К некоторым вопросам, связанным с лемматизацией корпуса чешских текстов

Лингвистическая разметка языковых корпусов, как лексическая (лемматизация), так и морфологическая, упрощает поиск и лингвистическую работу с корпусом; особую важность она приобретает для языков с высокой степенью флективности. Во многих случаях, тем не менее, мы сталкиваемся с проблемами, однозначно решить которые затрудняются даже опытные лингвисты, но конкретное решение которых может непосредственным образом отразиться в различных языковых пособиях (напр. словарях), опирающихся на материал корпуса. В чешском языке к таким проблемным случаям относятся, напр., частное отрицание или разметка существительных, имеющих только формы множественного числа, и существительных, единственное и множественное число которых различаются семантически. Во всех приведенных случаях первостепенное значение имеет то, какая лемма будет закреплена за той или иной словоформой.

Разметка корпуса

В целях облегчения поиска тексты в корпусах по-разному размечаются. Важным условием поиска является предоставление данных о том, из какого текста почерпнут конкретный материал - т.е. предоставление обычной библиографической информации. Наряду с ней в ЧНК используются обозначения типа текста (напр. беллетристика, специальная литература, публицистика и т.п.), жанра (напр., путевые записки, история, право, сельское хозяйство, химия и т.п.) источник информации (книга, журнал, интернет), пол автора, в случае перевода - и переводчика, язык, на котором написано оригинальное произведение, и год издания. Эти обозначения в ЧНК служат для составления репрезентативного корпуса (100 млн. SYN2000: 60% публицистика, 25% специальная литература, 15% беллетристика), а по некоторым из них можно и вести поиск (тип текста, год, название произведения). Этот тип обозначений называется **внешней аннотацией**, их лингвисты вносят в специальную базу данных. Материалы этой базы данных программисты затем соединяют с текстом в форме так наз. **шапки**, имеющей формат SGML (Standard Generalized Markup Language). Одновременно фиксируется первоначальная структура всего текста посредством так наз. **структурных знаков** (обозначение предложений, абзацов, типов шрифта и т.п.).

Следующим типом разметки является так наз. лингвистическая разметка - **лемматизация** и **морфологическая характеристика слов текста (тэгги).** Этот процесс в ЧНК до настоящего времени производится с помощью автоматической статистической разметки , но в будущем он будет усовершенствован использованием разметки опирающейся на лингвистические правила.

В чешском языке, как и в остальных славянских языках, процесс лемматизации осложнен обширной омонимией словоформ. Напр., форма *metrem* может быть произведена от лемм имени существительного *metro* (средний род), *metr* (можской род), *metrum* (средний род), имеющих разный род. Еще больше проблем связано с определением отдельных морфологических категорий – напр., падежа: именительный и винительный падеж неодушевленных имен мужского рода имеют в единственном числе одинаковую форму у имен прилагательных и существительных: *nový dům*. Эту и так довольно сложную ситуацию может еще больше усложнить само определение звучания леммы у существительных, прилагательных и наречий, начинающихся с приставки *ne-,* обозначающей отрицание.

Если существует положительная форма, то все зависит от того, действительно ли имеется в виду противоположное значение или отрицательная форма является лексикализованной и ее значение является расширением или сдвигом значения. С этими вопросами мы столкнулись при создании частотного словаря чешского языка. Более подробно я рассмотрю данную проблематику на примере имен прилагательных.

Отрицательные прилагательные.

В чешском языке словесное отрицание образуется при помощи присоединения к слову приставки *ne-* (*umět-neumět, herec-neherec, přemožitelný-nepřemožitelný, sladce-nesladce*). При спряжении глагола она иногда переносится на вспомогательный глагол: *nebude umět x neuměl jsem*. У этих слов леммой является положительная форма. Чешский язык содержит, конечно, и слова, начинающиеся на *ne-*, которые после отрыва этого начала в языке не существуют: *nenávidět, nerost, neustálý, nejapně*. В таком случае при лемматизации нужно следить за тем, чтобы в результате устранения начального *ne-* не возникли несуществующие леммы. Простое распознавание прилагательных и наречий путем простого отделения начального *ne-* осложнено их превосходной степенью, которая начинается с *nej-*. Но превосходную степень в корпусе образует лишь 2% всех встретившихся прилагательных, 96% имеют первую степень. Прилагательные могут иметь также промежуточные варианты:

1) **слово существует в положительной форме (без начального *ne-*), но имеет иное значение**: *chutný - nechutný, smyslný- nesmyslný, mocný- nemocný*

Пример:

chutný 432 употреблений в ЧНК

Наиболее частотные существительные: *maso, jídlo,pokrm, oběd, večeře*

*Jiřina Bohdalová je skvělou kuchařkou a **chutným** pokrmům se odolává nejhůř. (Story 1997)*

nechutný 562 употреблений в ЧНК

Наиболее частотные существительные: *tahanice, scéna, kampaň, útok, věc,*

Marně si však kladu otázku, kdo vyprovokovával ony <**nechutné** tahanice> o název republiky. (Respekt 1990)

2) одно из значений формы с начальным *ne-* отличается от положительной формы:

falšovaný –nefalšovaný, počítaný – nepočítaný

Пример:

počítaný 106 употреблений в ЧНК

Наиболее частотные существительные: *(předmět, nit, hodnota, objekt)*

*V pádech pro <**počítaný** předmět> má čeština opravdu pestrost.* (Naše řeč 1997)

nepočítaný 14 употреблений в ЧНК

– фразеологизм: *dostat nepočítaných (na zadek)*

*A o to mi vlastně šlo , jinak bych schytala <**nepočítaných**> na zadek.*

(Kludská, Dagmar: Srdcové eso. 1996)

3) для выражения отрицания, или скорее противоположного качества, используется иная лексема, а в негативной форме произошел определенный сдвиг значения:

velký – malý – nevelký.

