Словацкий национальный корпус

Radovan Garabík

Jazykovedný ústav Ľ. Štúra SAV

813 64 Bratislava, Slovakia

korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

Príspevok bol prednesený na medzinárodnej konferencii Корпусная лингвистика – 2004 (Sankt Peterburg 2004) a bude publikovaný v zborníku.

Abstract. First preliminary work on text corpus of Slovak language started in 1992, with primary intention to provide enough searchable material for lexicographic purposes. In 2002, the Government of the Slovak Republic approved the project to build a representative corpus of contemporary Slovak language to be used in further linguistic research. Currently, Slovak National Corpus contains 192 million lemmatised and morphologically annotated words, with the plans of reaching 200 million words of balanced corpus in 2005. A new morphosyntactic tagset for the Slovak language has been developed and is used to create a manually lemmatised and annotated core, useful for further analysis and NLP tools training. Incoming texts are converted into common text format based on XCES (TEI) standard, with additional linguistic annotation added at this stage, and then converted into internal format for the corpus manager. Access to the corpus can be given to any person in exchange for a simple registration procedure, provided that the corpus is used for educational or personal non-commercial purposes. Texts in the corpus include complete bibliographic and style-genre annotation, described here in detail.

История

Корпусная (и вообще математическая) лингвистика не пользовалась большим вниманием в словацкой лингвистике, хотя первые шаги были сделаны уже в 1962 – 1970 годах после возникновения отдела Математической лингвистики Института словацкого языка Словацкой академии наук. Отдел занимался преимущественно до тех пор не исследованными аспектами словацкого языка, в том числе главным образом статистическим анализом.

Возобновление интереса к компьютером как к средству обработки лингвистических данных произошло в связи с проникновением компьютеров в научное общество после 1989-ого года[1], но главная польза компьютеров состояла в поддержке лексикографических работ. Лишь в 1991-ом году появились первые мысли[2] о необходимости коллекции достаточного количества текстов в цифровом виде с возможностью сложных поисков. Эта коллекция, без всякой следующей лингвистической обработки, стала называться Корпус текстов словацкого языка[3], и ею пользовались только в институте, главным образом как дополнительный источник данных к подготовке Короткого словаря словацкого языка.

Текстовой корпус словацкого языка продолжал увеличиваться в количестве словоупотреблений, но скоро стало ясно что для подготовки большого, современным стандартом соответствующего корпуса есть необходимо принять более систематический подход. В начале 2002-ого года правительство Словацкой республики одобрило проект создания представительного корпуса современного словацкого языка для будущего лингвистического исследования, вместе с созданием отдела Словацкого национального корпуса в рамках Института языкознания Словацкой академии наук.

Принципы постройки корпуса

Словацкий национальный корпус, чтобы предоставить желающую информацию своим пользователям, должен быть не только механический сбор текстов (хотя и это может своим образом использоваться в лингвистических исследованиях, как показывает бывшая версия текстового корпуса словацкого языка), но он должен отвечать нескольким определённым требованиям. [4]

Во первых, Словацкий национальный корпус должен быть корпусом *современного* словацкого письменного языка. Под современным словацким языком подразумевается язык после последней значительной реформы орфографии в 1953 году. Постепенные малейшие поправки не вносили в язык большие изменения, и самая последняя реформа в 1994-ом году, хотя и не так незначительная, не проникла достаточно в общество. В корпус решено включать тексты возникшие в 1955 – 2005 годах. Нижний предел назначен промежутком времени который истёк пока ещё новая орфография не стала общепринятой. Тексты из более давних пор в принципе не исключены из обработки, но они не входят в окончательный корпус – постройка диахронического корпуса Словацкого языка будет отдельным проектом. Принято решение включать в корпус прежде всего современные тексты, поскольку их можно нетрудно приобрести. Но будет сделано тоже сечение по текстах прошлого времени.

Во вторых, корпус должен быть *представительным*, т.е. он должен включать самые разные формы языка и стилей, пропорционально к их распространённости в письменном общении.

