УДК 811Д61.1'42:004.738.5

М. В. Копотев, А. Мустайоки (Хельсинки)

Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет

Описывается аннотированный корпус русских текстов ХАНКО, который создается на Отделении славянских и балтийских языков и литератур Хельсинкского университета. Излагаются принципы подготовки корпуса; определяются типы релевантной лингвистической информации; обсуждаются связанные с подготовкой корпуса теоретические проблемы (аналитические формы и так называемые "эквиваленты" слов, множественность интерпретации). В приложении дается список лингвистических параметров корпуса.

ОСНОВНЫЕ ПРИНЦИПЫ СОЗДАНИЯ КОРПУСА

1. Направленность на широкий круг пользователей. При составлении корпуса ХАНКО, а также при разработке компьютерного интерфейса мы исходим из того, что он должен быть доступным не только узкому кругу специалистов, но и студентам и учителям русского языка. Это, разумеется, не значит, что мы полностью избегаем употребления лингвистических терминов, но выбор параметров поиска осуществляется так, что их знание минимизируется.

2. Направленность на максимальный охват грамматической информации, а не на объем материала. Параллельно с развитием компьютерной технологии объем корпусов постоянно увеличивается. Из этого, однако, не следует, что качественные характеристики корпусов улучшаются. Наоборот, в этом направлении относительно мало прогресса. Это и понятно, поскольку, несмотря на новые программы-анализаторы, работающие с текстом на естественном языке, более тщательный и точный лингвистический анализ попрежнему требует последующей ручной работы. В этой ситуации мы не хотим конкурировать с создателями других корпусов в количестве материала, наша цель — предоставить для широкого пользования аннотированный корпус, содержащий более точную грамматическую информацию по сравнению с тем, как она представлена в существующих или создаваемых корпусах. Из этого следует, что объем ХАНКО сравнительно небольшой.

3. Направленность на многоуровневую грамматическую информацию. Как известно, поиск на основе лексических единиц можно делать сравнительно легко, используя обыкновенные инструменты поиска. Корпус ХАНКО будет содержать многостороннюю **грамматическую** информацию, включающую морфологические, синтаксические и функциональные (семантические) характеристики. В процессе поиска их можно будет комбинировать. Так, в полном варианте корпус позволит вести такие варианты поиска: "несклоняемые существительные женского рода", "собирательные числительные в дательном падеже", "глаголы несовершенного вида в общефактическом значении", предложения типа инфинитив 4- существительное в именительном падеже", "косвенные дополнения с предлогом в", "наречия как обстоятельства места", "выражение обладания конструкциями, отличными от *'у* + родительный падеж'".

4. **Направленность на устоявшиеся лингвистические представления**. С потребностью в доступности корпуса связано то обстоятельство, что при его создании мы опираемся на устоявшиеся теоретические концепции, которые используются в известных лингвистических трудах и/или учебной литературе по русской грамматике. Исключение составляет функциональная часть и лишь только потому, что в этой области еще нет широко признанной терминологии, общепринятых классификаций и т. д. (см. подробнее ниже).

5. **Возможность более чем одной интерпретации языковых фактов.** Наконец, еще одно принципиальное решение, принятое создателями ХАНКО, требует особых комментариев. Любому исследователю, работающему с конкретным языковым материалом, приходилось сталкиваться с тем, что тот или иной языковой факт с трудом подается однозначной характеристике. Такая ситуация обусловлена двумя обстоятельствами. Во-первых, фрагмент языкового материала может допускать более чем одну содержательную интерпретацию, например, в предложении *Двери не открывать ни в коем случае* невозможно однозначно определить, является ли словоформа *двери* формой родительного падежа единственного числа или формой винительного падежа множественного числа [1, с. 15-16]. Во-вторых, существуют случаи, когда содержательная интерпретация очевидна, но критерии лингвистической классификации столь шаткие, что языковая единица может быть отнесена к разным классам слов. Так, слово *тысяча* имеет как черты существительных, так и черты числительных и может быть, таким образом, определено как представитель обеих категорий. В таких случаях в корпусе будут отмечены оба варианта. Такая "нечеткость" , как нам кажется, облегчает поиск нужной для пользователя информации.

