаций, описанную в работах Айда (Ide) и Ромари (Romary) [14, 15, 16]. Разработка этой основы строилась на достижениях сложившейся практики и обобщении различных принципов лингвистического аннотирования, использовавшихся на протяжении последних 15-20 лет. Ядром системы является спецификация опорной абстрактной модели, из которой получаются аннотации, ориентированные на конкретные цели информационного обмена.

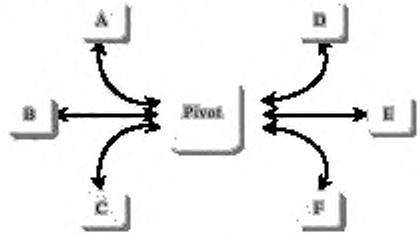

Рисунок В.1 — LAF как базовый формат

Рисунок Б.1 иллюстрирует принципы использования LAF применительно к шести различным форматам пользовательских аннотаций (с метками от A до F), которые требуют двух преобразований для каждой схемы: одно — для отображения в базовый формат и одно – для преобразования из базового формата, представленного разработчиком схемы. Следовательно, максимальное число схемных преобразований составит $2n$, в отличие от ($n^2 - n$) преобразований при отсутствии базового формата.

Для преобразования к опорной модели схема аннотирования должна быть изоморфна абстрактной модели (или сделана таковой в процессе преобразования); эта модель состоит, во-первых, из эталонной структуры для привязки внешних аннотаций к первичным данным, представленной ориентированным графом (орграфом), и, во-вторых, из представления структуры свойств содержимого аннотации. Таким образом, аннотация представляет собой орграф ссылок на n-мерные области первичных данных, равно как и на другие аннотации; в этом графе узлы аннотируются структурами элементов, которые, в свою очередь, формируют содержимое (контент) аннотации. Формально LAF включает в себя следующие компоненты:

– модель данных для аннотаций, основанную на вышеуказанных орграфах и определённую как граф аннотаций: граф аннотаций G – это совокупность вершин $V(G)$ (термин "вершина" является синонимом термина "узел") и множества дуг $E(G)$. Вершины и дуги могут снабжаться одним или несколькими свойствами. Свойство определяется четвёркой элементов (G', VE, K, V), где G' – это граф, VE – вершина в G', K – имя свойства и V – значение свойства;

– *первичные данные с базовой сегментацией*, которая определяет дуги, идущие от одного виртуального узла к другому и расположенные, между "символами" первичных данных, где символ определяется как смежная байтовая последовательность конкретной длины (по умолчанию для текста принимается значение UTF-16). Результирующий граф G трактуется как *рёберный граф* G', узлами которого являются дуги графа G и который представляется листовыми ("стоковыми") вершинами. Эти вершины образуют основу для однослойного или многослойного аннотирования. Над первичными данными может определяться множество сегментаций, а одна и та же сегментация может соотноситься с множественными аннотациями;

– публикации модели данных, одна из которых обозначается как опорная;

– методы манипулирования моделью данных.

Следует иметь в виду, что LAF не порождает спецификаций для категорий содержимого аннотаций (то есть аннотаций, описывающих соответствующие лингвистические явления); стандартизация таких аннотаций требует гораздо более сложных разработок. Архитектура LAF ориентирована на взаимодействие с реестром категорий данных [*Data Category Registry* (DCR)], содержащим предопределённые элементы данных и схемы, которые могут использоваться непосредственно в аннотациях вместе со средствами определения новых категорий и модификации существующих [14, 15].

Приложение ДА
(справочное)

Сведения о соответствии ссылочных международных стандартов ссылочным национальным стандартам Российской Федерации

Сведения о соответствии ссылочных международных стандартов ссылочным национальным стандартам Российской Федерации приведены в таблице ДА.1.

