Морфологічна анотація текстів Словацького національного корпусу

Аґата Карчова
Марія Шимкова

Інститут мовознавства імені Л. Штура
Словацька академія наук
813 64 Братислава, Словацька республіка

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

Abstract. From 2004 to 2005, four different kind of texts have been manually disambiguated and morphologically annotated: Orwell's 1984 novel, selected Slovak Wikipedia texts, Internet magazine InZine and SME daily newspaper articles. Empirically estimated precision of manual annotation is presented, together with description of the annotation process. The result is high quality manually annotated corpus of Slovak language texts, containing over 300 000 words.

Одночасно із першими кроками побудови словацького корпусу текстів в 2002 – 2003 рр. розпочалася підготовка морфологічної анотації текстів Словацького національного корпусу (СНК). Паралельно зі збиранням текстів та їх первісним опрацюванням формувалася внутрішня структура корпусу, йдеться про технічний аспект, та здійснювалося стильово-жанрове опрацювання текстів. Важливою справою було придбання придатного корпусного програмного засобу чи менеджера. Не менш важливим було також запитання: розпочинати морфологічну анотацію чи ні, і якщо так, то яку саме?

На момент початку робіт над створенням СНК у Словаччині комп'ютерна або корпусна лінгвістика як галузь науки не були достатньо розбудованими і ще не існували відповідні до морфологічної анотації програмні забезпечення, не кажучи вже про лінгвістичні описи, які можна було б використати для комп'ютерного опрацювання словацької мови. У той же час у світі ті, хто створювали та використовували електронні корпуси, щораз більше орієнтувалися на лінгвістичні анотації корпусів, на відміну від минулого, коли корпуси будували на взірець комп'ютерних баз даних слів чи текстів із майже виключно лексикографічним призначенням, а лінгвістичні інформації, які вносилися в корпуси, вважалися зайвими через змінність граматичних категорій. Лінгвістичне анотування корпусів стимулювало появу знаряддя для морфологічної анотації флективних мов. Таке ж знаряддя безпосередньо для словацької мови було створено на Фізико-математичному факультеті Карлового університету в Празі, і

виникло як другорядний продукт знаряддя на лематизацію, аналіз та зняття неоднозначності для чеської мови, яке застосовували в Чеському національному корпусі.

На першому етапі побудови Словацького національного корпусу – етапі підготовки морфологічної анотації текстів, – було здійснено об'єднання вже наявних підходів до морфологічних анотацій корпусних текстів (Forróová, Horák 2004), проведено тестове перевіряння чеського знаряддя, яке для нашого відділу СНК віддала на підставі договору вже згадувана інституція, а також вирішено, яким чином здійснюватиметься анотація текстів СНК.

На підставі тестового аналізу празького програмного забезпечення було виявлено, що:

а) програмне забезпечення, створене на початку 90-х років у межах проекту Чеського національного корпусу SYN 2000, було доволі часто критиковане за надто строгий технічний підхід; за відсутність мітки для категорії виду у дієслова, яка є важливою специфічною ознакою більшості слов'янських дієслів; під критику підпав також позиційний спосіб запису морфологічних знаків, який є важкий для читання та розшифровування, і наявність помилок при анотуванні за статистичним принципом, наприклад, часто вживане чеське слово, яке функціонує як займенник або сполучник *jak* було переважно визначуване як іменник зі значенням «назва тибетського тура» тощо;

б) словацьке знаряддя для анотації, яке по суті виникло як другорядний продукт, базувалося на основі словника, який не був цілком коректний, а саме тому помітно знижувалася успішність анотації;

в) словацька мовна система та словацька граматична теорія в порівнянні із чеською мовою має кілька суттєвих розбіжностей.

Виходячи зі сказаного, ми дійшли до таких висновків: чеське знаряддя для анотації буде використане в першій фазі на автоматизоване попереднє анотування словацьких текстів та для внутрішніх досліджень, поки не буде створено власного анотаційного знаряддя, а також вручну анотований корпус достатнього розміру. Таке рішення означало підготовку власного таґсету (правил токенізації, лематизації та маркування), який розпочали розробляти приблизно в середині 2003 р.

