Today: Week 4, lecture
Today’s topic: Overview of Language Data Resources
Today’s teacher: Zdeněk Žabokrtský

E-mail: zabokrtsky@ufal.mff.cuni.cz
WWW: http://ufal.mff.cuni.cz/zdenek-zabokrtsky
Why language data?

In general, when studying any language phenomenon, there are two basic ways to go:

- thinking about it in the context of one’s language experience, using **introspection**...
- or using **empirical evidence**, statistical models based on real world usage of language ...
  
  ▶ side remark: this includes also using brain-imaging methods or at least eye-tracking devices, but such approaches are still rare in the real NLP industry
Armchair linguistics or data crunching?

- 1957: Noam Chomsky’s attack: “Any natural corpus will be skewed. Some sentences won’t occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.”
- 1988: Frederick Jelinek: ”Every time I fire a linguist, the performance of the speech recognizer goes up” (perhaps not an exact citation)
- but 2004: Frederick Jelinek: “My colleagues and I always hoped that linguistics will eventually allow us to strike gold.”
- 2005: Tony McEnery: “Corpus data are, for many applications, the raw fuel of NLP, and/or the testbed on which an NLP application is evaluated.”
- 200?: Eric Brill: “More data is more important than better algorithms.”
- 200?: Eugene Charniac: “Future is in statistics.”
The world of language data resources today

- Today’s Language data resources map - hopelessly diverse.
- A very very tiny fragment for illustration: only ontologically-oriented data collections, just those adhering to the linked open data principles (credit: Wikipedia)

2016: 1,250 submissions to LREC 2016 (International Conference on Language Resources and Evaluation, biannual)
Why is that so complicated?

Why researchers need so many different pieces of data?

- Is the natural language really so complex? Well, yes.
- In addition,
  - thousands of languages (plus dialects), different writing systems. . .
  - many underlying theories
  - many end-application purposes
Let’s try to systematize the space of data resources

Basic dimensions:

- corpus vs. lexicon
  - lexicon in the broad sense, as a repertory of tokens’ types
- modality: spoken vs. written
  - and other, eg. sign languages
- covered languages: monolingual vs. multilingual
  - if multilingual, then possibly parallel
- time axis: synchronic vs. diachronic
  - if annotated, then what on which “level”, with which underlying theory, what tag set . . .
- time axis: synchronic vs. diachronic
  - if annotated, then what on which “level”, with which underlying theory, what tag set . . .
- plain vs. annotated
  - if annotated, then what on which “level” (which language phenomena are captured), with which underlying theory, with what set of labels (tag set) . . .
- other language variables:
  - original vs. translation
  - native speaker vs. learner
Corpora
Full Definition of CORPUS

plural corpora

1: the body of a human or animal especially when dead

2a: the main part or body of a bodily structure or organ <the corpus of the uterus>
b: the main body or corporeal substance of a thing; specifically: the principal of a fund or estate as distinct from income or interest

3a: all the writings or works of a particular kind or on a particular subject; especially: the complete works of an author
b: a collection or body of knowledge or evidence; especially: a collection of recorded utterances used as a basis for the descriptive analysis of a language
A historical remark

- linguists recognized the need for unbiased empirical evidence long before modern NLP
  - excerption tickets collected systematically for Czech from 1911
Corpus size

- typically measured in tokens (words plus punctuation marks)
- sampling is inescapable
  - an I-want-it-all corpus is far beyond our technology (even in a strictly synchronic sense)
- but still, the corpora sizes have been growing at an exponential pace for some time:
  - Brown Corpus in 1964 $\approx 1$MW
  - (electronic corpus of Czech texts in 1970s: 500kW)
  - British Natural Corpus in 1994 $\approx 100$ MW
  - English Gigaword in 2004 $\approx 1$GW
  - Google’s 5-gram for 10 European Languages in 2009 based on $\approx 1$TW
Balanced corpora

- an elusive goal: a balanced corpus whose proportions correspond to the real language usage
- criteria for choosing types of texts their relative proportion in the corpus (and eventually concrete texts)?
  - style, genre
  - reception vs. perception (a few influential authors vs. production of a large community)?
- actually no convincing generally valid answers for an optimal mixture . . .
- . . . but at least some strategies seem to be more reasonable than others
- an example of a clearly imbalanced corpus: Wall Street Journal Corpus
  - unfortunately used as a material source for the Penn Treebank, which is undoubtedly among the most influential LR
  - “NLP = Wall Street Journal science”
Corpus annotation