Дальше, корпус должен содержать как минимум информацию о лемме лексемы и соответствующих грамматических категориях.

Тексты в корпусе также должны сохранять информацию о их происхождении и включать остальные метатекстовые данные о стиле, жанре, авторе, полу автора, источнике текста итд. Эта информация очень важна для эвентуальной постройки разных субкорпусов.

И наконец, корпус должен быть общедоступный для всех интересующихся, как профессиональных лингвистов, так и для любителей языка или учащихся словацкий язык..

Желанная версия представительного Словацкого национального корпуса должна содержать приблизительно 200 миллионов словоупотреблений с полной морфологической разметкой и с добавленной библиографической информацией.

Сбор текстов

Учитывая проблемы которые встретили остальные корпусы, тексты в Словацкий национальный корпус входят только когда удастся приобрести разрешение включить тексты в состав корпуса. Хотя в словацком авторском законе есть возможность использовать тексты для научных целей, следующее распространение текстов не позволено ни в каком случае, и такой корпус не мог бы быть общедоступен.

Главными источниками текстов являются издательства, газеты и журналы, но довольно часто тексты (или разрешение для включения текстов в корпус)

приобретены прямо от авторов.

Обработка текстов

Тексты в корпус входят в четырёхуровневую иерархию. На первом уровне – архив, в котором данные сохраняются в оригинальном виде, т.е. так как они были получены.

Следующий уровень – банк, где тексты переведены в общий формат вроде простого XML, содержавшего нужную информацию о типографии в оригинальном формате. От действительного XML этот формат отличается возможностью перекоса тэгов, который иногда очень трудно удалить. На этом уровне тексты аннотированы библиографической разметкой. Один текст в банке соответствует одной логической единицы документа, например один роман, одна статья газетного текста или один закон из сбора. Из этого следует что одному документу в архиве часто соответствует большое количество единиц в банке.

Следует корпусоид, в котором тексты находятся в формате XCES[5], содержавшие уже тексты с полной морфосинтаксической разметкой. Иначе структура корпусоида совпадает структуре банка.

Наконец тексты поступают в секцию «дата», в цифровом формате конкретного корпусного менеджера. Здесь тексты уже распределены по разным субкорпусам.

Метатекстовая разметка

Каждая единица текста в банке аннотирована с учётом библиографического источника, стиля и жанра текста. Для каждого текста, аннотация находится в самостоятельном текстовом файле, где в каждой строке записана комбинация ключ-значение. Значение может быть либо свободное, либо избранное из множества возможных значений. Аннотация в архиве ведётся подобным образом, но немного проще.

У каждого ключа возможны эти специальные значения:

- пустое место: неопределённое значение – о значении не можно сказать ничего точного
- XXX: неизвестное значение – его не удалось открыть, напр. имя автора в газетной статьи
- YYY: неопределимое значение – в принципе невозможно определит его, напр. пол переводчика, если произведение не является переводом
- MIX: смешение других значений
- MSC: другое значение
- TTT: неизвестное значение которое надо определить позже

Полное описание аннотации архива и банка в приложении.

Лингвистическая разметка

Основным исходным пунктом лингвистической обработки корпуса является токенизация. В корпусе принято решение разделять текст на самые малейшие

логические единицы которые можно определить в письменных документах. Это прежде всего значит, что в корпусе ни каким образом не выделены составные названия – например *Sankt-Peterburg* будут три токена, *Peterburg* с полной морфосинтаксической пометкой, *Sankt* без морфологической информации и „-“ как пунктуационный знак. Главная причина таково решения была невозможность точного определения границы когда составное название уже не можно считать одной лексической единицей текста.

Но с другой стороны, соединённые слова не разделяются, так например местоимения с агглютинативными предлогами (*preň, oň, zaňho...*) оставлены как особый вид местоимений. В словацком языке, в отличие от чешского, они встречаются достаточно редко чтобы такой подход был обоснованным.