Таблица ДА.1

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ИСО 1087-1:2000	–	*
ИСО 1087-2:2000	–	*
ИСО 12620:2009	–	*
ИСО 24611:2012	–	*

* Соответствующий национальный стандарт отсутствует. До его утверждения рекомендуется использовать перевод на русский язык данного международного стандарта. Перевод данного международного стандарта находится в Федеральном информационном фонде технических регламентов и стандартов.

Библиография

[1] ИСО 639-1:2002, Коды для представления названий языков. Часть 1. Двухбуквенный код

[2] ИСО 639-2:1998, Коды для представления названий языков. Часть 2. Трехбуквенный код

[3] ИСО 639-3:2007, ISO 639-3:2007, Коды для представления названий языков. Часть 3. Трехбуквенный код для всестороннего охвата языков

[4] ИСО/МЭК 10646-1:2000, Информационные технологии. Универсальный многооктетный набор кодированных знаков. Часть 1: Архитектура и основная многоязычная матрица

[5] ИСО/МЭК 11179-3:2003, Информационные технологии. Реестры метаданных (MDR). Часть 3: Метамодель системного регистра и базовые атрибуты

[6] ИСО 24610-1:2006, Управление языковыми ресурсами. Структуры элементов. Часть 1. Представление структуры элементов

[7] ИСО 24612, Управление языковыми ресурсами. Система лингвистической аннотации

[8] ИСО 24613:2008, Управление лингвистическими ресурсами. Схема лексической разметки

[9] Abeillé, A. (ed.) Building and Using Syntactically Annotated Corpora. Kluwer, Dordrecht, 2001

[10] Abeillé, A., Hansen-Schirra, S. and Uszkoreit, H. (eds.), Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), 2003

[11] Calzolari, N., McNaught, J. and Zampolli, A. (eds). *EAGLES: Introduction, 1996.* gost.e-ecolog.ru

[12] Francopoulo, G., Declerck, T., Sornlertlamvanich, V., de la Clergerie, E. and Monachini, M. 2008. Data Category Registry: Morpho-syntactic and Syntactic profiles, *LREC Workshop on use and usage of language resource-related standards*

[13] Ide, N. and Romary, L. A common framework for syntactic annotation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France, July 6-11, 2001. Association for Computational Linguistics, Morristown, NJ, 2001, pp. 306-313. DOI= gost.e-ecolog.ru

[14] Ide, N. and Romary, L. A Registry of Standard Data Categories for Linguistic Annotation. *Proceedings of the Fourth Language Resources and Evaluation Conference* (LREC), Lisbon, 2004, pp. 135-139

[15] Ide, N. and Romary, L. International Standard for a Linguistic Annotation Framework. *Journal of Natural Language Engineering*, **10**:3-4, 2004, pp. 211-225

[16] Ide, N. and Romary, L. Representing Linguistic Corpora and Their Annotations. *Proceedings of the Fifth Language Resources and Evaluation Conference* (LREC), Genoa, Italy, 2006

[17] Ide, N. *GrAF: A Graph-based Format for Linguistic Annotations*. Proceedings of the LAW Workshop at ACL 2007, Prague, 2007

[18] Montemagni, F. *et al*. Building the Italian Syntactic-Semantic Treebank. In: *Building and using Parsed Corpora* (ed. Abeillé, A.), Language and Speech series, Kluwer, Dordrecht, 2003

[19] Rumbaugh, J., Jacobson, I. and Booch, G. *The Unified Modeling Language Reference Manual*, 2nd edition. Addison Wesley, 2004

[20] Веб-сайты проекта:
Группа EAGLES Initiative: gost.e-ecolog.ru
Проект LIRICS: gost.e-ecolog.ru
Проект SPARKLE: gost.e-ecolog.ru
Проект TIGER: gost.e-ecolog.ru TIGER/
TIGER Corpus/

УДК 001.4:006.354 ОКС 01.020

Ключевые слова: менеджмент языковых ресурсов, синтаксическое аннотирование, категории данных, терминологическая работа

Подписано в печать 01.04.2014. Формат 60x84 1/8.

Усл. печ. л. 2,79. Тираж 31 экз. Зак. 1102.

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