Kapitola 20

Státnice I3: Jazykové korpusy a lingvistická anotace

20.1 Zdroje dat

- strojově čitelné soubory dat
  - noviny, sborníky, romány, technická dokumentace, dialogy, internet, ...
  - závisí na úkolu, který chci řešit
  - čím větší, tím lepší

Kódování znaků

- současné počítače jsou číslicové — všechno (program, data) je reprezentováno jako číslo
- pro práci s textem je potřeba zavést konvenci pro transformaci číslo ↔ znak abecedy

Základní pojmy:

- **znak** (character)
  - abstraktní pojem
  - nemá sám o sobě žádnou číselnou reprezentaci ani pevnou grafickou podobu — např. `velké písmeno A s čárkou`
- **repertoár znaků** (character repertoire)
  - množina znaků
  - otázka identity: stejně vypadající znaky mohou být považovány za logicky odlišné (A v latince a alfabetě)
- **kódování** (encoding)
  - algoritmus pro převod posloupnosti znaků na posloupnost oktetů
- **kódová pozice znaku** (code position)
  - číselná reprezentace znaku (nezáporné celé číslo)
- **kódování** (character code)
  - 1-1 relace mezi prvky repertoáru znaků a nezápornými celými čísly
- **glyf**
  - vizuální prezentace znaku
- **font**
  - repertoár glyfů pro množinu znaků

ASCII

- American Standard Code for Information Interchange (od 1950's)
- sedm bitů – hodnoty 0-127
- 0-31,127 — kontrolní znaky (Escape, Line Feed)
- 32-126 — mezera, speciální znaky, číslice, velká a malá písmena latinky:

```
! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z [ \ ] ^ _ ` a b c d e f g h i j k l m n o p q r s t u v w x y z { | } ~
```

- výhody:
  - velice jednoduché kódování: jeden znak – jedna kódová pozice
  - minimální objem: 1 znak – 1 oktet (jeden bit zbyde — oktety 128-255 zůstávají nevyužité)
- zásadní nevýhoda:
  - naprosto nedostačující pro repertoáry znaků jiných národních abeced

8-bitová kódování

- potřeba dalších znaků → vznikají nová kódování obsahující ASCII jako podmnožinu a navíc využívají oktety 128-255 (stále platí jeden znak – jeden oktet)
- International Standard Organisation vydává skupinu standardních kódování pro některé skupiny jazků — rodina ISO 8859 (1980's)
- ISO 8859-1 (ISO Latin 1) – západoevropské jazyky
- u češtiny a ostatních středo- a východoevropských jazyků vládne anarchie:
  - ISO 8859-2 (ISO Latin 2)
  - Windows-1250
  - Koi-8
  - Bratři Kameničtí
  - vlastní „standardy" IBM, Apple

Unicode

- jediné řešení: víceoktetová kódování (neplatí ale mýtus, že Unicode je 16-bitové kódování!!!)
- 1991 — Unicode Consortium vytváří normu Unicode (resp. ISO 10646), která určuje repertoár znaků a jejich kódové pozice
- v současnosti 30 světových abeced užívaných v několika stovkách jazyků, cca 40000 znaků — arabština, sanskrt, čínština, japonština, korejština, ... (ambice: 250 abeced pro několik tisíc jazyků)
- znaky v Unicode: „LATIN CAPITAL LETTER A WITH ACUTE"
- různé typy fyzické reprezentace kódů:
  - UTF-8
    - proměnná délka: 1-6 oktetů na znak (použitých max. 4)
    - v prvním oktetu posloupnosti určuje počet bitů zleva po první nulu celkový počet oktetů
    - další byty zápisu stejného znaku vždy začínají `10xxxxxx`, jiný byte je začátek zápisu dalšího znaku
    - výhoda: znaky ASCII se v UTF-8 kódují stejně → kompatibilita
    - výhoda pro češtinu: všechny znaky české abecedy se kódují jedním nebo dvěma oktety
  - UTF-16 — každý znak v BMP (Basic Multilingual Plane) je reprezentován dvěma oktety; ostatní znaky jsou reprezentovány čtyřmi oktety
  - UTF-32 — každý znak jsou čtyři oktety
- problémy Unicode v porovnání s 8-bitovými kódováními:
  - řetězcová ekvivalence vizuálně totožných znaků různých abeced: např. A v latince, azbuce a alfabetě
  - abecední řazení: např. „LATIN CAPITAL LETTER A WITH ACUTE" vs. „LATIN CAPITAL LETTER A WITH GRAVE"