ТИПЫ ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ

В ХАНКО предполагается представить следующую лингвистическую информацию (полный список параметров см. в Приложении).

1. Морфологическая информация

Полная морфологическая характеристика каждой текстоформы с возможностью указать спорные случаи, имеющие неоднозначную трактовку. Работа осуществляется автоматически с последующей ручной обработкой. Автоматический анализатор создан в рамках теории "двухуровневой грамматики" (Two-level Grammar) [2] и "грамматики ограничений" (Constraint Grammar) [3] как составная часть системы TWOL (основные сведения об анализаторе можно найти в статье [4]). Автоматическая дизамбигуация выполнена с помощью отдельного "постморфологического" модуля, при этом использовались правила, дающие только стопроцентную точность при устранении грамматических омонимов. Одно из таких правил связано с предложным управлением. Например, в предложении *Без друга пропадешь* программа автоматического анализа выдает для формы существительного два падежных варианта Род. и Вин. Пад. Однако падежное управление предлога *без* позволяет убрать омонимию, устранив Вин. Пад. из списка интерпретаций. Именно это и делает модуль "постморфологии" .

Основной акцент при создании корпуса ставится на предоставление синтаксической информации, поскольку морфологическое аннотирование является на сегодняшний день стандартным минимумом аннотирования и в достаточно полном виде представлено и в других, более представительных корпусах. Словообразовательная информация в корпусе отсутствует полностью, а из словоизменительных явлений во внимание принимаются только те, которые проявляются на синтаксическом уровне. По этой причине не отмечаются типы склонения, спряжения, ударения и т. п., но указываются, кроме традиционных морфологических категорий (род, падеж, число, лицо, одушевленность и т. д.), в частности, формы второго родительного и второго предложного падежей.

2. Синтаксическая информация

Синтаксическая информация маркирует три типа единиц: словосочетание, клаузу, предложение. На уровне словосочетания указываются тип связи (управление, примыкание, согласование, координация, сочинение), тип опорного слова (глагольный, субстантивный, адъективный, адвербиальный) и тип зависимого слова (глагольный, субстантивный, адъективный, адвербиальный, нумеральный, местоименный). На уровне клаузы отмечаются, в частности, структурные схемы простого предложения (согласно Академической грамматике 80-го г. издания), а также второстепенные члены предложения. На последнем, синтаксическом, уровне указывается тип предложения (сложное, простое, бессоюзное) .

Синтаксическая обработка осуществляется частично автоматически, на основе морфологической информации. Однако, естественно, необходима будет и существенная ручная обработка.

3. Функционально-семантическая информация

Создание ХАНКО мыслится как составная часть проекта "Функциональный синтаксис русского языка" (ср. [5-8]). Из этого следует, что наш "проект в проекте" нацелен, кроме прочего, на решение задач функциональной грамматики и должен предлагать максимально широкие возможности для описания языка "от значения к форме". Как известно, это весьма сложная задача, поскольку семантические категории репрезентируются на поверхностном уровне самыми разными языковыми средствами[1]. Однако, используя морфологическую и синтаксическую информацию как полигон, создание функциональной части корпуса можно частично автоматизировать. Естественно, понадобится и трудоемкая ручная работа.

Принципы отбора текстов

В силу сложности предполагаемого аннотирования и достаточно большого количества неавтоматизируемой работы, было принято решение ограничиться небольшим объемом текстов (около 100 тысяч текстоформ). Для этого был выбран один журнал, достаточно полно представляющий современную публицистику. Конечно, выбор журнала в таком случае всегда субъективен, однако издание должно отвечать определенным критериям:

- в журнале должен быть представлен широкий спектр жанров публицистики: аналитические материалы, интервью, рецензии и т. д.;
- тематика текстов должна быть достаточно многообразна: финансы, экономика, культура, политика, общественная жизнь, государственные институты и т. д.;
- выбранные тексты должны быть написаны людьми, владеющими стилистическими ресурсами русского языка.