При подготовке формального описания грамматических категорий словацкого языка подходящего для автоматической обработки мы опирались преимущественно на два уже существующие проекта автоматической морфологической разметки словацкого языка. Первый из них – автоматический анализатор построен на Институте формальной и прикладной лингвистики Карлова университета в Праге[6, 7]. Структура помет использована в этом анализаторе (и именно и анализатор) идентична с пометами использованными в Чешском национальном корпусе, с некоторыми мелкими отличиями совпадающими с отсутствием редких грамматических явлений чешского языка в словацком. Пометы характеризованы позиционной системой, каждой позиции ж помете соответствует одна грамматическая категория. Категории не имеющее смысл для указанной части речи обозначены знаком „-“. Анализатор включает модуль дисамбигуации и он хорошо проверен входя в многие проекты нуждающееся в морфологическом анализе, в том числе и коммерческие. Словарь корневых морфем и парадигм словацкого языка достаточно разработан, хотя и не достигает размер чешского словаря.

Второй анализатор Ajka построен на Факультете информатики в Брно[8]. Он независимый на языке и в принципе подходит для всех языков которые используют главным образом аффиксы для реализации грамматических категорий. Вопрос только в существовании словарей. Структура помет состоит из пар символов, где первый символ пары определяет грамматическую категорию и второй – значение категории. Если категория не имеет смысл для указанной части речи, соответствующий символ просто не включён в помету. Ajka не включает моду;ь дисамбигуации, и в настоящее время проводятся работы над созданием достаточного словаря словацких морфем в сотрудничестве отдела Словацкого национального корпуса и Факультета информатики.

Есть ещё третья система помет, которая использована в проекте Multext-East[9]. Это система была разработана для общей разметки славянских языков. Словацкая версия не существует, но система достаточно универсальная чтобы было можно ей воспользоваться.

При проектировке собственной системы было необходимо учесть несколько противоречивых требований, преимущественно основанных на непригодности формальных теоретических описаний морфологии словацкого языка к машинной обработке. Главный принятый принцип – стремление к описанию чисто морфологических категорий, без учёта синтаксических или даже семантических ролей словоупотреблений. К сожалению, не всегда было возможно соблюдать этот

принцип.

Наша система помет имеет черты как позиционных, так и ключевых систем.. Пометка начинается знаком определяющим часть речи, следует знак соответствующий парадигму (если это флективная часть речи), и следуют знаки грамматических категорий, причём значение каждой грамматической категории соответствует одно-однозначно одному знаку, знаки не повторяются и следуют в определённом порядке, и в пометке находятся только знаки релевантных категорий.

Строго морфологический принцип ведёт к тому что прошедшее время глаголов не выделено как особая категория, вместо того выделяется L-причастие. Будущее время назначено только в совершенных глаголах, и будущее время глаголов несовершенного вида назначено только в помете вспомогательного глагола „byt“

Вопрос звательного падежа – в словацком языке звательный падеж не осуществлён морфологией, он почти всегда совпадает с именительным. Но, есть несколько десяток слов, у которых сохранилась древняя морфология в звательном падеже, и некоторые из этих слов встречаются очень часто, чтобы ими было можно пренебречь. Сверх того, в разговорном языке появляется „новый звательный падеж“, возникший из сокращённых или уменьшительных вариантов прежде всего собственных имён или названий семейных отношений – эволюцией звательного падежа словацкий язык близко напоминает русский. Из-за этого, было решено включит в систему помет тоже звательный падеж, хотя это вносит довольно большую омонимию с именительным.

Причастия, хотя своей морфологией не отличаются от прилагательных имён, введены как собственная часть речи.

Омонимы, которых морфология та же самая, не отличаются (у них та же самая лемма).

Описание системы помет в Таб.1.

Для автоматической разметки в корпусе используется сейчас словацкая версия анализатора сделанного в Праге. Кроме того есть „ядро“ корпуса, состоящее из текстов с ручной морфосинтаксической разметкой, полезное для будущего исследования и тренировки программных средств языковой обработки.