Jiná řešení

- transliterace
- escape notation — znaky mimo kódování, které je k dispozici, mohou být nahrazeny dohodnutou posloupností znaků (různé pro různé systémy)

Ä: `\"{A}` (TeX) `&Auml;` (HTML)

20.2 Anotace

- přidávání vybrané lingvistické informace k již existujícímu korpusu
- **(semi)automatická** nebo **ruční**
- několik fází zpracování
  1. shromáždění materiálu — skenování + OCR, WWW jako korpus
  2. konverze + čištění — jednotný formát a kódování
  3. klasifikace dokumentů
  4. segmentace dokumentu — hranice vět, hranice slov (tokenizace; problém co je slovo — číselné výrazy, jazyky bez mezer)
- textová reprezentace: jednoduchá implementace, uložené přímo v nízkoúrovňovém datovém formátu
- grafická reprezentace: intuitivnější (minimálně pro lingvisty)
  - potřeba speciálních nástrojů pro anotaci, správu dat, dávkové zpracování příkazů
- tým lidí, kteří provádí anotaci
  - vedoucí (rozděluje úkoly, najímá lidi)
  - editor (člověk) pro tvorbu anotačních pravidel
  - lingvista pro řešení lingvistických jevů
  - anotátoři — nejpočetnější skupina
  - technická podpora/programátor
  - kontrola dat

20.3 Datové formáty

Rozmanitost datových formátů

- rozmanitost datových zdrojů vedla k rozmanitosti datových formátů
- postupně ale dochází ke konvergenci k XML

Ukázky různých formátů

- Negra Treebank – tabulka (odkazy na syntaktické rodiče)

```
%% word                 tag     morph           edge parent
#BOS 1 1 985275570 1
Mögen                  VMFIN   3.Pl.Pres.Konj  HD      508
Puristen               NN      Masc.Nom.Pl.*   NK      505
aller                  PIDAT   *.Gen.Pl        NK      500
Musikbereiche          NN      Masc.Gen.Pl.*   NK      500
```

- Penn Treebank – závorková notace stromů

```
( (S (NP-SBJ The proposed changes)
    (ADVP also)
    (VP would
        (VP allow
            (S (NP-SBJ executives)
               (VP to
```

```
                    (VP report
                        (NP (NP exercises)
                            (PP of
                                (NP options)))
                        (ADVP-TMP (ADVP later)
                                  and
                                  (ADVP less often)))))))
.))
```

- WordNet – databáze s odkazy na ID příbuzných pojmů (ukázka: index a datový řádek)

```
entrance v 2 3 @ ~ + 2 0 01806505 00020926

00044113 04 n 05 entrance 0 entering 0 entry 0 ingress 0 incoming 0 012 @ 00043484 n 0000 + 01958650 v 030
+ 01958650 v 0201 + 01671620 v 0201 + 01958650 v 0101 ~ 00044454 n 0000 ~ 00044639 n 0000 ~ 00044745 n 000
~ 00044898 n 0000 ~ 00045146 n 0000 ~ 00046615 n 0000 ~ 01178182 n 0000 | the act of entering; "she made a
```

- Valenční slovník VALLEX

```
* HLÁSIT
 ~ ned: hlásit
 + ACT(1;obl) ADDR(3;opt) PAT(o+6;opt) EFF(4,jak,že;obl) LOC(;typ)
    -synon: oznámit
    -example:  hlásit někomu o něčem zprávu
    -reciprex:  hlásili si navzájem o sob? nové zprávy
    -reciprocity: ACT-ADDR-PAT
    -use: prim
    -class: communication
    -freq: 8;9
 ~ ned: hlásit
 + ACT(1;obl) ADDR(3;opt) PAT(na+4;opt) EFF(4,jak,?e;obl) LOC(;typ)
    -synon: udat
    -example:  hlásit na něj, že přechovává zakázané knihy
    -reciprocity: ACT-ADDR-PAT
    -use: posun
    -class: communication
    -freq: 3
```