- raw texts – difficult to exploit
- solution: gradual “information adding” (more exactly, adding the information in an explicit, machine tractable form)
- annotation = adding selected linguistic information in an explicit form to a corpus
Corpus annotation criticism

- some critics: an annotated corpus is worse than a raw corpus because of forced interpretations
  - one has to struggle with different linguistic traditions of different national schools
  - example: part of speech categories
- relying on annotation might be misleading if the quality is low (errors or inconsistencies)
## Variability of PoS tag sets

### Penn Treebank POS tagset (for English)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordinating conjunction (<em>and</em>)</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number (<em>1, third</em>)</td>
</tr>
<tr>
<td>DT</td>
<td>determiner (<em>the</em>)</td>
</tr>
<tr>
<td>EX</td>
<td>existential there (<em>there is</em>)</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word (<em>d’heure</em>)</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/subordinating conjunction (<em>in, of, like</em>)</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective (<em>green</em>)</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative (<em>greener</em>)</td>
</tr>
<tr>
<td>JJ$S$</td>
<td>adjective, superlative (<em>green est</em>)</td>
</tr>
<tr>
<td>LS</td>
<td>list marker (<em>I</em>)</td>
</tr>
<tr>
<td>MD</td>
<td>modal (<em>could, will</em>)</td>
</tr>
<tr>
<td>NN</td>
<td>noun, singular or mass (<em>table</em>)</td>
</tr>
<tr>
<td>NNS</td>
<td>noun plural (<em>tables</em>)</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular (<em>John</em>)</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural (<em>Vikings</em>)</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer (<em>‘both, is the boy’s</em>)</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending (<em>friend’s</em>)</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun (<em>I, he, it</em>)</td>
</tr>
<tr>
<td>PRP$S$</td>
<td>possessive pronoun (<em>my, his</em>)</td>
</tr>
<tr>
<td>RB</td>
<td>adverb (<em>however, usually, naturally, here, good</em>)</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative (<em>better</em>)</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative (<em>best</em>)</td>
</tr>
<tr>
<td>RP</td>
<td>particle (<em>give up</em>)</td>
</tr>
<tr>
<td>TO</td>
<td>to (<em>to go, to him</em>)</td>
</tr>
<tr>
<td>UH</td>
<td>interjection (<em>uhuhuhuh</em>)</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form (<em>take</em>)</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense (<em>took</em>)</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund/present participle (<em>taking</em>)</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle (<em>taken</em>)</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, sing. present, non-3d (<em>take</em>)</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, 3rd person sing. present (<em>takes</em>)</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-determiner (<em>which</em>)</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun (<em>who, what</em>)</td>
</tr>
<tr>
<td>WP$S$</td>
<td>possessive wh-pronoun (<em>whose</em>)</td>
</tr>
<tr>
<td>WRB</td>
<td>wh-abverb (<em>where, when</em>)</td>
</tr>
</tbody>
</table>
Negra Corpus POS tagset (for German)

