М. Шимкова, Р. Гарабик

СИНТАКСИЧЕСКАЯ РАЗМЕТКА ТЕКСТОВ СЛОВАЦКОГО НАЦИОНАЛЬНОГО КОРПУСА

Введение

Словацкий национальный корпус[1] (СНК) в настоящее время доступен в интернете со стилево-жанровой аннотацией, лемматизирован, а также имеет полную автоматическую морфологическую разметку. Актуальная версия prim-2.1 (primary) содержит более 300 миллионов слов (токенов). Для автоматизированной морфологической разметки использовался меньший корпус текстов, которые были размечены вручную на базе системы тегов[2], разработанной в СНК. Этот корпус (r-mak-1.0) также доступен в интернете. С июня 2005 г. в Словацком национальном корпусе началось аннотирование данных текстов и на синтаксическом уровне. Синтаксическая разметка вместе с морфологической позволяет проводить другие исследования с использованием метод статистического анализа, и является полезной базой данных синтаксических структур и атрибутов. Особенно значительны эти разметки в настоящее время для словацких лингвистов, которые составляют новое описание грамматики современного словацкого языка (морфологического и синтаксического уровней) и ведут работы над 8-томным Словарем современного словацкого языка.

Морфологическая разметка

Тексты в Словацком национальном корпусе автоматически

---

[1] **http://korpus.juls.savba.sk**

[2] *Garabík R., Gianitsová L., Horák A., Šimková M.* Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu // URL: **http://korpus.juls.savba.sk/publications/block2**

лемматизированы и морфологически размечены. Система морфологических меток (тегов) описывает все грамматические категории слова, со стремлением к чёткой и ясной форме строки. Каждой отдельной грамматической категории соответствует один символ, и одному символу соответствует только одна определённая категория. На первой позиции находится знак части речи (включая символы для сокращений, знаков препинания, цифр, иностранных слов и неопределимых или неправильных элементов текста). Далее следуют символы, которые обозначают остальные обыкновенные грамматические категории, но и (на второй позиции) парадигму склонения – некоторые слова не склоняются (иностранные слова) или склоняются не по своей парадигме (имена существительные как прилагательные, например, *hlavný (главный* [бухгалтер]*)*, *odsúdený* (*подсудимый)*, местоимения и числительные как имена существительные или прилагательные и т.д.).

Синтаксическая разметка

Синтаксическая разметка СНК использует программное обеспечение и следует принципам Пражского корпуса зависимостей (Prague Dependency Treebank)[3]. В синтаксической разметке различаем два уровня. Уровень аналитический, на котором размечена поверхностная структура предложения (подлежащее, сказуемое и т.д., также союзы, предлоги и т.д., знаки препинания и другие несловесные знаки), и уровень тектограмматический, отражающий глубинную структуру предложения (agens, patiens, валентность, тема, рема высказывания и т.д.). Синтаксическая разметка состоит из дерева зависимостей, где узлам соответствуют отдельные единицы текста (слова и знаки препинания на аналитическом уровне). В настоящее время в СНК проводится синтаксическая разметка на аналитическом уровне – в этом

[3]Anotace na analytické rovině. Návod pro anotátory / *Hajičová E.*, *Sgall P*. (eds.). Praha, 1999. URL: **http://ufal.mff.cuni.cz/pdt**

субкорпусе почти 35 000 предложений, содержащих 570 000 слов. Все предложения аннотированы дважды, значит, у нас синтаксически размечено почти 70 000 предложений. Тексты те самые, которые были размечены вручную на морфологическом уровне: художественные – 78%, научные – 13%, публицистические – 9%.

Разметку на аналитическом уровне называем и структурной, классической, но от общеизвестной она отличается некоторыми особенностями. Кроме уже упомянутых (разметка несинтаксических слов, т.е. предлогов, союзов, частиц, междометий и также всех несловесных знаков) это, например, определение сказуемого придаточных предложений не меткой предиката, а меткой типа придаточного предложения, не различение согласованного и не согласованного определения особенным символом, а только позицией, различение специальных функций местоимения *to (это)* и т.д.

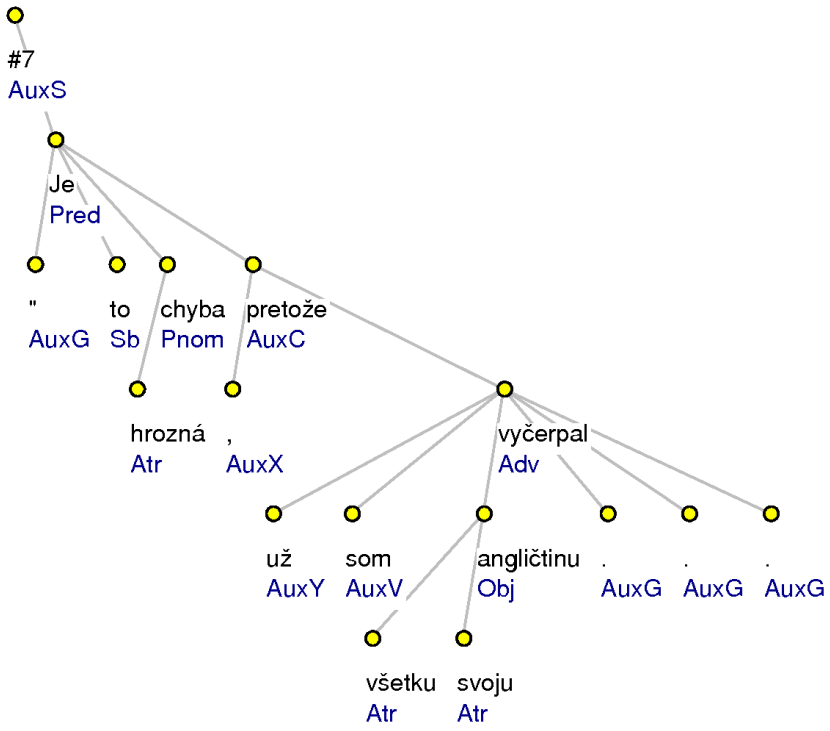

Рис. 1: Пример размеченного предложения «Je to hrozná chyba, pretože už som všetku svoju angličtinu vyčerpal…» («Это кошмарная ошибка, так как я уже весь свой английский исчерпал…)