PDT

- Pražský závislostní korpus (Prague Dependency Treebank)
- původně vlastní formáty
  - **CSTS** (Czech Sentence Tree Structure) – hlavní formát PDT 1.0, založený na SGML
    - původně pouze morfologická rovina, pozd. i analytická rovina
    - struktura zachycena pomocí odkazů

```
< s id="ln95048:053-p6s48">
  <f cap>Černý<l>černý-1_^(barva)
    <t>AAMS1----1A----<A>Sb<r>1<g>3
  <f>se<l>se_^(zvr._zájmeno/částice)
    <t>P7-X4----------<A>AuxT<r>2<g>3
  <f>vzdal<l>vzdát
    <t>VpYS---XR-AA---<A>Pred<r>3<g>0
  <d>.<l>.
    <t>Z:-------------<A>AuxK<r>4<g>0
```

- **FS** (Feature Structures)
  - pro analytickou a morfologickou rovinu, dvojice atribut – hodnota
  - zachycení struktury pomocí závorek — nutná linearizace stromu
  - problémy při přesouvání uzlu z konce věty na začátek další věty (při opravě chyb v segmentaci) — nutno přepočítat celou strukturu

```
[#10,AuxS,#10,SENT,0,0](
  [vzdal,Pred,vzdát_se,PRED,3,3,0](
    [Černý,Sb,černý,ACT,1,1,3],
    [se,AuxT,se,2,2,3,hide]
  ),[.,AuxK,&Period;,4,4,0,hide]
)
```

- později přechod na formát **PML**
  - reprezentace pomocí XML
  - obecný jazyk poskytující nástroje k popisu omezení a vlastností svých aplikací pomocí **PML schématu** (popis toho, co se může v instancích vyskytovat, jaká je struktura, role)
  - PML instance
    - hlavička: odkaz na externí schéma nebo schéma samo
    - další odkazy na externí soubory
    - ostatní prvky podle příslušného schématu

```
<?xml version="1.0"?>
  <annotation xmlns="http://ufal.mff.cuni.cz/pdt/pml/">
  <head>
    <schema href="example1_schema.xml"/>
  </head>
  <meta>
    <annotator>Jan Novak</annotator>
    <datetime>Sun May 1 18:56:55 2005</datetime>
  </meta>
  <trees>
    <LM ord="2">
      <func>Pred</func>
      <form>loves</form>
      <governs>
        <LM ord="1">
          <func>Subj</func>
          <form>John</form>
        </LM>
        <LM ord="3">
          <func>Obj</func>
          <form>Mary</form>
        </LM>
      </governs>
    ...
```

XML

- GML (Generalized Markup Language), SGML (Standard GML), HTML, XML
- **well-formed** (odpovídá XML) vs **valid** (odpovídá druhu konkrétního dokumentu) dokument – validace pomocí DTD definic
- jmenné prostory, XPath, XSL + XSLT
- API pro XML: **SAX**, **DOM**

20.4 Typologie korpusů

- **korpus** — strukturovaný, unifikovaný a (často též označkovaný) rozsáhlý soubor jazykových dat
- **korpusová lingvistika**
  - zabývá se studiem jazyka a obsahuje všechny procesy týkající se zpracování, používání a analýzy psaných a mluvených (strojově čitelných) korpusů
  - relativně moderní pojem, který označuje přístup založený na příkladech použití jazyka v "reálném světě"
- kritéria klasifikace korpusu

  - **médium** — tištěný, elektronický text, digitalizovaná řeč, video
  - **metoda tvrby** — vyvážený (je vůbec možné vytvořit vyvážený korpus?) vs speciální (oborový)
  - **jazykové proměnné**
    * jednojazyčný vs. mnohojazyčný
    * původní text vs. překlady z cizího jazyka
    * rodilý mluvčí vs. nerodilý mluvčí
  - **jazykový vývoj** — synchronní vs. diachronní
  - prostý vs. anotovaný