- ADJA Attributives Adjektiv
- ADJD Adverbiales oder prälükatives Adjektiv
- ADV Adverb
- APPR Praposition; Zirkumposition links
- APPRART Praposition mit Artikel
- APPRO Postposition
- APZR Zirkumposition rechts
- ART Bestimmter oder unbestimmter Artikel
- CARD Kardinalzahl
- FM Fremdsprachliches Material
- ITJ Interjektion
- KOUI Unterordnende Konjunktion mit zu und Infinitiv
- KOUS Unterordnende Konjunktion mit Satz
- KON Nebenordnende Konjunktion
- KOKOM Vergleichspartikel, ohne Satz
- NN Normaler Nomen
- NE Eigennamen
- PDS Substituierendes Demonstrativpronomen
- PDAT Attribuierendes Demonstrativpronomen
- PIS Substituierendes Indefinitpronomen
- PIAT Attribuierendes Indefinitpronomen
- PIDAT Attribuierendes Indefinitpronomen mit Determiner
- PPER Irreflexives Personalpronomen
- PPPOS Substituierendes Possessivpronomen
- PPOSAT Attribuierendes Possessivpronomen
- PRELS Substituierendes Relativpronomen
- PRELAT Attribuierendes Relativpronomen
- PRF Reflexives Personalpronomen
- PWS Substituierendes Interrogativpronomen
- PWAT Attribuierendes Interrogativpronomen
- PWAV Adverbiales Interrogativ- oder Relativpronomen
- PROAV Pronominaladverb
- PTKZU zu vor Infinitiv
- PTKNEG Negationspartikel
- PTKVZ Abgetrennter Verbzusatz
- PTKANT Antwortpartikel
- PTKA Partikel bei Adjektiv oder Adverb
- TRUNC Kompositions-Einheit
- VVFIN Finites Verb, voll
- VVVIMP Imperativ, voll
- VVIZU Infinitiv mit zu, voll
- VVPP Partizip Perfekt, voll
- VAFIN Finites Verb, aux
- VAIMP Imperativ, aux
- VAINF Infinitiv, aux
- VAPP Partizip Perfekt, aux
- VMFIN Finites Verb, modal
- VMINF Infinitiv, modal
- VMPP Partizip Perfekt, modal
- XY Nichtwort, Sonderzeichen
- Z, Komma
- S, Satzeinende Interpunktion
- $, Sonstige Satzzeichen; satzütern
- VVINF Infinitiv, voll

Zdeněk Žabokrtský (ÚFAL MFF UK) Overview of Language Data Resources Week 4, lecture 15 / 48
Variability of PoS tag sets, cont.

Prague Dependency Treebank morphologitagset (for Czech), several thousand combinations using 15-character long positional tags

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>Morphological tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Některé</td>
<td>některý</td>
<td>PZFP1--------------</td>
</tr>
<tr>
<td>kontury</td>
<td>kontura</td>
<td>NNFP1---A--------</td>
</tr>
<tr>
<td>problému</td>
<td>problém</td>
<td>NNIS2------A-----</td>
</tr>
<tr>
<td>se</td>
<td>se_<code>(zvr._zájemco/částice)</code></td>
<td>P7-X4------------</td>
</tr>
<tr>
<td>však</td>
<td>však</td>
<td>J^---------------</td>
</tr>
<tr>
<td>po</td>
<td>po-1</td>
<td>RR--6------------</td>
</tr>
<tr>
<td>oživeni</td>
<td>oživeni_<code>(3it)</code></td>
<td>NNNS6------A-----</td>
</tr>
<tr>
<td>Havlovým</td>
<td>Havlův_;S_<code>(3el)</code></td>
<td>AUIS7M-----------</td>
</tr>
<tr>
<td>projevem</td>
<td>projev</td>
<td>NNIS7------A-----</td>
</tr>
<tr>
<td>zdaji</td>
<td>zdát</td>
<td>VB-P---3P-AA-----</td>
</tr>
<tr>
<td>být</td>
<td>být</td>
<td>Vf--------A------</td>
</tr>
<tr>
<td>jasnější</td>
<td>jasný</td>
<td>AAFP1---2A------</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>Z:--------------</td>
</tr>
</tbody>
</table>
Treebanks
Treebanks

- a treebank is a corpus in which sentences’ syntax and/or semantics is analyzed using tree-shaped data structures
- a tree in the sense of graph theory (a connected acyclic graph)
- sentence syntactic analysis ... it sounds familiar to most of you, doesn’t it?

Credit: http://konecekh.blog.cz
Honesty: trees are irresistibly attractive data structures.

We believe sentences can be reasonably represented by discrete units and relations among them.

Some relations among sentence components (such as some word groupings) make more sense than others.

In other words, we believe there is an latent but identifiable discrete structure hidden in each sentence.

The structure must allow for various kinds of nestedness (… a já mu řek, že nejsem Řek, abych mu řek, kolik je v Řecku řeckých řek …).

This resembles recursivity. Recursivity reminds us of trees.

Let’s try to find such trees that make sense linguistically and can be supported by empirical evidence.