Засадничим обрано формально-морфологічний підхід, скорочений – атрибутивний запис, у якому кожний тип мітки має свою власну стабільну позицію, і звичайний поділ слів на окремі частини мови. Словацький теґсет, який був захищений в кінці 2003 р. нашими лінгвістами в рамках проекту Словацького національного корпусу і

представлений на міжнародній конференції SLOVKO 2003 (Forróová, Garabík, Gianitsová, Horák, Šimková 2005), містить дев'ятнадцять класів слів, з яких десять відповідає звичайним частинам мови а дев'ять інших класів слів позначає специфічні мовні елементи, наприклад інтерпункцію, скорочення та ремарки, цитатні звороти запозичені з інших мов, цифри, елементи, які не входять до категорії слів. Окремо відділяються формальні дієприкметники типу *napísaný, prijatý*, що з одного боку ускладнює анотацію на морфологічному рівні, але з іншого боку є вдалою вихідною точкою для подальшого опрацювання текстів. Навпаки певним полегшенням є спосіб інтерпретації слів *sa*/*si*, які мають різні значення і різні функції в системі морфології. За нашими правилами вони завжди вважаються зворотними займенниками і мають спеціальну мітку «R». Специфічною рисою теґа в іменних частинах мови є встановлювання типу парадигми на другій позиції мітки: окремо так відзначаються іменники з прикметниковою парадигмою (*obžalovaný, vedúci* – частина мітки «SA»), прикметники із не повною парадигмою (*super* – частина мітки «AF») тощо. Наш теґсет можна знайти за адресою http://korpus.juls.savba.sk/.

Другий етап (лютий 2004 р. – жовтень 2005 р.) був етапом ручної анотації з використанням словацького теґсету, розвитку та модифікації анотаційних знарядь. Паралельно відбувалося удосконалення в теґсету, а саме, вносилися до нього необхідні зміни та доповнення.

Першим текстом обраним для анотації став переклад новели Джорджа Орвелла «1984». Таке рішення мотивовано двома чинниками, по-перше, ця новела була частиною міжнародного проекту Multex-East, по-друге, існував чеський сегмент проекту Multex-East з вдало зробленою ручною анотацією, який за необхідності можна було б використати для порівняння і тестової перевірки можливостей автоматизованої анотації (Hlaváčová, 2005).

Анотацію вручну виконували студенти факультетів словацької мови університетів в містах Пряшів, Ружомберок та Братислава. Але анотації зроблені під час навчання не дали достатніх результатів, вдалішим та успішнішим виявилося індивідуальне анотування обраними студентами у відділенні СНК поза навчанням, тобто студенти працювали на підставі угоди про виконану роботу і за свою роботу одержували відповідну платню. Анотовані тексти ще проходили фазу остаточного виправлення, тому текст анотований та виправлений вручну був готовий аж у середині 2005 р.

Після анотації перекладу новели Джорджа Орвелла «1984», яка містить 102 000 токенів, ми вибрали для морфологічної ручної анотації тексти із щоденної газети SME, інтернетного журналу InZine та словацької версії інтернетної енциклопедії Wikipedia.

Таке рішення мотивовано тим, щоби вручну утворений анотований корпус складався із текстів різноманітних стилів, оскільки його призначення використовуватися для тренування маркування великого корпусу. У червні, липні та серпні 2005 р. було здійснено морфологічну анотацію текстів обсягом приблизно 115 000 токенів, а саме тексти SME разом містили 21 268 токенів поділених до 51 файлів, InZine містив разом 42 703 токенів в 51 файлах та Wikipedia 50 077 токенів в 48 файлах.

На початковому етапі із згадуваних джерел були вибрані ті тексти, які були найоптимальнішими для анотування, тобто такі, які були написані як одне ціле, словацькою мовою, з оптимальним числом слів і ці тексти були поділені на окремі файли. Файл містив найменше 100 токенів, максимальна кількість токенів у файлі становила 6 500. Найкращою з погляду зосередженості на анотації та організації процесу анотування виявилася довжина файла приблизно 1 500 токенів.