Поиск в корпусе

В главную очередь, корпус разделяется на три субкорпуса, в зависимости от лицензии текстов. Первый, доступный для широкого общества, состоит из текстов для которых есть разрешение использования в корпусе без добавленных ограничений. Однако, некоторые текстодатели разрешили включить тексты в корпус при условии что ими могут пользоваться только сотрудники института языкознания, и эти тексты вместе с первой группой создают второй субпорпус.

Третий субкорпус состоит из совершенно всех текстов, даже из тех для которых не существует разрешение предоставления другим лицам (потому что переговоры о приобретении разрешения ещё не завершились, или есть другие препятствия). Этот субкорпус доступен только членам отдела словацкого национального корпуса.

Доступ к первому, публичному корпусу может приобрести любое частное лицо, подписавшее договор о пользовании корпусом. Право пользоваться корпусом

личное, не переносное и действует в течении одного года, после чего его надо возобновить. Доступ организациям или другим группам не разрешается, с исключением студентов на уроках связанных с корпусной лингвистикой, которым предоставлен общий доступ, действителен обыкновенно в течении одного семестра. Доступ позволен исключительно для научно-исследовательских или частных целей, и ни в каком случае не может быт предоставлен для коммерческой пользы.

Для поиска в корпусе пользуется система Manatee/Bonito[10], которая состоит из сервера (Manatee) и клиента (Bonito). Пользователи корпуса подключаются к серверу и задают вопросы, которые проводятся на сервере. Сервер позволяет простой поиск одного слова, или фразы (нескольких слов в очередном порядке), или произвольных регулярных выражений. К каждому словоупотреблению может быть добавлена информация в виде атрибутов, по которым тоже можно проводить поиск. В Словацком корпусе добавлены атрибуты *lemma* и *tag*, соответствующие лемме и морфосинтаксической помете.

Manatee дальше позволяет проводить разные статистические исследования в рамках колокаций, как MI-score, T-score, и также частотный анализ разных форм отысканных единиц.

К каждому тексту может быть получена полная информация о метатекстовой (библиографической) разметке, которая находится в банке, и в контексте можно включить изображение структурных помет текста, показывавших или разные типографские черты текста (как например заголовки, курсив), или внутреннюю структуру текста (конец абзаца, конец предложения). Поиск можно ограничить только в рамках желающих структур. Manatee позволяет работать с несколькими корпусами, и каждый пользователь имеет доступ к версии корпуса содержавшей все тексты подходящей лицензии, и к версии с исключёнными текстами, которые по разным причинам мешали бы в лингвистическом анализе – например тексты из специализированных лингвистических журналов или тексты с отсутствующими диакритическими знаками.

Заключение

В настоящее время Словацкий национальный корпус состоит из 192 миллионов словоупотреблений. Публично доступный субкорпус содержит 183 миллионов словоупотреблений. Корпусы включают метатекстовую библиографическую и стилистическо-жанровую разметку, и лингвистическую разметку с показанием леммы и морфосинтаксических категорий. В корпус постоянно добавляются новые тексты и дополнительная лингвистическая обработка.

Отдел Словацкого национального корпуса принимает участие в разных проектах, например в подстройке словацкой версии WordNet (вместе с Университетом в Прешове), синтаксически аннотированного корпуса словацкого языка (вместе с Университетом в Прешове и Институтом формальной и прикладной лингвистики Карлова университета в Праге), параллельных Словацко-Хорватского (вместе с Хорватским национальным корпусом) и Словацко-Русского (вместе с Кафедрой математической лингвистики Филологического факультета Санкт-Петербургского государственного университета) корпусов.