На основе этих критериев для создания корпуса были выбраны все крупные статьи из журнала "Итоги" за январь 2001 года. Этот журнал представляет содержательно и стилистически широкую палитру языковых средств, в нем представлены работы журналистов, уровень владения языком которых достаточно высок.

Некоторые теоретические проблемы

Обычно создатели языковых корпусов идут на более или менее серьезные компромиссы, выбирая между скоростью обработки материала и точностью интерпретации. Так, большинство создателей русскоязычных корпусов игнорируют аналитические конструкции, сложные для обработки и плохо поддающиеся автоматическому анализу. Например, аналитические формы сослагательного наклонения глагола в предложении *Студент прочитал бы книгу, если бы она была в библиотеке* определяются как формы прошедшего времени изъявительного наклонения и — отдельно — частицы. Однако при таком подходе достаточно большая

[1] Возможность создания корпуса, ориентированного на функциональный подход к описанию языка, и сложности, возникающие при этом, обсуждались неоднократно (см. [9; 10]).

и важная часть грамматической информации искажается или не учитывается.

Обычно остаются вне поля зрения следующие аналитические конструкции.

о Формы сослагательного наклонения глагола: *прочитал бы, сходил бы* и более сложные случаи, как в предложении *Я хочу, чтобы студенты прочитали эту книгу*, в котором слившиеся частица и союз не отменяют аналитизма сослагательной формы *прочитали [бы]*. о Формы сложного будущего времени: *буду читать, буду ходить*.

о Аналитические формы прилагательных и наречий: *более быстрый, более быстро*. о Составные и дробные числительные: *сто сорок восемь, две третьих*.

о Аналитические формы местоимений: *ни от кого*. Естественно, что при создании корпуса ХАНКО предполагается уделить достаточно много внимания ручной обработке текстов. С одной стороны, это работа, требующая значительных затрат времени, но, с другой стороны, это увеличивает информационную насыщенность материала. Подобная задача выполнима для корпуса такого относительно небольшого объема, как наш.

Еще одна задача, которая, насколько нам известно, никогда не ставилась при аннотировании текстов русского языка, заключается в том, что мы планируем уделить специальное внимание так называемым "эквивалентам слова". Речь идет о единицах типа *потому что, в течение, к сожалению*. Эта группа языковых единиц, занимающая промежуточное положение между словом и словосочетанием, не пользуется особым вниманием лингвистов, тем более трудно говорить об учете этих единиц при разметке текстов в ходе создания корпуса. Мы считаем, что эквиваленты слова необходимо ввести в корпус как особый ярус рассмотрения языкового материала. Так, например, предложение *Во время сессии не стоит забывать об отдыхе* состоит из восьми слов, однако, так же очевидно, что комплекс *во время* функционально является предлогом, управляющим родительным падежом в предложно-именной группе *во время сессии*. Сложность выделения и неопределенность статуса подобных единиц заставляет прийти к компромиссному решению: в создаваемом корпусе эти единицы размечаются и как самостоятельные (предлог во, существительное *время* в данном случае), и как единый комплекс — эквивалент слова (предлог *во время)*. Списки таких единиц и их частеречная квалификация определены по ([11; 12]; ср. также [13]).