Відібрані тексти були перетворені на так звані анотаційні файли, які відповідали XML-стандарту. У нашому відділенні була створена спеціальна GUI-програма написана за допомогою програмної мови Python, котра використовує для розбору XML-файлів ElementTree бібліотеки. Програма зображає реєстр усіх слів у файлі і для кожного обраного слова список відповідних лем та міток. Анотатор має можливість обрати до токена відповідну пару лема-мітка, або самостійно доповнить відповідну пару. Цей метод попереднього позначення текстів великою мірою допомагає анотаторам і пришвидшує їх роботу (≈300 токен $\cdot$ година$^{-1}$).

Це анотування здійснювало десять студентів Філософського факультету Університету імені Коменського в Братиславі. Переважна більшість файлів опрацьовувалася двічі, тобто той самий файл був окремо анотований двома студентами, щоб можна було пізніше порівнювати між собою дві анотації. Як правило анотатори робили помилки в приписувані міток, інколи лема і мітка, що відповідали одному токенові, були помилковими. Помилки спричиняло також недостатність знань про внутрішні закономірності граматичної системи словацької мови, а також внутрішньомовні неоднозначності, наприклад не цілком ясна межа при встановленні незмінних частин мови. Студенти також повинні були достосувати свої знання до обраного нами методу, який в деяких пунктах відрізняється від класичної морфології, а деякі правила доповнюють її, тому анотування виявилося для студентів не надто простою справою. Важливим фактором появи помилок були і інші обставини, наприклад втома анотаторів – анотація є працемісткою роботою, яка вимагає постійного зосередження уваги.

Часто помилковим було встановлення відмінка іменника у випадках, коли йшлося про омоформи, інколи при дієсловах був неправильно визначений знак афірмації або негації, багато разів помилково визначався вид дієслів, часто сплутувався числівник із займенником, частка з прислівником, помилково встановлювався рід іменників та тип парадигми змінюваних частин мови. У мітках інколи бракувало обов'язкової частини, наприклад встановлення ступеня порівняння у прикметниках та дієприкметниках тощо.

Документи	**кількість токенів**	**кількість словесних елементів**	**порівняння**	**кількість помилок [%]**		
				лема	**мітка**	**обидва**
SME	21 268	77,35 %	A_1 - Б	0,87	7,87	0,33
			A_2 - Б	1,08	7,21	0,40
			A_1 - A_2	1,73	11,45	0,63
WIKI	50 077	82,86 %	А - Б	1,01	6,74	0,50
INZINE	42 703	82,01 %	A_1 - Б	1,23	9,40	0,51
			A_2 - Б	1,68	11,07	0,96
			A_1 - A_2	2,52	14,56	1,36

Таблиця 1. (A_1 - перший анотатор, A_2 – другий анотатор, Б – остаточно виправлений текст)

Зважаючи на велику кількість помилок ми вважали необхідним додатково виправляти оці помилки. Перевірка текстів, здійснювана протягом серпня – жовтня 2005 р., складалася з трьох окремих фаз. За допомогою наших власних напівавтоматизованих знарядь було можливе швидке та надійне виявлення й виправлення більшості помилок.

У першій фазі були знайдені такі помилки: мітка не мала правильної довжини (не була складена з усіх обов'язкових знаків – деякого знаку бракувала, а деякий був зайвим), у лемі фіксувався пробіл між буквами або велика буква (лема повинна складатися із малих букв), у мітці була неправильна комбінація декотрих знаків (не правильне встановлення роду, числа тощо). Результати пошуку помилок були зображені в наочних таблицях.

номер рядка	**токен**	**лема**	**мітка**	**опис помилки**
126	nejakej	nejaký	PAfs6x	Bad length
267	"	"	Z	Spaces in lemma/orth
503	bola	byť	VLesc+	Bad length
651	mi	ja	PPms3	Bad gender
1333	tučné	tučný	AAfp4	Bad length

1456	sú	byť	VKefp+	Bad number
2500	zažili	zažiť	VLdpbm+	Bad gender
2932	—	—	Z	Spaces in lemma/orth
2990	niekoľko	niekoľko	PU	Bad length
3057	deviatich	deväť	NUip2w	Bad length
3332	ich	on	PPmp4	Bad gender
5856	pohrá	pohrať	VKmsc+	Bad aspect
6057	slúžia	slúžiť	VKepci+	Bad length
7818	II	II	C}----------	Bad POS
11637	,	,	None	Not string