Korpusy

Jednojazyčné korpusy

- Brown Corpus — 1 MW (1964)
  - děrné štítky, publikovaná statistika, později otagován; výběr amerických textů z r. 1961
- British National Corpus — 100 MW (1994)
  - 100 MW soubor psaného (90 procent) a mluveného (10 procent) jazyka obsahující současnou britskou angličtinu, morfologická anotace
- Deutsches Referenzkorpus/Cosmas IDS-Mannheim (2004)
  - >4 GW současné psané němčiny (bez anotace);
  - automatická lematizace, databanka kookurencí – předpřipravený seznam výskytových vzorů pro cca 220000 lemmat
- Český národní korpus
  - diachronní část 13-19.století — DIAKORP
  - Synchronní část — cca od 1900
  - psaný jazyk – 100MW v SYN 2000/2005 ("vyvážený"), 700MW SYN2009PUB (pouze publicistika)
  - mluvený jazyk – Pražský mluvený korpus (PMK), Brněnský mluvený korpus (BMK)
  - dialekty

Paralelní korpusy

- text a jeho překladové ekvivalenty v jednom nebo několika jazycích (hub language)
- přidaná hodnota — párování (alignment)
- InterCorp (ÚČNK)
  - čeština + cca 20 evropských jazyků, některé z nich s morfologickou anotací
- MULTEXT-EAST
  - Multilingual Text Tools and Corpora for Central and Eastern European Languages
  - překlad "1984" od George Orwella, kolem 100kW
  - bulharština, čeština, estonština, maďarština, rumunština, slovinština a angličtina (hub language) (a nedávno také chorvatština, litevština, rumunština, ruština)
  - manuálně ověřené zarovnání vět
- Europarl
  - texty extrahované ze sborníků Evropského parlamentu
  - 11 paralelních jazyků: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish
  - až 50 MW na jazyk
- JRC-Acquis
  - Dokumenty evropská komise, 23 jazyků

  - Automatický alignment, často dost hrubý

- PCEDT
  - český překlad 21,600 anglických vět z části obsahující texty z Wall Street Journal z Penn Treebank 3 korpusu
  - automatická morfologická anotace a parsování na analytickou a tektogramatickou úroveň

- CzEng
  - česko-anglický automaticky anotovaný korpus (8 MS, cca 80 MW)
  - cca polovina textu jsou filmové titulky (3 MS); zbytek pak legislativa EU (1,5 MS), technická dokumentace (1 MS), romány (1 MS), paralelní webové stránky (0,5 MS), novinové články

Treebanky

- **treebank**
  - databáze syntaktických stromů
  - korpus obsahující informaci o morfologické a syntaktické (a někdy i sémantické) struktuře

Penn Treebank

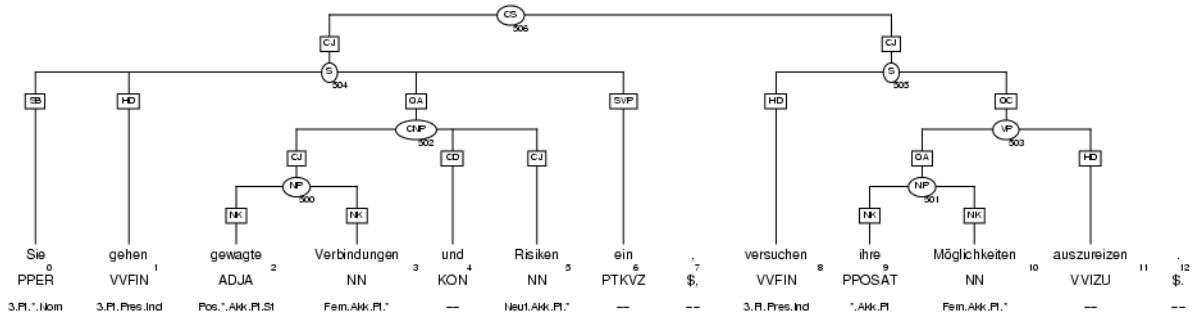
- obsahuje 1,5 MW anglických novinových článků
- syntax bezprostředních složek

