

*О.А. Казакевич, Ю.Е. Галямина, Е.Л. Клячко, Е.Л. Рудницкая,
Е.М. Будянская
O.A. Kazakevich, Ju.E. Galyamina, E.L. Klyachko, E.L. Rudnitskaya,
E.M. Budyanskaya*

СИНТАКСИЧЕСКАЯ РАЗМЕТКА ДЛЯ КОРПУСОВ
СЕЛЬКУПСКИХ, ЭВЕНКИЙСКИХ И КЕТСКИХ ТЕКСТОВ¹
SYNTACTIC ANNOTATION FOR SELKUP, EVENKI AND KET
TEXT CORPORA

Аннотация: В основе разметки лежит грамматика непосредственных составляющих. Рассматриваются основные синтаксические отношения в синтаксическом дереве. Излагаются принципы сегментации текста, обозначения узлов, базовая структура предложения и др. Разметка отражает базовый порядок слов в рассматриваемых языках и отклонения от него. Показывается лингвистическая обоснованность разметки.

Ключевые слова: синтаксическая разметка, корпус, селькупский язык, эвенкийский язык, кетский язык

Abstract: The annotation is based on the immediate constituent analysis for sentence structure. Main syntactic relations in the syntactic tree are introduced. Principles of text segmentation, of tree nodes notation, of the base clause structure, etc. are outlined. The annotation accounts for the basic word order in the languages considered and for deviations from it. It is shown that the annotation is linguistically well grounded.

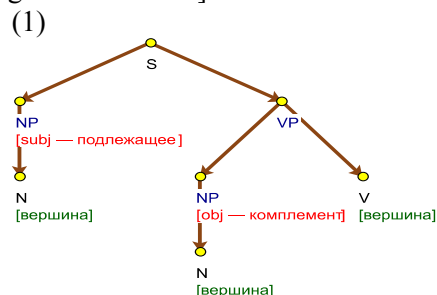
¹ Доклад подготовлен в рамках проекта «Корпусное исследование синтаксических структур в устной и письменной речи носителей исчезающих языков (на материале селькупских, эвенкийских и кетских говоров бассейнов Среднего Енисея и Среднего и Верхнего Таза)», грант РФФИ 14-06-00449.

Keywords: syntactic structure, corpus, Selkup, Evenki, Ket

1. Введение

Работа над созданием аннотированных корпусов селькупских, кетских и эвенкийских говоров ведется в лаборатории автоматизированных лексикографических систем НИВЦ МГУ им. М.В. Ломоносова уже около 10 лет, однако до недавнего времени речь шла в основном о морфологическом аннотировании². В 2014 г. мы начали работу над синтаксической аннотацией наших корпусов.

В соответствии с [Kernes 1980: 604-605], наш формализм можно определить как основанную на грамматике непосредственных составляющих аннотацию не с частичным, а с полным синтаксическим анализом (то есть анализом, при котором единица членения текста — это всегда предложение, а не меньшая составляющая, см. п. 3). Наша разметка не только информативна и понятна читателю, но и (как нам представляется) лингвистически обоснована. В качестве исходной использовалась следующая схема разворачивания фразовой структуры (1), ср. [Haegeman 1994: 34]³.



² См., напр., [Казакевич, Клячко 2013].

³ Отправная точка для (1) — ранние версии генеративной грамматики, например, [Chomsky 1957].

В (1) представлены только те правила развертывания структуры, которые связаны с основными компонентами предложения: подлежащим, сказуемым и прямым дополнением⁴. При построении синтаксической структуры (схемы предложения, названия узлов, расположение графов) мы используем чисто синтаксические характеристики – порядок слов в предложении, состав именных и глагольных конструкций. Морфологическая информация дается в строке глоссирования⁵, которая приводится вместе со словом, соответствующим терминальному узлу, см. *wa:-rə-n* ‘убивать-NFUT-3SG’ в (3)⁶.