Сравнение аннотаций

Аннотация каждого файла (содержавшего около 10 – 30 – 50 предложений в зависимости от типа и длины текста) проводится двумя аннотаторами. Все эти предложения исправляются технически, т.е. исправляются неправильные разметки однозначных узлов и атрибутов вместе с однозначными зависимостями – эти

ошибки возникают в связи с усталостью и недостаточной внимательностью аннотатора, иногда и различием между чешским и словацким подходами. После эти две аннотации сравниваются и на их основе создаётся третья версия дерева зависимостей, вручную исправленная ответственным аннотатором.

Различия между двумя аннотациями могут быть или различия между атрибутами узлов, или различия в структуре дерева зависимостей. Количество совпадений узлов в аннотации указано в матрице в табл. 1): диагональ соответствует тем же самым атрибутам, элементы вне диагонали обозначают количество различий.

	Adv	Apos	Atr	AuxC	AuxO	AuxP	AuxR	AuxT	AuxV	AuxX	Coord	ExD	Obj	Pred	{OTH}
[[Adv	13581	2	1203	284	14	88	44	13	6	9	24	501	1827	763	1741]
[Apos	0	219	2	0	0	0	0	0	0	87	137	13	0	5	121]
[Atr	0	0	17226	39	32	12	4	6	8	5	4	281	1401	763	1717]
[AuxC	0	0	0	4064	6	51	0	1	2	18	123	25	192	2	892]
[AuxO	0	0	0	0	48	45	48	28	1	5	6	2	59	2	78]
[AuxP	0	0	0	0	0	10358	101	85	4	3	7	22	22	9	325]
[AuxR	0	0	0	0	0	0	700	2281	25	0	0	1	351	0	12]
[AuxT	0	0	0	0	0	0	0	949	100	0	1	3	222	2	28]
[AuxV	0	0	0	0	0	0	0	0	1056	1	0	4	36	81	95]
[AuxX	0	0	0	0	0	0	0	0	0	9383	953	261	3	4	279]
[Coord	0	0	0	0	0	0	0	0	0	0	4724	112	1	47	1246]
[ExD	0	0	0	0	0	0	0	0	0	0	0	2106	381	141	1452]
[Obj	0	0	0	0	0	0	0	0	0	0	0	0	12416	755	1714]
[Pred	0	0	0	0	0	0	0	0	0	0	0	0	0	11149	400]
[{OTH}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28948]]

Таблица 1: Соответствие атрибутов, используемых аннотаторами. На диагонали сходные атрибуты.

Заметно, что много ошибок связано с атрибутами Adv-Atr, Adv-Obj, которые обыкновенно трудно анализируются, и с атрибутах AuxO-AuxR, AuxX-Coor, AuxR-AuxT, которые обозначают особые элементы текста (некоторые местоимения, частицы, запятая и т. д.).

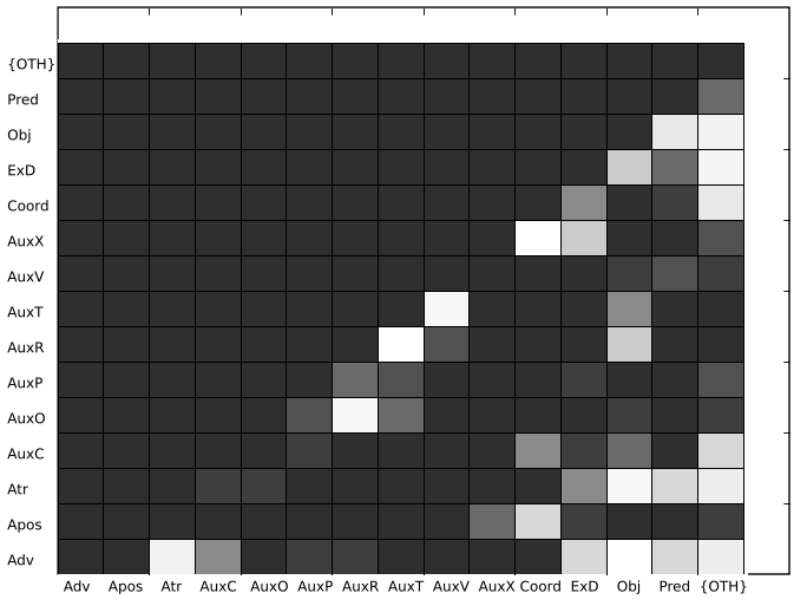

Рис. 2: Изображение различий между аннотациями атрибутов. Светлые пятна обозначают высокое число различий.

Следующий этап исследования

В первую очередь мы будем продолжать разметку (минимум до 50 000 предложений) и исправлять ошибки в уже размеченных предложениях. Необходимо сделать корпус доступным в интернете, с использованием клиент-сервер системы NetGraph для сложных поисков, и простого интерфейса для быстрого просмотра разметки. Пример синтаксической разметки Словацкого национального корпуса находится на странице

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