Let’s hope they’ll be useful in developing NLP applications such as Machine Translation.
So what kind of trees?

There are two types of trees broadly used:
- constituency (phrase-structure) trees
- dependency trees

Constituency trees simply don’t fit to languages with freer word order, such as Czech. Let’s use dependency trees.
How do we know there is a dependency between two words?

- There are various clues manifested, such as
  - word order (juxtaposition): “...přijdu zítra ...”
  - agreement: “...novými_{pl.instr} knihami_{pl.instr}...”
  - government: “...slíbil Petrovi_{dative}...”

- Different languages use different mixtures of morphological strategies to express relations among sentence units.
Basic assumptions about building units

If a sentence is to be represented by a dependency tree, then we need to be able to:

- identify **sentence boundaries**.
- identify **word boundaries** within a sentence.
Basic assumptions about dependencies

If a sentence is to be represented by a dependency tree, then:

- there must be a **unique parent word** for each word in each sentence, except for the root word
- there are **no loops** allowed.
Even the most basic assumptions are violated

- Sometimes **sentence boundaries are unclear** – generally in speech, but e.g. in written Arabic too, and in some situations even in written Czech (e.g. direct speech).
- Sometimes **word boundaries are unclear**, (Chinese, “ins” in German, “abych” in Czech).
- Sometimes its **unclear which words should become parents** (A preposition or a noun? An auxiliary verb or a meaningful verb? ...).
- Sometimes there are too many relations (“Zahlédla ho bosého.”), which implies **loops**.

Life’s hard. Let’s ignore it and insist on trees.
Counter-examples revisited

If we cannot find linguistically justified decisions, then make them at least consistent.

- Sometimes sentence boundaries are unclear (generally in speech, but e.g. in written Arabic too...)
  - OK, so let’s introduce annotation rules for sentence segmentation.
- Sometimes word boundaries are unclear, (Chinese, “ins” in German, “abych” in Czech).
  - OK, so let’s introduce annotation rules for tokenization.
- Sometimes it’s not clear which word should become parent (e.g. a preposition or a noun?).
  - OK, so let’s introduce annotation rules for choosing parent.
- Sometimes there are too many relations (“Zahlédla ho bosého.”), which implies loops.
  - OK, so let’s introduce annotation rules for choosing tree-shaped skeleton.
Treebanking

• Is our dependency approach viable? Can we check it?
• Let’s start by building the trees manually.
• A treebank - a collection of sentences and associated (typically manually annotated) dependency trees
• For English: Penn Treebank [Marcus et al., 1993]
• For Czech: Prague Dependency Treebank [Hajič et al., 2001]
  ▶ Layered annotation scheme: morphology, surface syntax, deep syntax
  ▶ Dependency trees for about 100,000 sentences
• High degree of design freedom and local linguistic tradition bias
• Different treebanks ➞ Different annotation styles
Case study on treebank variability: Coordination

- coordination structures such as "lazy dogs, cats and rats" consists of
  - conjuncts
  - conjunctions
  - shared modifiers
  - punctuations
- 16 different annotation styles identified in 26 treebanks (and many more possible)
- different expressivity, limited convertibility, limited comparability of experiments...
- harmonization of annotation styles badly needed!

<table>
<thead>
<tr>
<th>Main family</th>
<th>Prague family (code pP) [14 treebanks]</th>
<th>Moscow family (code pM) [5 treebanks]</th>
<th>Stanford family (code pB) [6 treebanks]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice of head</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head on left (code hL) [10 treebanks]</td>
<td><img src="lazy_dogs_cats_rats_hL_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_hL_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_hL_treebank.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Head on right (code hR) [14 treebanks]</td>
<td><img src="lazy_dogs_cats_rats_hR_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_hR_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_hR_treebank.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Mixed head (code hM) [1 treebank]</td>
<td><img src="lazy_dogs_cats_rats_hM_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_hM_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_hM_treebank.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Attachment of shared modifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shared modifier below the nearest conjunct (code dN) [15 treebanks]</td>
<td><img src="lazy_dogs_cats_rats_dN_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_dN_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_dN_treebank.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Shared modifier below head (code dH) [11 treebanks]</td>
<td><img src="lazy_dogs_cats_rats_dH_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_dH_treebank.png" alt="Diagram" /></td>
<td><img src="lazy_dogs_cats_rats_dH_treebank.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Attachment of coordinating conjunction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordinating conjunction before previous conjunct (code cP) [2 treebanks]</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Coordinating conjunction before following conjunct (code cF) [1 treebank]</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Coordinating conjunction between two conjuncts (code cB) [8 treebanks]</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Coordinating conjunction as the head (code cH)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Placement of punctuation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>values pP [7 treebanks], pF [1 treebank] and pB [15 treebanks] are analogous to cP, cF and cB (but applicable also to the Prague family)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How many treebanks are there out there?