Технически разметка осуществляется при помощи TrEd – инструмента, разработанного Карловым университетом в Праге⁷. Выбор TrEd обусловлен тем, что он позволяет:

⁴ Как отмечается в работе [Brants et. al.: 76-77], такие принципы наглядно отражают предикатно-аргументную структуру предложения и также снижают вероятность синтаксической неоднозначности – см. п. 3. Правила присоединения адъективных и адвербиальных модификаторов рассматриваются отдельно также в п. 3.

⁵ Особенность нашей синтаксической разметки в том, что она по инструментарию полностью разведена с морфологической – такое соотношение нетипично для аннотаций, основанных на непосредственных составляющих. О представлении кетского материала см. ниже.

⁶ Отгlossированные тексты на всех трех языках см. на сайте «Малые языки Сибири: наше культурное наследие»; URL: <http://siberian-lang.srcc.msu.ru/> (10.03.2015).

⁷ Petr Pajas, Peter Fabian. Tree Editor TrEd 2.0; URL: <http://ufal.mff.cuni.cz/tred/> (10.03.2015)

- конфигурировать разметку в любом формализме (при условии, что синтаксические структуры в нем являются древовидными);
- конфигурировать визуализацию дерева: графически (цветом, формой и т. п.) маркировать нужные вершины дерева или ребра;
- осуществлять поиск по синтаксической структуре на специальном языке запросов;
- использовать построенные деревья для создания на платформе Pml-Tq⁸ синтаксического корпуса, доступного онлайн. Мы импортировали в TrEd тексты, уже размеченные морфологически⁹.

2. Языковые характеристики, определяющие принципы построения аннотации.

Селькупский и эвенкийский языки по своей морфологии агглютинативные, относятся к синтаксическому типу SOV [Comrie 1981: 80-97]. Это значит, что вершина фразовой проекции, или группы, должна следовать за комплементами (дополнениями) или модификаторами, см. (1) и (2а). При этом в селькупском и эвенкийском, в отличие, например, от японского, допускаются отклонения от данного правила: комплемент или модификатор может следовать за вершиной.

В эвенкийском и селькупском объект может следовать за глаголом, например, при фокусировании глагола (2б). Часто встречается топиализация объекта (2в)¹⁰.

⁸ Institute of Formal and Applied Linguistics. PML-TQ; URL: <https://ufal.mff.cuni.cz/tools/pml-tq> (10.03.2015)

⁹ Программные средства, используемые при подготовке корпуса; URL: <https://github.com/lalsnivts/syntax> (10.03.2015)

- (2) а. *bəju:mimni:* (S) *mo:ty:-wa* (O) *wa:-rə-n*(V)
 охотник лось-ВИН убить-ПРОШ-3.ЕД
 «Охотник (S) убил (V) лося (O)»
 б. *bəju:mimni:* (S) *WA:-RƏ-N* (V) *mo:ty:-wa* (O)
 «Охотник (S) УБИЛ (V) лося (O)»
 в. *MO:TY:-WA* (O) *bəju:mimni:* (S) *wa:-rə-n* (V)
 «ЛОСЯ (O) охотник (S) убил (V)» (а медведя нет)

Эти и другие факты плохо согласуются с единообразной синтаксической разметкой по параметру «порядок вершины и комплемента».

В кетском языке ситуация несколько иная. В нем актанты выражены прономинальными элементами, встроенными в полисинтетическую словоформу, или инкорпорированными именами, а полнозначные референциальные имена можно считать топиальными элементами¹¹.

3. Принципы построения синтаксической разметки

1. Единицей при сегментации текста считается предложение (S). Критерий для выделения предложения грамматический – наличие сказуемого (финитного, нефинитного, именного). Например, VP не является допустимой отдельной составляющей. Тут наша разметка отличается от таких корпусов устной речи, как VERMOBIL ([Kawata, Bartels: 5] – японский язык), в которых сегментация речи проводится с учетом характеристик

¹⁰ В (2б) глагол фокусируется *in situ*; фокус и (контрастивный) топик также часто стоят в начальной позиции [Nedjalkov 1997: 129-130].

¹¹ В этом случае разметка в рамках лексико-функциональной грамматики представляется адекватной, см. сноску 5.

