Overview of Language Data Resources

Zdeněk Žabokrtský

October 1, 2024




Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

Why data, and why is it important?

Corpora

Specialized corpora

Lexicon-like Data Resources

And many other types of language resources

Troubles with choosing an annotation scheme: a case study on problematic corpus/treebank design decisions

Final remarks

Why data, and why is it important?

Why data?

- an elegant answer by Tony McEnery (in 2005): "Corpus data are, for many applications, the raw **fuel of NLP**, and/or the **testbed** on which an NLP application is evaluated."
- this is important: you typically need the data not only for developing your application, but also for measuring its quality

Why evaluation?

Why so much focus on performance evaluation right at the beginning?

- Some people from other IT fields think that presenting some weird measures is only about saturating ambitions of academic scholars...
- No, it's not!
- So what's so special about NLP?
- In fact we still don't understand how language works (compared, e.g., to the level of understanding in mechanics, electromagnetic field or anorganic chemistry).
- In NLP, all our solutions are still only approximative nowadays, and often far from perfect.
- Subjective evaluation is slow and costly.

Data as accumulated "ground truth"

- Typically, the data shows us the "ground truth" which our application tries to mimic (e.g., a possible correct translation of a sentence in language A into language B)
- Whenever possible, we use a fully automatized comparison with the ground-truth data as a performance evaluation measure which tells us what works and what doesn't!
- The data gives us the "gradient" during the development of the application, both in the metaphorical sense (for more abstract design decisions) and in the literal sense (if machine learning is used, but this is almost always the case nowadays).

Sometimes, the evaluation role of the data is more crucial.

Basic methodology

In general, when studying a specific language phenomenon or developing an end-user natural language applications, 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

From scepticism about linguistic data to modern 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."
- 200?: Eugene Charniac: "Future is in statistics."
- 200?: Eric Brill: "More data is more important than better algorithms."

The world of language data resources today

- Today's language data resources map - hopelessly diverse.
- more than 1,000 submissions to every LREC (International Conference on *Language Resources and Evaluation*, biannual)
- data centers offering zillions of various language data packages

Why is that so complicated?

Why researchers develop/use 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...
  - sometimes $O(L^2)$ when pairs of languages considered (especially in parallel data for machine translation)
  - many underlying theories
  - many end-application purposes

Let's try to systematize the space of data resources

Classification along basic dimensions:

- corpus vs. lexicon
  - corpus - a collection of authentic utterances in a given language (texts composed of sentences composed of words)
  - lexicon in the broad sense, as a repertory of tokens' types
- modality: spoken vs. written
  - and other, e.g. sign languages
- number of contained languages: monolingual vs. multilingual
  - a single language only vs. several languages
  - if multilingual, then possibly parallel

Classification along basic dimensions, cont.

- time axis: synchronous vs. diachronic
  - only the contemporary language (such as a few last decades) vs. a collection of texts along longer time space (such as several centuries)
- raw 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
  - various kinds of language disorders ...

Corpora

CORPUS according to Merriam-Webster

Full Definition of CORPUS

plural **corpora** \-p(ə-)rə\

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

2 **a :** 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

3 **a :** 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, numerals, puntuation marks ...)
- sampling is inescapable
  - an I-want-it-all corpus is far beyond our technology (even in a strictly synchronous sense)
- but still, the corpora sizes have been growing at an exponential pace for some time:
  - Brown Corpus in 1964 $\approx$ 1MW
  - (electronic corpus of Czech texts in 1970s: 500kW)
  - British Natural Corpus in 1994 $\approx$ 100 MW
  - English Gigaword in 2004 $\approx$ 1 GW
  - HPLT 1.2, English, 3 TW
- so very roughly more than an order of magnitude per decade
- but no exponential growth lasts forever - what's the upper limit?
- probably at least 1 TW for "big" languages: web-based data - Google's 5-gram dataset for 10 European Languages in 2009 based on $\approx$ 1TW

Balanced corpora

- a discussion that used to be heated some decades ago
- an (IMHO 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
- the most famous example of clearly imbalanced corpora: the 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"

Balanced corpora, cont.