Таблиця 2 – приклад результатів із програми виконуючої порівняння

Друга фаза виправлення полягала в порівнянні відмінностей між текстами, анотація яких була виконана незалежно один від одного двома різними анотаторами. Таким чином виявилися всі випадки, у яких студенти не узгоджувалися у встановленні леми або мітки. Також були знайдені важливі помилки, які випливали із недостатнього порозуміння правил теґсету. Виявлено велику кількість неправильних інтерпретацій, але не всі, тому що в декотрих випадках помилялися обидва анотатори.

1	Zakázaný	zakázaný	AAis1x
	Zakázaný	zakázaný	Gtis1x
7	BRATISLAVA	bratislava	SSfs1:r
	BRATISLAVA	bratislava	SSfs1
23	osobných	osobný	AAip2x
	osobných	osobný	Gtip2x
40	ho	on	PFis4
	ho	jeho	PUis4
54	autora	autor	SSms2
	autora	autora	SSms2
57	B	b	Q
	B	B	Wms1
59	Manna	mann	SSms2:r
	Manna	Mann	SSms2:r
64	je	nebyť	VKesc-
	je	byť	VKesc+
74	SAV	sav	W?
	SAV	sav	Wfs2
95	aj	aj	O
	aj	aj	T

Таблиця 3 – приклад результатів порівняння анотацій між двома анотаторками

У третій фазі знов йшлося про порівняння двох текстів. У цей час ми уже мали і могли використовувати остаточно виправлений текст Дж. Орвелла «1984». На підставі порівняння триґрам токен – лема – мітка до таблиці були виписані всі відмінності між текстами, а не тільки неправильні, але й такі, яких у тексті новели не виявилося. Словниковий запас слів новели є само собою обмежений – лише 40 % триґрамів знаходилося в наступних текстах, тому ми ще остаточно розподіляли мітки за частинами мови, щоб читання інформацій в таблицях прискорилося.

Після закінчення цих фаз перевіряння всіх текстів було зреалізовано і особлива увага приділялася так званим проблемним словам, яким виявилися частки, сполучники, дієприкметники, номери, частини текстів іншомовного походження, у яких ще інколи знаходилися помилки і, відповідно, які були остаточно виправлені.

На початок листопада 2005 р. був відкритий доступ до нової версії СКН prim-2.0 (первісний корпус), який є більшим за кількістю текстів (понад 250 мільйонів токенів), а також першу версію автоматизованої морфологічної анотації словацькими мітками.

Наша система опрацювання текстів є гібридна: для лематизації було використано первісне чеське програмне знаряддя, для маркування мітками та розпізнавання омонімів було використано TNT (Brants), натреноване на вище згадуваному тренувальному корпусі анотованому вручну обсягом ≈130 000 токенів. За інформаціями автора цього знаряддя, для статистичної анотації на базі наявності триґрамів у наперед анотованому зразку, зреалізованому на англійських та німецьких текстах, це знаряддя при 100 тисячному тренувальному корпусі досягає успішність 95 %, при 200 тисячному корпусі успішність 96 % і т.д. Правда, словацька мова типологічно відрізняється від німецької та англійської мови і успішність знижується передусім через високу частоту омонімічних форм, тому успішність першої версії автоматизованої анотації була приблизно 90 %.

Перші фільтри показали, що в автоматизованих анотаціях повторюються переважно помилки подібні до тих, які з’являються у анотованих вручну текстах, наприклад, не правильно позначається відмінок у омонімічних формах, помилковим є встановлення мітки для істот та неістот в іменниках чоловічого роду, не правильно встановлюється вид дієслів, не розрізняються формальні дієприкметники та прикметники. Окремі проблеми випливають із неоднозначності займенників, а також встановлення афірмації та негації в дієсловах та ступінь порівняння прикметників та прислівників, проблеми випливають із частої омонімії незмінних частин мови – ці слова краще анотувати вручну, за допомогою списку слів укладених за окремими ознаками, наприклад, займенники із іменниковою парадигмою, окремою займенниковою,

змішаною, не повною парадигмами та прислівниковою парадигмою, і також списків виключних – неомонімічних часток, прийменників тощо.