акустического потока – пауз, междометий, хезитационных вставок¹².

2. Количество фразовых категорий, или групп, и соответствующих вершин минимально и определяется принадлежностью вершины к той или иной части речи с минимальным количеством дополнительных грамматических характеристик (например, финитность). Так, для любых глагольных сочетаний вводятся группы VP (глагольная группа), CVP (группа конверба), PartP (группа причастия) и соответствующие вершины, независимо от модели управления глагола; нет вершин, название которых уточняет лично-числовую форму глагола.

3. Базовые синтаксические отношения сводятся к минимуму: кроме отношений «вершина V – комплемент» и «предложение S – подлежащее», вводится не относящееся к базовой/аргументной структуре отношение адъюнкции/модификации (обстоятельство, именное определение). Это отношение передается через присоединение к S, VP, NP модифицирующих групп AdjP (группа прилагательного), AdvP (группа наречия) и т.д. Сюда же относятся и дислоцированные именные группы (см. объект NP_i в (3) ниже). Для избежания неоднозначности мы выделяем адъюнкты полужирным шрифтом.

4. В основе разметки лежит понятие о базовой структуре (1) предложения в языке типа SOV. Структура (1) входит в каждое предложение. Мы вводим нулевые узлы в позициях аргументов

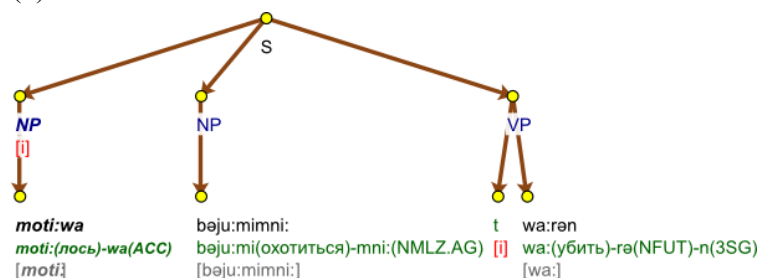
¹² Это так называемый *chunk/partial analysis* (неполный синтаксический анализ). Наш корпус, как и VERMOBIL, - корпус устной речи, однако мы ориентируемся на работы по созданию аннотации для корпусов устной речи (см. [Cresti et al. 2005]), которые уделяют достаточно большое внимание нормализации устной речи и для синтаксической разметки используют те же инструменты, что и при анализе письменных текстов.

базовой структуры, если они не заполнены словами – например, нулевые подлежащие и дополнения в (дее)причастном обороте.

Если порядок слов отклоняется от SOV, например, в связи с топикализацией (2в), вводится передвижение¹³.

Топикализованная NP_i присоединяется к узлу S слева, является адьюнктом и выделяется полужирным шрифтом. Внутри VP остается нулевой узел (дополнение), обозначаемое как след передвижения t_i – см. схему (3) для примера (2в).

(3)



Передвижение (топикализация или выдвижение на правую периферию предложения) вводится только для ядерных актантов: подлежащего и прямого дополнения. Модификаторы присоединяются в том месте, где они находятся.

5. Кроме основных групп (NP, VP и т.д.), вводятся ConnP (коннектор: ‘затем’, ‘тогда’), PrntP (парентетический оборот: ‘видимо’, ‘к счастью’) и др. Они, как и оговорки (SLIP) и междометия (INTJ), подчиняются узлу S или другому узлу (NP, VP,...), в зависимости от их линейной позиции в предложении.

¹³ В большинстве систем, использующих синтаксические разметки, передвижение не вводится, поскольку оно связано с большим объемом ручной работы. Топикализация в этих системах обозначается начальной позицией имени или другой части речи, а также посредством групп с вершиной – фокусной / топикальной частицей.

Эти единицы, по существу, не относятся к синтаксической структуре предложения, а являются элементами дискурса, или текста. Поэтому они выделяются курсивом.

