Co to je korpus a co v něm můžeme najít?

(volný překlad dle : ***Tony McEnery & Andrew Wilson:*** **Corpus Linguistics,** Edinburgh Teextbooks in Empirical Linguistics 1996, 1997)

Korpus versus počítačově čitelné texty

Jak jsme už naznačili v minulé hodině, při empirickém výzkumu jazyka badatel používá psaných nebo mluvených textů, které ovšem samy o sobě nemohou být označeny jako korpusy. Zkoumání korpusu se od zkoumání řekněme souboru textů (textu) – literárního díla, excerpt z novin a časopisů, analýz mluvených projevů atd. liší tím, že se nejedná jen o zkoumání náhodně vybraných textů nebo jednotlivého textu. V širším slova smyslu můžeme sice označit každý soubor textů jako korpus – tedy ryze etymologicky na základě fundujícího latinského slova corpus – tělo-těleso – tvořené texty. V moderní lingvistice se ovšem slova korpus používá jakožto termínu, který je přesně definovatelný a zužuje výše zmíněné široké neterminologické pojetí.

Při definování korpusu je potřeba zdůraznit čtyři základní rysy, které sbírka textů musí mít, abychom ji mohli v přísném slova smyslu pokládat za korpus odpovídající požadavkům moderní lingvistiky (KL).

- vzorky (sampling) a reprezentativnost
- konečná velikost (omezený a vymezený rozsah)
- strojově čitelná forma (MRF)
- standardní reference

Vzorky a reprezentativnost korpusu

V lingvistice se klade většinou důraz na jazyk v jeho rozrůzněnosti ve všech varietách, v nichž funguje, než na jazyk jednoho autora. Z toho plyne, že se buď můžeme snažit o to, abychom zachytili všechny vyskytnuvší se výpovědi pronesené nebo zaznamenané v daném jazyce, nebo vykonstruovat malé vzorky, které by pokud možno zachytily všechny vyskytnuvší se varianty. Zatímco první přístup je těžko představitelný (snad jen u mrtvých jazyků nebo přesně vymezitelných podmnožin jazyka, protože počet textů jazyka je prakticky nekonečný), zdá se, že druhý je reálným řešením.

Právě odpůrci KL v čele s N.Ch. namítali proti korpusu to, že NUTNĚ je OMEZENÝ (skewed).

Tuto kritiku je možno překonat pouze vypracováním strategie vzorkování a ověřením v praxi. Jak zařídit, aby vzorky vybrané z nekonečného množství jazykových projevů, reprezentovaly v úplnosti celek jazyka? To je otázka, kterou si lingviste na poli KL kladou a na kterou se snaží s větší či menší mírou určitosti a úspěšnosti odpovídat. Podrobněji se touto problematikou budeme zabývat v příštích přednáškách.

Rozsah korpusu

Druhým úkolem je stanovení rozsahu korpusu. Korpus musí být konečný. Např. může mít 1 000 000 slovních tvarů. Narozdíl od vymezených konečných korpusů existují i tzv. monitorovací korpusy (např. COBUILD Johna Sinclaira), někdy se jim také říká sbírky textů (collections of texts), knihovny, archivy. Důležité jsou především pro lexikografii (nová slova a změna významu již užitých slov).

Jejich nevýhodou je, že vzhledem k tomu, že se stále mění, nejsou spolehlivá pro kvantitativní analýzy.

Rozsah korpusu by měl být hlavní otázkou řešenou při projektu korpusu. Lze vycházet ze zkušeností (rozsahu) existujících korpusů a přihlížet k specifickým cílům projektu.

Elektronicky čitelná podoba korpusu
Dnes je MRF podmínkou. Papírové korpusy jsou věcí historie. Souvisí s rostoucími možnostmi dalšího zpracování (automatického).
Budování – texty v MRF formě
scanování
přepis nahrávek
přepis netištěných dokumentů nebo dokumentů s velmi složitou tištěnou formou.
MRF forma umožňuje rozšíření holého textu o tzv. anotace – obecně vzato veškeré přídatné informace.
Standardizace
Základní informací, již musí každý korpus obsahovat, je odkaz k tomu, jak vznikl, jak byl vytvořen, jak široký vzorek variet jazyka zahrnuje, tedy jak je reprezentativní, rozsáhlý.

V moderní lingvistice je tedy třeba chápat korpus nejen jako jakési tělo složené z textů (soubor nebo sbírku textů), nýbrž jako rozsahem omezený soubor počítačově čitelných textů v podobě vzorků vytvořených se snahou postihnout maximálně reprezentativním způsobem co možná nejvíce jazykových variet na základě jistých vymezitelných a vymezených kriterií.
(definicí je mnoho, srov. lit.)