Пример:

velký 177667 употреблений в ЧНК

Наиболее частотные существительные: *(Británie, část, množství, počet, problém)*

malý 74 457 употреблений в ЧНК

Наиболее частотные существительные: *(dítě, strana, počet, množství,část)*

nevelký 1009 употреблений в ЧНК

Наиболее частотные существительные: *(počet, skupina, prostor, množství, zájem)*

*Dvorek byl **nevelký**, obdélníkový, vypadal stísněně mezi vysokými zdmi domu.*

(Pecka, Karel: Malostranské humoresky. Atlantis 1992)

4) обе формы или форма с отрицанием используются в качестве терминов:

movitý – nemovitý majetek nevidomý

Пример:

nevidomý 740 употреблений в ЧНК

Наиболее частотные существительные: člověk, hudebník, dítě, občan, chlapec

*Petr byl jediný **nevidomý**, ale přesto se mu věnovali alespoň dvakrát týdně v individuální výuce. (Lidové noviny 1993)*

5) негативная форма более частотна *nebezpečný, nedomyšlený*

Пример:

bezpečný 3741 употреблений в ЧНК

Наиболее частотные существительные: *zóna, vzdálenost, místo, provoz, jízda*

*V centru města je **bezpečná** zóna OSN (Lidové noviny 1995)*

nebezpečný 10 259 употреблений в ЧНК

Наиболее частотные существительные: *odpad, známost, látka, situace, místo()*

*Po dvou letech by měl projekt ukázat , jak nejvýhodněji by se dal **nebezpečný** odpad sbírat v celé Praze.*

(Mladá fronta 1994)

При формировании словника необходимо решить, должны ли оба слова образовать самостоятельную вокабулу или вокабульным словом может быть положительная форма. Дело в том, что вид словоформы определяет место слова в словнике, поэтому для пользователей словаря с алфавитным расположением материала оно очень важно.

В современной чешской лексикографической практике приводились обе формы как вокабулы в случаях a-d. Тип e нельзя было продемонстрировать, такую возможность дал только корпус. Решение данной проблемы при лемматизации корпуса является еще более важным, т.к. непосредственно сказывается на абсолютной частотности а других статистических показателях. Разумеется, и здесь нельзя всегда с абсолютной надежностью установить, где давать обе формы, а где оставить формы объединенными под одной леммой - в одной, чаще всего положительной форме. В некоторых случаях новое значение проявляется очень нечетко даже на основе корпусных данных.

Попыткой такого различения было отделение лемм типа e в новом частотном словаре, созданном на основе данных корпуса SYN2000. Затем каждая из спорных лемм была дополнительно индивидуально разыскана (с помощью корпусного менеджера Bonito). Обычно в языке более частотны положительные формы. Итак, мы предполагали, что частотное возрастание негативных форм служит свидетельством определенного сдвига в значении, о котором пользователь словаря должен быть информирован. Если у имен прилагательных преобладала отрицательная форма (51% и более), то в словаре эта форма приводилась на первом месте, а за косой чертой давалась положительная форма: *nebezpečný/bezpečný.* Это служит пользователю сигналом того, что здесь речь идет об объединенной частотности обеих форм. В корпусе же в качестве леммы приводится часть, находящаяся перед косой чертой, т.е. отрицательная - *nebezpečný*. Здесь мы исходим из того, что пользователь при поиске в корпусе может, исходя из своих потребностей, легко отделить обе формы друг от друга и установить их частотность.

Для установления и указания частотности в корпусе SYN2000 такой метод представлялся наиболее подходящим - однако это совсем не значит, что это имеет силу в случае абсолютно всех лемм. Употребление отрицательных форм может быть обусловлено составом корпуса, в котором преобладают газетные тексты, часто описывающие негативные события. Было бы важно проанализировать употребление этих прилагательных в корпусах разного жанрового состава. И в таком случае было бы необходимо произвести еще семантический анализ корпусных данных, чтобы обогатить словник лемм некоторыми негативными прилагательными.

Literatura:

Čermák, F. – Křen, M.: Frekvenční slovník češtiny. Praha 2004
Čermák, F.: Jazykový korpus: Prostředek a zdroj poznání. SaS 56, 1995.
Čermák, F.: Czech National Corpus: A Case in Many Contexts, International Journal of Corpus Linguistics Vol. 2, 181-197, 1997
Kocek, J. – Kopřivová, M. – Kučera, K.: Český národní korpus. Úvod a příručka uživatele. Praha 2000
Kocek, J. – Kopřivová, M. – Schmiedtová: The Czech National Corpus. In:Proceedings of the 9th EURALEX International Congress, Heid U., Evert S., Lehmann E., Rohrer Ch. (eds.), Stuttgart 2000, s. 127 - 132.
Schmiedtová, V. – Nováková, M :The Czech National Corpus. In: Materialy Tretjej mezhdunarodnoi shkoly-seminara, Ivanovo, Rossija, 14-16 sentyabra 1999 g. Ivanovo 2000, P.20-22. (Proceedings of 3-d International School-Seminar "Dictionary in Contemporary World/)

Korpus
Český národní korpus - SYN2000. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: <http://ucnk.ff.cuni.cz>.