PropBank / NomBank

- přidání sémantické vrstvy jako nadstavby nad Penn Treebank – propozice (predikát a argumenty) pro slovesa, resp. substantiva

```
Ukázka anotace: Mr.Bush met him privately, in the White House, on Thursday.
Rel: met
Arg0: Mr.Bush
Arg1: him
ArgM-MNR: privately
ArgM-LOC: in the White House
ArgM-TMP: on Thursday
```

NEGRA



Obrázek 20.1: Příklad stromu z NEGRA Treebank

- 350 kW novinových textů v němčině, syntaktická anotace podobná PTB

Tiger

- 700 kW novinových textů v němčině
- větší velikost než NEGRA treebank, jiná lematizace a morfologie, sekundární hrany (koordinace)

BulTreeBank

- bulharština (200 kW) anotovaná pomocí HPSG formalismu — složkové stromy

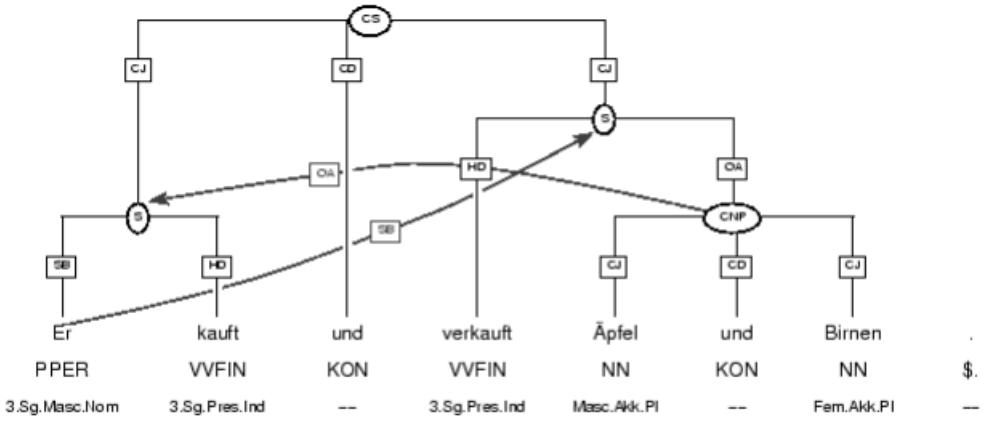


Abbildung 2: Annotation von Koordination durch sekundäre Kanten

Obrázek 20.2: Příklad stromu z Tiger Treebank se sekundárními hranami

- 4 typy prvků: lexikální (N, V, Prep, ...), frázové (VPAdjunct, NPComplement,...) , funkcionální (Conj, CongArg,...) , textové (vlastní řetězce)

Szeged Treebank

- syntaktické struktury vyvinuté pro maďarštinu
- 1,2 MW, ruční morfologická disambiguace a syntaktická anotace

Penn Chinese Treebank
Pražský závislostní korpus

- neboli Prague Dependency Treebank (PZK, PDT)
- obsahuje velké množství českých textů doplněných rozsáhlou a provázanou morfologickou (2 MW), syntaktickou (1,5 MW) a sémantickou (0,8 MW) anotací
- na sémantické rovině jsou navíc anotovány aktuální členění věty a koreferenční vztahy