Приложение

Описание аннотации в архиве (перечень возможных ключей):

- Name: Название документа
- Acqdate: Число когда текст попал в корпус
- Id: Однозначный идентификационный номер текста
- Source: Источник текста (напр. название издательства)
- License: Тип лицензии текста
- Comment: Произвольный комментарий

Описание аннотации в банке:

- Name: Название текста
- Origname: Оригинальное название текста (если это перевод)
- Author: Имя автора (авторов), как напечатано в произведении
- Origauthor: Оригинальное имя автора (на оригинальном языке, если оно отличается от предыдущего значения)
- Translator: Имя переводчика
- Translation:

trn	перевод
org	оригинальный текст
ftr	свободный перевод

- ISBN: ISBN номер
- ISSN: ISSN номер
- SourceId: Идентификационный номер документа архива, который является источником этого текста
- Rhyme: рифм текста

nrh	не рифмический
rhy	рифмический
MIX	частично рифмический

- Type: тип текста, подробнее определён ключом Subtype:

	Type		**Subtype**
img	художественный		
		poe	поэзия
		pro	проза
		dra	драма
inf	информативный текст		
		pub	публицистика
		adv	реклама, объявления
		adm	администрация
prf	профессиональный текст		
		sci	научные тексты, статьи, профессиональные журналы, учебники для ВУЗ-ов
		pop	популярно-научные тексты
		txb	учебники для основных и средних школ

	Type		Subtype
		enc	энциклопедии
		man	справочники
liv	коммуникация		
		spk	разговорная речь
		wri	письменная коммуникация

- Genre: жанр текста

	Genre
ver	стихи
son	песня
scd	театральный сценарий
scf	киносценарий
scr	перезапись радиопередач
nov	роман
col	рассказ
ess	очерк
mem	воспоминания
let	письмо
chr	хроника
dis	диссертация
std	студия
tcl	статья
rfl	рассуждение
ref	реферат
lct	доклад

- Subgenre: точнее определяет жанр текста для жанров *nov, col, ess*

	Subgenre
crm	детективный роман
scf	фантастика
bel	художественная литература
jun	литература для молодёжи
trv	путёвые записки
fac	литература факта

- Domain

	Domain		Subdomain
ars	художественная наука		
		mus	музыка, балет, опера
		cin	фильм
		arc	архитектура
		art	изобразительные исскуства, фотография, скульптура
		the	театр,театральные науки, критика
		lit	литература
hum	гуманитарная наука		
		his	история, археология
		psy	психология
		edu	педагогика
		soc	социология, коммуникация
		phi	философия
		inf	библиотековедение
		pol	политология
		lin	языкознание
		eth	этнология, этнография

	Domain		**Subdomain**
		cul	культурология
law	право		
		jur	правосудие
nat	естественные науки		
		agr	сельское хозяйство
		med	медицина
		zoo	зоология
		bot	ботаника
		bio	биология
		che	химия
		mat	математика
		ggr	география
		phy	физика
		met	метеорология
		geo	геология
		env	экология
tec	техника		
		tra	транспорт, средства связи
		ene	энергетика
		ind	промышленность
		com	информатика
		bui	строительство
		sta	нормализация
ecn	экономия, управление		
		eco	экономия, банковое дело, торговля
		mng	управление, менеджмент
		mer	товароведение
blf	вера, сверхъестественное		
		rel	религия, вера, секты
		teo	теология
		exc	магия, сверхъестественное
lif	быт		
		hou	домашнее хозяйство
		fsh	мода
		spo	спорт, физкультура
		sct	общественная жизнь
		amu	хобби, свободное время
		min	национальности
		reg	регионы
		cnl	консультация
		clt	культура

- Medium:происхождение текста

	Medium
lib	книга
nws	газета
jou	журнал
ste	лекции
net	интернет и другие компьютерные сети
for	формуляр
occ	документы вышедшее по случаю
npu	рукопись
tvf	телевидение
rad	радио

- Authsex: пол автора

	Authsex
msc	мужской
fem	женский

- Lang: язык документа

  Всегда „slk“ (словацкий)

- Varieta: вариант языка

	Varieta
std	стандартный
nst	нестандартный (диалект, архаичный, сленг...)