- growing interest in dependency treebanks in the last decade or two
- existing treebanks for about 50 languages now (but roughly 7,000 languages in the world)

- UFAL participated in several treebank unification efforts:
  - 13 languages in CoNLL in 2006
  - 29 languages in HamleDT in 2011
  - 37 languages in Universal Dependencies in 2015:
Other specialized corpora
Parallel corpora

- specific feature: alignment between corresponding units in two (or more) languages
  - document level alignment
  - sentence level alignment
  - word level alignment
  - (morpheme level alignment?)

- example: The Rosetta Stone
- example: CzEng - a Czech-English parallel corpus, roughly 0.5 words for each language, automatically parsed (using PDT schema) and...
Named entity corpora

- specific feature: instances of proper names, such as names of people, geographical names,
- example: Czech Named Entity Corpus - two-level hierarchy of 46 named entity types, 35k NE instances in 9k sentences

Coreference corpora

- specific feature: capturing relations between expressions that refer to the same entity of the real world

example: Prague Dependency Treebanks (around 40k coreference links in Czech texts)
Sentiment corpora

- specific feature: capture the attitude (in the sense of emotional polarity) of a speaker with respect to some topic/expression.
- simply said: “is this good or is it bad?”
- obviously over-simplified, but highly demanded e.g. by the marketing industry.

(credit: SemEval 2014 documentation)

example: MPQA Corpus
Highly multi-lingual corpora

- specific feature: as many languages as possible
- examples:
  - W2C - at least 1MW for more than 100 languages
  - The Bible Corpus - translations of the Bible into 900 languages
Examples of Lexicon-like Data Resources
Inflectional lexicons

- specific feature: capturing the relation between a lemma and inflected word forms, ideally in both directions
- example: MorfFlex CZ, around 120M word forms associated with 1M lemmas

```
podle-1^_*(3'y-1) Dg-------3N----6 nejnpodlejč
podle-1^_*(3'y-1) Dg-------3N---- nejnpodlejí
podle-1^_*(3'y-1) Dg-------3A----6 nejpodlejč
podle-1^_*(3'y-1) Dg-------3A---- nejpodlejí
podle-1^_*(3'y-1) Dg-------1N---- nepodle
podle-1^_*(3'y-1) Dg-------2N----6 nepodlejč
podle-1^_*(3'y-1) Dg-------2N---- nepodlejí
podle-1^_*(3'y-1) Dg-------1A---- podle
podle-1^_*(3'y-1) Dg-------2A----6 podlejč
podle-1^_*(3'y-1) Dg-------2A---- podlejí
podle-2 RR--2------------ podle
```
Derivational lexicons

- specific feature: capturing the relation between a base word and a derived word (typically by prefixing and/or suffixing)
- example: DeriNet, 1M lemmas, 700k derivation links
Thesaurus

- specific feature: capturing semantic relations between words, such as synonymy and antonymy

- example:

  **Main Entry:** great
  **Part of Speech:** adjective
  **Definition:** excellent, skillful
  **Synonyms:** able, absolute, aces, adept, admirable, adroit, awesome, bad*, best, brutal, cold*, complete, consummate, crack*, downright, dynamite, egregious, exceptional, expert, fab, fantastic, fine, first-class*, first-rate, good, heavy*, hellacious, marvelous, masterly, number one, out of sight, out of this world, out-and-out, perfect, positive, proficient, super-duper, surpassing, terrific, total, tough, transcendent, tremendous, unmitigated, unqualified, utter, wonderful