20.5 Počítačová lexikografie

- vytváření strojově čitelných slovníků za účelem jejich využití v různých úlohách NLP
- ruční tvorba nebo automatická extrakce
- problémy při tvorbě slovníků:
  - neúplnost slovníku — opomenutí nebo přehlédnutí některého hesla
  - jaký zvolit přístup k vlastním jménům — vynechat nebo zahrnout?
  - ghost words – neexistující slova, ve slovníku omylem (př. Dord = density ('D or d'))
  - rozlišení významu hesel (P.Hanks: A serious problem for computer applications is that dictionaries compiled for human users focus on giving lists of meanings for each entry, without saying much about how one meaning may be distinguished from another in text.)
- rozlišování hesel
  - **homografie** — vlastnost dvou nebo více termínů, které mají stejnou grafickou formu, ale rozdílnou výslovnost — např. pro-udit vs. proud-it
  - **polysémie** — jednomu lexému odpovídají dva nebo více významů (v souvislosti s jejich výskytem v různých kontextech), např. vyšel z místnosti vs. vyšel z předpokladu
  - **homonymie** — šetřit (= spořit) vs. šetřit (= zjišťovat)
  - pozor — homonymie a polysémine nejsou to samé!
  - P.Hanks: No generally agreed criteria exist for what counts as a sense, or for how to distinguish one sense from another

Slovníky

- Longman Dictionary of Contemporary English (LDOCE) – three-level embedded structure for sense distinctions (homographs,senses,optional subsenses)
- Roget's Thesaurus (1852)
- Cambridge International Dictionary of English
- COBUILD English Language Dictionary
- WordNet
- VerbNet
  - největší on-line slovník anglických sloves
  - rysy: hierarchický, doménově nezávislý (vyvážený), široké pokrytí
  - propojení do PropBank (90% pokrytí), WordNetu, FrameNetu

FRAMENET ANNOTATION:

[Goods A car] was *bought* [Buyer by Chuck].

[Goods A car] was *sold* [Buyer to Chuck] [Seller by Jerry].

[Buyer Chuck] was *sold* [Goods a car] [Seller by Jerry].

PROPBANK ANNOTATION:

[Arg1 A car] was *bought* [Arg0 by Chuck].

[Arg1 A car] was *sold* [Arg2 to Chuck] [Arg0 by Jerry].

[Arg2 Chuck] was *sold* [Arg1 a car] [Arg0 by Jerry].

Obrázek 20.3: Rozdíly v anotaci mezi ProbBank a FrameNet.

- FrameNet
  - on-line zdroj anglických slov — slovesa, podstatná jména, přídavná jména, předložky
  - cílem je dokumentovat sémantické a syntaktické možnosti kombinace možných významů jednotlivých slov
  - **sémantické rámce**, **sémantické role**
- PDT-VALLEX
  - valenční rámce / významy pro slovesa, podstatná jména a přídavná jména
  - významy, které se vyskytly v textech v PDT (bottom-up přístup — důraz na pokrytí textu)
- VALLEX
  - valenční slovník jako zdroj syntakticko-sémantické informace
  - valenční rámce / významy pro slovesa — komplexní zpracování sloves (všechny významy daného slovesa — top-down přístup)

20.6 Wordnety

Princeton WordNet

- lexikální sémantická síť strukturovaná okolo pojmu synset
- **synset** — soubor literálů se stejným slovním druhem zaměnitelných v určitém kontextu (množina synonym)
- inspirován psycholingvistickou teorií lidské lexikální paměti
- příliš detailní pro běžné NLP úlohy
- vztahy mezi synsety: homonymie, hyperonymie, meronymie, ...

EuroWordNet

- vícejazyčná databáze obsahující několik jednojazyčných wordnetů strukturovaných podobným způsobem jako Princeton WordNet
- angličtina, holandština, němčina, španělština, francouzština, italština, čeština, estonština

Sense tagged corpora

- **sense tagging** — přiřazování významu z nějakého slovníku slovům v textu
  - je potřeba nějaký **sense-enumerative** slovník, který zachycuje všechny možné významy slova
- úloha **WSD** — word sense disambiguation
- interest corpus
  - 2 kS obsahující slovo interest
- SENSEVAL
  - organizace pořádající vyhodnocovací soutěže pro WSD (word sense disambiguation) systémy
- SEMCOR
  - více než 200 kW z anglického Brownova korpusu s označkovaným významem podle Princeton WordNetu 1.6

20.7 Poznámky

- kW, MW, GW — tisíce, miliony, miliardy slov
- kS, MS, GS — tisíce, miliony, miliardy vět