- Paragraphs: сообщает сохранены-ли в тексте правильно абзацы

tru	да, абзацы сохранены
fls	нет, информация о абзацах отсутствует

- Emphasis: сообщает сохранена-ли в тексте информация о изменении начертания текста (напр. жирный, курсив, другой шрифт)

tru	да, информатиция сохранена
fls	нет, информация отсутствует

- Diacritics: сообщает сохранены-ли в тексте диакритические знаки

tru	да, диакритические знаки сохранены
fls	нет, диакритические знаки отсутствуют

- Transsex: пол переводчика. Смотри Authsex.
- Origlang: оригинальный язык документа. Следует норму ISO 639-2.
- Date: число издания
- Dateorig: число первого издания (возникновения документа)
- Conglomerate: множество документов к которому этот принадлежит (например один месяц газеты)
- Bogocong: сокращённое название Coglomerate.
- Comment: Произвольный комментарий

Position: 1.		2.		3.		4.		5.		6.		7.	
S	Substantives		**Paradigm**		**Gender**		**Number**		**Case**				
		S	substantive	**m**	masc. animate	**s**	singular	**1**	nominative				
		A	adjective	**i**	masc. inanim.	**p**	plural	**2**	genitive				
		F	mixed	**f**	feminine	**o**	unknown	**3**	dative				
		U	incomplete	**n**	neutrum			**4**	accusative				
								5	vocative				
								6	locative				
								7	instrumental				
								o	unspecified				
A	Adjectives		**Paradigm**		**Gender**		**Number**		**Case**		**Grade**		
		A	adjective	**m**	masc. animate	**s**	singular	**1**	nominative	**x**	positive		
		F	mixed	**i**	masc. inanim.	**p**	plural	**2**	genitive	**y**	comparative		
		U	incomplete	**f**	feminine	**o**	unknown	**3**	dative	**z**	superlative		
				n	neutrum			**4**	accusative				
				o	unspecified			**5**	vocative				
								6	locative				
								7	instrumental				
								o	unspecified				
P	Pronouns		**Paradigm**		**Gender**		**Number**		**Case**		**Agglutinated**		
		S	substantive	**m**	masc. animate	**s**	singular	**1**	nominative	**g**	agglutinated		
		A	adjective	**i**	masc. inanim.	**p**	plural	**2**	genitive				
		P	pronoun	**f**	feminine	**o**	unknown	**3**	dative				
		F	mixed	**n**	neutrum			**4**	accusative				
		U	incomplete	**o**	unspecified			**5**	vocative				
		D	adverbial	**h**	general			**6**	locative				
								7	instrumental				
								o	unspecified				
N	Numerals		**Paradigm**		**Gender**		**Number**		**Case**				
		S	substantive	**m**	masc. animate	**s**	singular	**1**	nominative				
		A	adjective	**i**	masc. inanim.	**p**	plural	**2**	genitive				
		N	numeral	**f**	feminine	**o**	unknown	**3**	dative				
		F	mixed	**n**	neutrum			**4**	accusative				
		U	incomplete	**o**	unspecified			**5**	vocative				
		D	adverbial					**6**	locative				
		X	solitaire use					**7**	instrumental				
								o	unspecified				
V	Verbs		**Form**		**Aspect**		**Number**		**Person**		**Gender**		**Negation**
		I	infinitive	**d**	perfective	**s**	singular	**a**	first	**m**	masc. animate	**+**	affirmative
		K	indicative	**e**	imperfective	**p**	plural	**b**	second	**i**	masc. inanimate	**–**	negative
		M	imperative	**j**	ambivalent			**c**	third	**f**	feminine		
		H	transgressive							**n**	neutrum		
		L	*l*-participle							**o**	unspecified		
		B	futurum							**h**	general		
G	Participles		**Type**		**Gender**		**Number**		**Case**		**Grade**		
		k	active	**m**	masc. animate	**s**	singular	**1**	nominative	**x**	positive		
		t	passive	**i**	masc. inanim.	**p**	plural	**2**	genitive	**y**	comparative		
				f	feminine	**o**	unknown	**3**	dative	**z**	superlative		
				n	neutrum			**4**	accusative				
				o	unspecified			**5**	vocative				
								6	locative				
								7	instrumental				
								o	unspecified				
D	Adverbs		**Grade**										
		x	positive										
		y	comparative										
		z	superlative										
E	Prepositions		**Form**		**Binds with**								
		v	vocalised	**2**	genitive								
		u	non-vocalised	**3**	dative								
				4	accusative								
				6	locative								
				7	instrumental								
				o	unspecified								
O	Conjunctions	**Y**	contains conditional morpheme *by*										
T	Particles	**Y**	contains conditional morpheme *by*										