Kódování a anotace
Korpusu existují v neanotované (surové) nebo anotované (označkované, taggované) podobě. Obsahují holé texty nebo texty s přídatnou především lingvistickou (morfologickou, slovnědruhovou, syntaktickou, sémantickou) informací. Už surové korpusy mohou výrazně pomoci lingvistovi v jeho bádání, obecně lze ovšem tvrdit, že anotované korpusu výrazně rozšiřují možnosti dojít k zajímavějším výsledkům v lingvistické práci a představují tudíž „lepší korpusy“.
Některé informace týkající se gramatiky lze ovšem správně zadaným dotazem vyčíst i z neanotovaného korpusu. Pokud třeba nás bude v anglickém korpusu zajímat člen, můžeme najít všechny slovní tvary (the, a, an) a pouze zkontrolovat, zda se jedná vždy o člen a materiál máme připravený, aniž bychom potřebovali mít slovnědruhově označkovaný korpus. Pokud budeme hledat slovesa ve 3.os.sg., tak ačkoliv mají implicitně výrazný znak (-s), nebude naše situace tak jednoduchá, protože (-s) signalizuje zároveň pl. sb. K vyhledávání odpovědí na složitější dotazy budeme tudíž potřebovat označkovaný korpus, v němž každé jednotce (slovu, větě,...) bude přiřazena anotace (tag – visačka) obsahující nějakou přídatnou lingvistickou informaci (POS, gram. význ., synt. popis,...) O jednotlivých typech lingvistických anotací se zmíním dále. Nyní bychom se měli podívat na 7 zásad, které formuloval přední korpusový lingvista G. Leech (1993), podle nichž by se měl anotátor řídit.

- zachovat vratnost anotovaného korpusu do surového stavu (autor značek je interpretem, s nímž nemusí každý potenciální uživatel souhlasit, přičemž by mělo být technicky možné se případné nežádoucí interpretace zbavit a moci pracovat bez ní)
- možnost extrahovat anotace z textu a uložit je zvlášť, aby bylo možné se k nim vrátit (formou nějaké relační databáze, nebo interlineárního formátu)
- Anotační schéma by mělo vycházet z teoretických východisek, která by měla být jasně formulovaná a přístupná každému konečnému uživateli korpusu. Mnohé korpusy byly anotovány ručně (existence subjektivních interpretací zaviněných osobou anotátora ve sporných případech). Značkování by pak mělo být doplněno komentáři, z nichž by byl důvod příslušné volby patrný.

- Mělo by být jasné JAK a KDO anotaci provedl (JAK – ručně x automaticky x poloautomaticky, s postkorekcí x bez korekce) (KDO – počítačový program, anotátor - člověk)
- Uživatel korpusu by si měl být vědom toho, že anotace nejsou nějakou nedotknutelnou neomylnou instancí. Anotace je pouze více či méně užitečným nástrojem. INTERPRETACE.
- Anotační schéma by mělo být založeno na široce schvalovaných a teoreticky nezatížených principech. Není na škodu i zjednodušující přístup.
- Žádné anotační schéma nemá právo být pokládáno za standardní. Je-li nějaké řešení uznávanější, děje se tak pouze z praktických důvodů.

Jedním z problémů ovšem zůstává užitečnost anotace pro konečného uživatele a snadnost anotace pro anotátora korpusu. Jev, který zajímá konečného uživatele může představovat značný problém pro automatickou analýzu, značné úsilí pro ručního anotátora. Při vlastní práci s korpusy se mi ovšem naskytla jistá zkušenost, s níž bych se s vámi ráda podělila. Pokud lze problém algoritmizovat, je mnohdy možné i bez anotací dostat příslušnou odpověď, pokud problém algoritmicky řešit nelze, je ruční analýza, kterou by musel provést anotátor dobře dostupná i konečnému uživateli. Zůstává pak otázkou, kolik konečných uživatelů daný problém zajímá a je-li jich tolik, aby se vyplatilo platit angažovat ručního anotátora.

Formát anotací

Neexistuje žádny široce přijímaný standard representace anotací. Existují jednotlivé pokusy, které se porůznu ujímají. Poměrně rozšířené jsou COCOA-references. COCOA je program pro extrakci extralingvistických informací z textu. Jeho konvence byly přeneseny do programu OCP. Použit byl pro Longman-Lancaster Corpus budovaný v Helsinkách. Je založen na použití <> uhlových závorek, které uzavírají informace typu atribut – hodnota, např. A- atribut označující autora textu má hodnotu konkrétní jméno autora. <A CHARLES_DICKENS>.