Під час автоматизованої анотації ще додалося технічних помилок, які були спричинені:

а) різницею обраних підходів: за словацькою лематизацією для дієслова в заперечній формі була встановлена лема в заперечній формі; власні імена, хоча починаються з великої букви, лематизуються з малою початковою буквою; дієприкметники лематизуються як прикметники, а не як дієслова. У корпусі анотованому вручну ці правила повинні бути дотримані, в корпусі, який був лематизований та анотований автоматично, леми встановлюються за чеським знаряддям, а тому в наведених випадках знаходяться відмінності, що інколи спричиняє також помилку в мітці;

б) не достатнім виявився розмір тренувального корпусу, в якому не виявилося, наприклад, частки *azda*, і тому була знаряддями визначена її належність до будь-якої частини мови, так само вигуки позначалися довільно. Наприклад до власного іменника-прізвища *Štúr* (Л. Штур був основоположником словацької літературної мови) була додана лема дієслова *štúrať* тощо; менш зрозумілими помилками були не системно призначувані відмінки чи інші граматичні категорії в триґрамах, наприклад, в іменному складі слів *na maďarský národný štát* прийменнику був приділений 6. відмінок, першому прикметнику 1. відмінок, другому прикметнику та іменнику 4. відмінок.

Третій етап морфологічної анотації текстів СНК, що власне зараз відбувається, передбачає поповнення анотованого вручну корпусу виправленнями вибраних текстів автоматично анотованих таким чином, щоб він досяг розміру найменше 1 мільйон токенів. З середини лютого 2006 р. вже існує відкритий в Інтернеті підкорпус анотований вручну з обсягом 322 600 слів, що представляє 56 465 окремих словесних форм, 26 498 лем та 1 428 різних міток, на якому знов тренувалася автоматизована анотація, результатом чого була нова доступна версія автоматизовано лематизованого корпусу prim-2.1.

Загалом, крім використання нашого корпусу для наступного опрацювання словацької мови, якісна морфологічна анотація може використовуватися, а в цьому часі вже й використовується, для граматичних досліджень, зокрема для створення нової морфології словацької мови. Тексти анотовані вручну далі передаємо на синтаксичну анотацію, яка цілком реалізується за принципами Празького корпусу.

Бібліографія

BRANTS, T.: *TNT – Statistical Part-of-Speech Tagging*
URL: http://www.coli.uni-saarland.de/~thorsten/tnt

FORRÓOVÁ, M., HORÁK, A.: Morfologická anotácia korpusu. In: *Slovenčina na začiatku 21. storočia.* Ed. Mária Imrichová. Prešov : Prešovská univerzita, Fakulta humanitných a prírodných vied, 2004, s. 174 – 186.

GARABÍK, R., GIANITSOVÁ, L., HORÁK, A., ŠIMKOVÁ, M.: *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu*.
http://korpus.juls.savba.sk/publications/

GARABÍK, R., GIANITSOVÁ-OLOŠTIAKOVÁ, L.: Manual Morphological Annotation of Slovak Translation of Orwell's Novel 1984 – Methods and Findings. In: *Proceedings of International Conference Slovko 2005 – Computer Treatment of Slavic and East European Languages.* Ed. R. Garabík. Bratislava : Veda, 2005.

HAJIČ, J.: *Popis morfologických značek – poziční systém.* http://ucnk.ff.cuni.cz/manual/znacky.html

HAJIČ, J., KRBEC, P., KVĚTOŇ, P., OLIVA, K., PETKEVIČ, V.: *Serial Combination of Rules and Statistics: A Case Study in Czech Tagging.* Prague Dependency Treebank. CD ROM. V. 1.0. Praha : Ústav formální a aplikované lingvistiky MFF UK, 2001.

HLAVÁČOVÁ, J.: Orwell's 1984 – Playing with Czech and Slovak Versions. In: *Proceedings of International Conference Slovko 2005 – Computer Treatment of Slavic and East European Languages.* Ed. R. Garabík. Bratislava: Veda, 2005.