J Interjection **R Reflexive particle/pronoun *sa, si*** **Y Morpheme *by*** **Z Punctuation** **W Abbreviation** **Q Unknown POS type**	**# Not a word** **% Citation element (e.g. foreign language word)** **0 Digits**

Таб.1) Полное описание морфосинтаксических помет в Словацком национальном корпусе

```
<tok>
  <orth>meč</orth>
  <disamb>
    <base>meč</base>
    <ctag>SSis1</ctag>
  </disamb>
  <lex>
    <base>meč</base>
    <ctag>SSis1</ctag>
  </lex>
  <lex>
    <base>mečať</base>
    <ctag>VMesb+</ctag>
  </lex>
</tok>
```

Фрагмент текста в корпусе в XCES формате

Литература

1. *Metódy výskumu a opisu lexiky slovanských jazykov*. Materiály zo sympózia konaného v rámci 7. zasadnutia Lexikologicko-lexikografickej komisie pri Medzinárodnom komitéte slavistov (Nové Vozokany 24. – 26. April 1989). Edited by V. Blanár. Bratislava: Jazykovedný ústav Ľ. Štúra SAV (1990)
2. Benko, V.: Slovak Language Lexical Database. In: Computational Lexicography. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary (1991) 5–8
3. Jarošová, A.: *Korpus textov slovenského jazyka.* In: Slovenská reč 2 (1993) 89–95
4. Šimková, M.: *Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus*. Budmerice: Počítačová podpora prekladu (2003)
5. Ide, N., Bonhome, P., Romary, L.: *XCES: An XML-based Encoding Standard for Linguistic Corpora*. In: Proceedings of the Second International Language Resources and Evaluation conference. Paris: European Language Resources Association (2000)
6. Hajič, J., Hladká, B.: *Czech Language Processing - POS Tagging*. In: Proceedings of the First International Conference on Language Resources and Evaluation. Granada,

Spain (1998) 931–936
7. Hajič, J., Hric, J., Kuboň, V.: *Machine Translation of Very Close Languages*. In: Proceedings of the ANLP 2000. Seattle, U.S.A. (2000) 7–12
8. Sedláček, R.: *Morfologický analyzátor češtiny*. PhD. thesis. Faculty of Informatics, Masaryk University Brno, (1999)
9. Erjavec, T. et all.: *The MULTEXT-East Morphosyntactic Specifications for Slavic Languages.* In: Proceedings of the EACL 2003 Workshop on the Morphological Processing of Slavic Languages.
10. http://www.textforge.cz
11. Przepiórkowski, A.: *The IPI PAN Corpus preliminary version.* Instytut Podstaw Informatyki PAN, Warsaw, Poland (2004)
12. Zakharov, V., Volkov, V.: *Morphological Tagging of Russian Texts of the XIXth Century.* In: Text, Speech and Dialogue. Proceedings of the 7th International Conference TSD 2004. Brno, Czech Republic (2004) 235–242
13. Forróová, M., Horák, A.: *Morfologická anotácia korpusu.* In: Slovenčina na začiatku 21. storočia. Prešovská univerzita, Prešov, Slovakia (2004) 174–183
14. Gianitsová, L.: *Dynamika skloňovania substantív (na materiáli SNK).* In: „Tzv. základní výzkum v lingvistice – desideratum, nebo realis?“ V. mezinárodní setkání mladých lingvistů (Olomouc 17 – 19 May 2004). To be published.