  **Antonyms:** ignorant, menial, poor, stupid, unskilled, weak

  * = informal/non-formal usage
Wordnets

- specific feature: hyponymy (hyperonymy) forest composed of synsets (sets of synonymous words)
- example: Princeton Wordnet
specific feature: wordnets of several languages interconnected through English as the hub language

**Architecture of the EuroWordNet Data Base**

(credit: intuit.ru)
Valency lexicons

- specific feature: capturing combinatory potential of a word (most frequently of a verb) with other sentence elements
- example: VALLEX - Valency Lexicon of Czech Verbs

**Example 1:**

\begin{verbatim}
1. odvěst; dávat odpověď
\end{verbatim}

\begin{verbatim}
frame: ACT\textsubscript{1} obli, ADDR\textsubscript{3} obli, PAT\textsubscript{opt} aeff\textsubscript{4} typed, MANN\textsubscript{typ}, MEANS\textsubscript{typ}
\end{verbatim}

\begin{verbatim}
example: impf: odpovídal mu na jeho dotaz pravdu / činem / smichem / že ... pf: odpověděl mu na jeho dotaz pravdu / činem / smichem / že ...
\end{verbatim}

**Example 2:**

\begin{verbatim}
2. impf: reagovat pf: reagovat
\end{verbatim}

\begin{verbatim}
frame: ACT\textsubscript{obl}, PAT\textsubscript{opt}, EFF\textsubscript{7}
\end{verbatim}

\begin{verbatim}
example: impf: pokožka odpovídala na chlad zarudnutím, gruzinští milicí neodpovídali střelbou (SYN) pf: vojáci odpověděli střelbou (SYN), na výzvu doby odpověděl změnou vlastního politického chování (SYN)
\end{verbatim}

**Example 3:**

\begin{verbatim}
3. mít odpovědnost
\end{verbatim}

\begin{verbatim}
frame: ACT\textsubscript{1} obli, ADDR\textsubscript{3} opt, PAT\textsubscript{obl} ytf, MEANS\textsubscript{typ}
\end{verbatim}

\begin{verbatim}
example: odpovídá za své děti, odpovídá za ztrátu svým majetkem
\end{verbatim}

**Example 4:**

\begin{verbatim}
4. být ve shodě / v souladu; korespondovat
\end{verbatim}

\begin{verbatim}
frame: ACT\textsubscript{1,3}, PAT\textsubscript{3} obli, REG\textsubscript{typ}
\end{verbatim}

\begin{verbatim}
example: řešení odpovídá svými vlastnostmi požadavkům
\end{verbatim}
... and many other types of language resources
Speech corpora

- specific feature: recordings of authentic speech, typically with manual transcriptions
- for training Automatic Speech Recognition systems
- example: The Switchboard-1 Telephone Speech Corpus, 2,400 telephone conversations, manual transcriptions
Datasets primarily unintended as corpora

- Web as a corpus
- Wikipedia as a corpus
- Enron corpus - 600,000 emails generated by 158 employees of the Enron Corporation
“Metainformation” about languages

- example: The World Atlas of Language Structures (WALS)
  - http://wals.info/
  - specific feature: various language properties (related e.g. to word order, morphology, syntax) captured for hundreds of languages

Feature 33A: Coding of Nominal Plurality

Values
- Plural prefix: 126
- Plural suffix: 513
- Plural stem change: 6
- Plural tone: 4
- Plural complete reduplication: 8
- Mixed morphological plural: 60
- Plural word: 172
- Plural suffix: 61
- No plural: 98

Map of World showing distribution of different coding types.
Final remarks
A final remark: current trends in language resources . . .

trends (in the last few years) according to Nicoletta Calzolari’s LREC 2016 foreword

- social media analysis
- discourse, dialog and interactivity
- treebanks
- under-resourced languages
- semantics
- multi-linguality
- evaluation methodologies
...and the last word

Be careful when you hear (or say) that some language data resource (or an annotation scheme, or a probabilistic model, or a technological standard...) is

- theory neutral, or
  - If fact we cannot “measure” language stuctures per se, and thus we always rely on some assumptions or conventions etc.

- language independent.
  - In fact it is impossible for an NLP developer to consider all variations in morphology/syntax/semantics of all language.