Основные данные об участниках проекта

ХАНКО создается на Отделении славянских и балтийских языков и литератур Хельсинкского университета при поддержке гранта Академии Финляндии. Руководителем проекта является Арто Мустайоки, а главным "архитектором корпуса" — Михаил Копотев. Кроме того, существует консультативная группа, в состав которой в настоящий момент входят Л. А. Бирюлин и А. Никунласси из Хельсинкского университета, а также И. Кюльмоя из Тартуского университета. В ручной обработке корпуса участвуют русскоязычные филологи, живущие в Финляндии, в том числе студенты и аспиранты Отделения. Дополнительную информацию о проекте можно найти в Интернете по адресу: www.slav.helsinki.fi/hanko. Комментарии и вопросы, касающиеся корпуса, можно посылать по адресу mihail.kopotev@helsinki.fi

ХАНКО создается поэтапно, первые результаты будут представлены в Интернете сразу после того, как будет подготовлена морфологическая часть. Мы надеемся, что эта стадия проекта будет завершена к концу 2003 года.

СПИСОК ЛИТЕРАТУРЫ

1. Mustajoki A. &; Heino H. Case Selection for the Direct Object in Russian Negative Clauses. Part 2: Report on a Statistical Analysis.— Helsinki: University of Helsinki, 1991.
2. Koskenniemi K. Two-level Morphology: a General Computational Model for Word-form Recognition and Production.— Helsinki: University of Helsinki, 1983.
3. Constraint Grammar: A Language-Independent Framework for Parsing Unrestricted Text // Ed. by F. Karlsson, A. Voutilainen, J. Heikkilä & A. Anttila, Mouton de Gruyter: Berlin/New York, 1995.
4. Viikki L. RUSTWOL: A System for Automatic Recognition of Russian words. 1997. www.lingsoft.fi/doc/rustwol/rustwol.txt.
5. Mustajoki A. Mielestä kieleen: kontrastiivisen funktionaalisen lauseopin teoriaa.— Helsinki: Yliopistopaino, 1993.
6. Мустайоки А. Возможна ли грамматика на семантической основе? // Вопросы языкознания.— 1997.— № 3.— С. 15-25.
7. Мустайоки А. Аспектуальность в тео рии функционального синтаксиса. Die grammatischen Korrelationen, / hrsg. von B. Tosovic, Graz: Inst. fur Slawistik, 1999.— P. 229-244. [=Grazer Linguistische Slawistentage, Bd. 1].
8. Мустайоки А. Теория функционального синтаксиса: от семантических структур к языковым сред ствам [в печати].
9. Dik S. C. Functional Grammar and its Potential Computer Applications. Corpus Linguistics and Beyond. Processing of the Seventh International Conference on English Language Research on Computerized Corpora.— Amsterdam, 1987.— P. 253-268.
10. S outer C. Systemic-functional Grammar and Corpora, Theory and Practice in Corpus Linguistics // Ed. by J. Aarts and Meis.— Amsterdam: Rodopi, 1990.— P. 179-211.
11. Рогожникова Р. П. Словарь эквивалентов слова.— М.: Русский язык, 1991.
12. Богданов С. И., Рыжова Ю. В. Русская служебная лексика. Сводные таблицы.— СПб: Изд-во СПб ун-та, 1997.
13. Мустайоки А. &; Копотев М. Эквивалент слова — необходимое понятие в описании языка? [в печати].

ПРИЛОЖЕНИЕ: ПАРАМЕТРЫ БА ЗЫ ДАННЫХ

Для каждого текста отмечаются название статьи, автор, номер журнала. 1. ОСНОВНЫЕ ЕДИНИЦЫ

1.1. Текстоформа (словоформа)
1.2. Лемма (начальная форма)
1.3. Словосочетание (синтаксема)