Литература

1. *Казакевич О.А., Клячко Е.Л.* (2013), Создание мультимедийного аннотированного корпуса текстов как исследовательская процедура // Труды Международной конференции «Корпусная лингвистика – 2013». Санкт-Петербург. С. 292-300.
2. *Brants T., Skut W., Uszkoreit H.* (2003), Syntactic annotation of a German newspaper corpus // Treebanks: Building and using parsed corpora / ed. by Abeillé A. et al. . P. 72-87.
3. *Chomsky N.* (1957), Syntactic structures. The Hague.
4. *Comrie, B.* (1981), Language universals and linguistic typology. Oxford – Chicago.
5. *Cresti E., Panunzi A, Scarano A* (2005), The Italian corpus // C-ORAL-ROM. Integrated reference corpora for spoken Romance languages / ed. By E. Cresti, M. Moneglia, Cambridge – Oxford. P. 71-110
6. *Haegeman L.* (1994), Introduction to government and binding theory. 2-nd edition. Cambridge – Oxford.
7. *Kawata Y., Bartels J.* (2000), Stylesheet for the Japanese Treebank in VERMOBIL, Report 240. Seminar für Sprachwissenschaft, Universität Tübingen, available at: http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/stylebook_vm_jap.pdf

8. *Kernes H.* (2008), Syntactic prepossessing // *Corpus linguistics. An International Handbook*. Vol. 1 / ed. by Anke Lüdeling, Merja Kytö. Berlin, New York. P. 598-613.

9. *Nedjalkov I.* (1997), *Evenki*. London – New York.

References

1. *Kazakevich O.A., Klyachko E.L.* (2013), Developing a multimedia marked text corpus as a research procedure // *Proceedings of the International conference «Corpus Linguistics – 2013»*. Saint-Petersburg. P. 292-300.

2. *Brants T., Skut W., Uszkoreit H.* (2003), Syntactic annotation of a German newspaper corpus // *Treebanks: Building and using parsed corpora* / ed. by Abeillé A. et al. . P. 72-87.

3. *Chomsky N.* (1957), *Syntactic structures*. The Hague.

4. *Comrie, B.* (1981), *Language universals and linguistic typology*. Oxford – Chicago.

5. *Cresti E., Panunzi A, Scarano A* (2005), The Italian corpus // *C-ORAL-ROM. Integrated reference corpora for spoken Romance languages* / ed. By E. Cresti, M. Moneglia, Cambridge – Oxford. P. 71-110

6. *Haegeman L.* (1994), *Introduction to government and binding theory*. 2-nd edition. Cambridge – Oxford.

7. *Kawata Y., Bartels J.* (2000), Stylesheet for the Japanese Treebank in VERMOBIL, Report 240. Seminar für Sprachwissenschaft, Universität Tübingen, available at: http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/stylebook_vm_jap.pdf

8. *Kernes H.* (2008), Syntactic prepossessing // *Corpus linguistics. An International Handbook*. Vol. 1 / ed. by Anke Lüdelig, Merja Kytö. Berlin, New York. P. 598-613.

9. *Nedjalkov I.* (1997), Evenki. London – New York.

Казакевич Ольга Анатольевна

Московский государственный университет им. М.В. Ломоносова (Россия).

Kazakevich Olga

Lomonosov Moscow State University (Russia).

E-mail: kazakevich.olga@gmail.com

Галямина Юлия Евгеньевна

Московский государственный университет им. М.В. Ломоносова (Россия).

Galiamina Julia

Lomonosov Moscow State University (Russia).

E-mail: jugjug@yandex.ru

Клячко Елена Леонидовна

ЗАО «Синимекс» (Россия).

Klyachko Elena

CINIMEX Inc. (Russia).

E-mail: elenaklyachko@gmail.com

Рудницкая Елена Леонидовна

Институт востоковедения РАН (Россия).

Rudnitskaya Elena

Institute of oriental studies (Russia).

E-mail: erudnitskaya@gmail.com

Будянская Елена Михайловна

Московский государственный университет им. М.В.
Ломоносова (Россия).

Budianskaya Elena

Lomonosov Moscow State University (Russia).

E-mail: davemonty2@gmail.com