Iniciativu při budování standardního způsobu anotací převzala iniciativa TEI (Text Encoding Initiative). Jedná se o aktivitu sponzorovanou hlavními vědecky orientovanými asociacemi zabývajícími se využitím komputerů v humanitních vědách. ACL (Association for Computional Linguistics), ALLC (the Association for Literary and Linguistic Computing), ACH (the Association for Computers and Humanities). Cílem TEI je vytvoření standardní implementace pro operace s počítačově čitelnými texty. TEI za tímto účelem používá již existující formu SGML (Standard Generalised Markup Language). Byl přijat proto, že je jednoduchý, jasný, formálně přísný a již mezinárodně uznávaný. Vlastním příspěvkem TEI je detailní návod k použití přísl. standardu.

V TEI obsahuje každý dokument povinně dvě části header (hlavičku) a vlastní text. Header obsahuje informace o textu jako takovém (autor, titul, datum vzniku,...) informace o původním formátu z něhož byl dekódováním text vytvořen (edice, MRF) a informaci o tom, jak se text dekódoval (kódoval). Současná praxe TEI je založena na tom, že header obsahuje tagy a entity references. Text je složen z elementů. Elementem může být libovolná jednotka textu (slovo, věta, odstavec, kapitola,...) Elementy jsou otaggovaný značkami SGML (NN1 – podstatné jméno, apelativum, nominativ). Začátek elementu je označen uhlovými závorkami <...> , konec rovněž </...>Např. začátek odstavce a konec odstavce . Reference jsou naopak ohraničeny na začátku znakem „&” a na konci znakem „;”. Reference jsou těsnopisnými detailnějšími informacemi. Používá se těsnopisných zkratek FSD (Feature System declaration) např. vvd – plnovýznamové sloveso v participiální formě –ed.

„polished&vvd;”

v Longman-Lancaster Oslo-Bergen Corpus vypadala značka podobně

“polished_vvd“

Text jako celek se v TEI popisuje pomocí DTD (Document type description). Jedná se o formální reprezentaci, která informuje uživatele nebo počítačový program o tom, které elementy text obsahuje , jak jsou tyto elementy kombinovány, obsahuje také sadu deklarací entit, např. reprezentaci nestandardních znaků. TEI už v DTD definovalo standardní typy jako báseň, dopis, drama atd. Např. tag pro drama má k dispozici značky pro označení různých typů dramatického textu, seznam výstupů atd. DTD se používá programem známým pod názvem SGML parser, který kontroluje, zda je text otaggován ve formátu kompatibilním s TEI.
Celá řada korpusů přijala TEI za své.
TEI vydává mnoho návodů pro kódování korpusových textů. EU založilo dozorčí skupinu EAGLES (Expert Advisory Groups on Language Egineering Standards), která má za úkol sledovat a pomáhat různým evropským iniciativám. Jde o to vytvořit systém značek, který by na jedné straně zachytil všechny zvláštnosti, potřeby, specifika všech evropských jazyků a na straně druhé zachoval jednotu systému.

Typy anotací

- Vnitrotextové a extratextové informace

Jaké typy anotací vlastně v korpusu mohou být uváděny?
První informací je, na jaký text se vlastně díváme, který korpus máme před sebou

- pravopis

TEI – WSD (writing system declaration) – snaha řešit problémy abeced jednotlivých evropských jazyků, přihlíží se i k nelatinkovým abecedám.

- lingvistické anotace

Gramatické anotace morfologického typu se v korpusové literatuře nejčastěji nazývají tagy. Hovoří se o gramatickém tagování spíše než o gramatických anotacích.
POS – slovnědruhové značkování (gramatické, morfologické, morfosyntaktické značkování)
LEMMATIZACE
Parsing
Sémantika
Anotace diskursu a textově lingvistické anotace
Fonetická transkripce
Prozódie

Vícejazyčné korpusy

Kromě jednojazyčných korpusů, o nichž jsme se v naší přednášce primárně zmiňovali, existují i vícejazyčné korpusy (Aarhus Corpus of Danish, French and English contract Law).
Někdy se hovoří o paralelních korpusech (srovnání různých biblických překladů).
V souvislosti s paralelními korpusy je důležité řešení otázky tzv. alingement (zarovnání) – vzájemné přiřazení jednotek, které si odpovídají (vět).

Závěr

V dnešní přednášce jsme si ukázali, co je to vlastně korpus z hlediska moderní lingvistiky a jaké typy dodatečných informací v něm lze nalézt.
Důležité je zapamatovat si 4 požadavky kladené na korpus a 7 zásad pro dodatkové informace.

Dále doporučuji stránky: http://www.athel.com/corpus.html