1.4. Клауза
1.5. Предложение
1.6. Абзац

2. МОРФОЛОГИЧЕСКИЕ ПРИЗНАКИ
2.1. Существительное
2.1.1. Род: мужской, средний, женский, общий
2.1.2. Число: единственное, множественное, pluralia tantum
2.1.3. Падеж: именительный, родительный-1, роди тельный-2, дательный, винительный, твори тельный, предложный-1, предложный-2
2.1.4. Несклоняемое (не имеющее форм рода, числа и падежа; но значение по контексту для таких слов будет определено)
2.1.5. Лексико-грамматические разряды: собствен ное или нарицательное; одушевленное или не одушевленное
2.2. Прилагательное
2.2.1. Лексико-грамматические разряды (качествен но-относительное, притяжательное)
2.2.2. Род: мужской, средний, женский
2.2.3. Число: единственное, множественное
2.2.4. Падеж: именительный, родительный, датель ный, винительный, творительный, предложный
2.2.5. Несклоняемое (не имеющее форм рода, числа и падежа; но значение по контексту для таких слов будет определено)
2.2.6. Краткая / полная форма
2.2.7. Степень сравнения: сравнительная синтети ческая, превосходная синтетическая, сравни тельная аналитическая, превосходная анали тическая
2.3. Местоимение
2.3.1. Разряд: личное, возвратное, притяжатель ное, вопросительное, относительное, указа тельное, определительное, отрицательное, не определенное
2.3.2. Лицо: первое, второе, третье
2.3.3. Род: мужской, средний, женский
2.3.4. Число: единственное, множественное
2.3.5. Падеж: именительный, родительный, датель ный, винительный, творительный, предложный
2.4. Числительное
2.4.1. Разряд: порядковое, количественное, собира тельное, дробное
2.4.2. Состав: простое, составное
2.4.3. Род: мужской, средний, женский
2.4.4. Число: единственное, множественное
2.4.5. Падеж: именительный, родительный, датель ный, винительный, творительный, предложный
2.5. Глагол
2.5.1. Личная форма, инфинитив, причастие, дее причастие, безличная форма
2.5.2. Вид: совершенный, несовершенный, двувидо вой
2.5.3. Время: настоящее, прошедшее, будущее про стое, будущее сложное
2.5.4. Залог: действительный, страдательный
2.5.5. Возвратность / невозвратность
2.5.6. Наклонение: изъявительное, повелительное, сослагательное
2.5.7. Род: мужской, женский, средний
2.5.8. Лицо: первое, второе, третье, безличная форма
2.5.9. Число: единственное, множественное
2.5.10. + для причастий краткая / полная форма
2.5.11. Падеж: именительный, родительный, датель ный, винительный, творительный, предложный
2.6. Наречие
2.6.1. Степень сравнения: сравнительная синтети ческая, превосходная синтетическая, сравни тельная аналитическая, превосходная анали тическая
2.7. Категория состояния
2.8. Предлог
2.8.1. Падеж, которым управляет предлог
2.9. Частица
2.10. Союз
2.11. Междометие
2.12. Аббревиатура

3. СИНТАКСИЧЕСКИЕ ПРИЗНАКИ
3.1. Синтаксис словосочетания / синтаксемы
3.1.1. Тип связи: управление, примыкание, согласо вание, координация, сочинение
3.1.2. Тип опорного слова: глагольное, субстантив ное, адъективное, адвербиальное, нумераль ное, местоименное
3.1.3. Тип зависимого слова: глагольное, субстан тивное, адъективное, адвербиальное, нуме ральное, местоименное
3.2. Синтаксис клаузы
3.2.1. Тип клаузы: самостоятельная клауза, подчи ненная клауза, главная клауза
3.2.2. Структурные схемы (примерный список)
3.2.2.1. Nnom—Vf *(Лес шумит; Дети веселятся);*
3.2.2.2. Vf3sInf (*Следует подождать; Запрещает ся шуметь);*
3.2.2.3. Ngen (neg) Vfas *(Воды убывает; Друзей не находится);*
3.2.2.4. Nnom — Nnom *(Брат — учитель; Моск ва — столица);*
3.2.2.5. Nnom — Adjкратк.ф. *(Ребенок умен);*
3.2.2.6. Nnom — Adjnom, полн.ф *(Ребенок умный);*
3.2.2.7. Nnom — Partкратк.ф. *(Дом построен);*
3.2.2.8. Nnom — N.../Adv *(Дом — у дороги; Конец близко);*
3.2.2.9. Nnom — Inf *(Задача — учиться);*
3.2.2.10. Nnom — Adv-o *(Экскурсия* — [это] *инте ресно);*
3.2.2.11. Inf — Nnom *(Трудиться — доблесть);*
3.2.2.12. Inf— Adv-o *(Кататься — весело);*
3.2.2.13. Inf — Inf *(Руководить значит прове рять);*
3.2.2.14. Praed Inf *(Пора ехать; Нельзя оставать ся)*',
3.2.2.15. Praed (neg) N_{gen}/N_{a}cc *(Видно следы; Не видно следы/следов);*
3.2.2.16. PraedpartNgen *(Наготовлено запасов; Подтверждения не получено);*
3.2.2.17. Adv_{qU}ant(Niquant) — N_{gen} *(Много дел; Ма ло времени; Довольно неприятностей);*
3.2.2.18. *Нет* N_{gen} *(Нет сомнений);*
3.2.2.19. *Ни* Ngen *(Ни облачка; Ни звука);*
3.2.2.20. *Никого (ничего)* N_{gen} *(Никого знакомых);*
3.2.2.21. *Никакого (ни одного, ни единого, ни малейшего)* Ngen *(Никакой надежды);*
3.2.2.22. $Pron_{neg}$ Inf *(Некому работать);*
3.2.2.23. Vf_{3sg} *(Светает; Трясет);*
3.2.2.24. Vf_{3pl} *(Стучат)]*
3.2.2.25. Nnom *(Ночь)]*
3.2.2.26. Ng_{en} *(Народу!);*
3.2.2.27. N_{g}en/Ngen2/Nacc (*Чаю! Хлеба и зрелищ!);* .
3.2.2.28. Praed ([Ему] *рады)]*
3.2.2.29. (Neg) Ing *(Не пройти);*
3.2.2.30. Praed *(Холодно)]*
3.2.2.31. Praedpart *(Натоптано);*
3.2.3. Второстепенные члены предложения и слова, не входящие в структуру предложения
3.2.3.1. Дополнение: прямое, косвенное
3.2.3.2. Обстоятельство: времени, места, причины, цели, условия, уступки, образа действия, меры и степени
3.2.3.3. Определение: согласованное, несогласован ное, приложение
3.2.3.4. Обращение
3.2.3.5. Вводное слово
3.2.4. Конструкции, осложняющие структуру пред ложения
3.2.4.1. Однородные ряды: открытые, закрытые

3.2.4.2. Обособленные конструкции: деепричастный оборот, причастный оборот, субстантивный оборот, адъективный оборот, поясняющие обороты, уточняющие обороты 3.3. Синтаксис сложного предложения
3.3.1. Сложносочиненное предложение
3.3.2. Сложноподчиненное предложение
3.3.3. Бессоюзное предложение

4. ФУНКЦИОНАЛЬНЫЕ ПРИЗНАКИ (предварительный список) 4.1. Ядерные положения вещей
4.1.1. Физические действия
4.1.2. Физиологические положения вещей
4.1.3. Эмоциональные положения вещей
4.1.4. Существование и его изменение
4.1.5. Обладание и его изменение
4.1.6. Характеризация
4.1.7. Идентификация и т. д.

4.2. Расширение ядра (компликаторы)
4.2.1. Каузация
4.2.2. Темпоральная фаза
4.2.3. Модальная фаза и т. д.

4.3. Комментарии говорящего (комментаторы)
4.3.1. Речевые функции
4.3.2. Авторизация и т. д.

4.4. Спецификаторы
4.4.1. Время
4.4.2. Аспектуальность
4.4.3. Модальность
4.4.4. Определенность
4.4.5. Количество
4.4.6. Место и т. д.

4.5. Связь между положениями вещей
4.5.1. Нейтральные отношения
4.5.2. Таксисные отношения
4.5.3. Логические отношения

ISSN 0548-0027

НАУЧНО-ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

СЕРИЯ 2

Информационные процессы и системы

СОДЕРЖАНИЕ

№ 6

2003