- 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
  - an occasional criticism of NLP: "NLP = a Wall Street Journal science"
- a lesson taken
  - no broadly accepted criteria for balancing a corpus
  - thus we can hardly reach a perfectly balanced corpus
  - but we should at least avoid building a strikingly imbalanced corpus

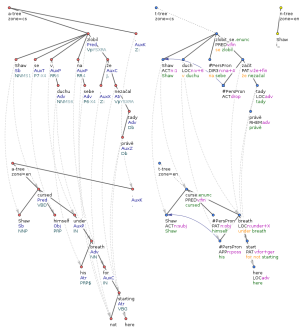
Corpus annotation

- 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
- some examples of possible annotations:
  - morphological annotation (assign a lemma, part of speech and other morphological categories to all tokens)
  - syntactic annotation – phrase-structure or dependency syntactic trees
  - semantic annotation – dozens of theoretical approaches, no consensus so far
  - anaphora – e.g. what pronouns refer to
  - word sense disambiguated corpora – which sense is used for a given polysemous word in a given context
  - sentiment corpora – positive/negative emotions induced by some expressions in a text

Specialized corpora

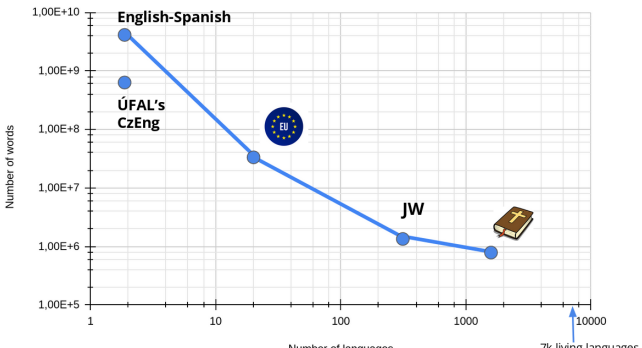
Parallel corpora

- 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 MW for each language,



automatically parsed (using PDT schema) and aligned

Sizes of the biggest multi-parallel corpora



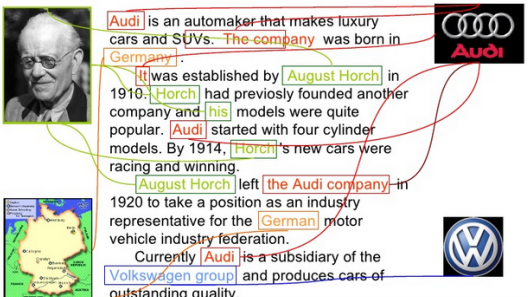
Named entity corpora

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

Dnes sehrají fotbalisté Slavie na Strahově od 17.30
hodin utkání Interpoháru s Bayerem Leverkusen , v
jehož barvách by se měl představit i bývalý olomoucký
útočník Pavel Hapal .
Cítím , že můj osud je zpečetěn .
Křesťanství pohanům .
ČINNOST POBOČKY EVROPSKÉ BANKY PRO OBNOVU A
ROZVOJ (BERD) v Praze slavnostním přestřižením stuhy
včera zahájil president BERD Jacques Attali .

Coreference corpora

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

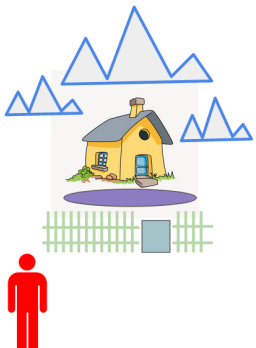


(credit: Shumin Wu and Nicolas Nicolov)

- example: Prague Dependency Treebanks (around 40k coreference links in Czech texts)
- breaking news: coreference datasets for 11 languages converted to a common unified format and published a week ago by UFAL!

Coreference, cont.

Во время своих прогулок в окрестностях Симеиза я обратил внимание на одинокую дачу , стоявшую на крутом склоне горы . К этой даче не было проведено даже дороги . Кругом она была обнесена высоким забором , с единственной низкой калиткой , которая всегда была плотно прикрыта . И ни куста зелени , ни дерева не виднелось над забором . Кругом дачи - голые уступы желтоватых скал ; меж ними кое - где росли чахлые можжевельники и низкорослые , кривые горные сосны . " Что за фантазия пришла кому - то в голову поселиться на этом диком , голом утесе ? Да и живет ли там кто - нибудь ? " - думал я , бродя вокруг дачи . Я еще никогда не видел , чтобы кто - нибудь выходил оттуда . Любопытство мое было так велико , что я , признаюсь , пытался заглянуть на двор таинственного жилища , взобравшись на вышележащие скалы . Но дача была так расположена , что , откуда бы я ни заходил , я мог видеть только небольшой угол двора . Он был так же пуст и невозделан , как и окружающая местность .



- **Coreference resolution** = the task of connecting expressions that refer to the same real-world entity

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

Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
"March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*," according to Chen, also Taiwan's chief WTO negotiator.
friday evening plans were great, but saturday's plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: "I'd like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

(credit: SemEval 2014 documentation)

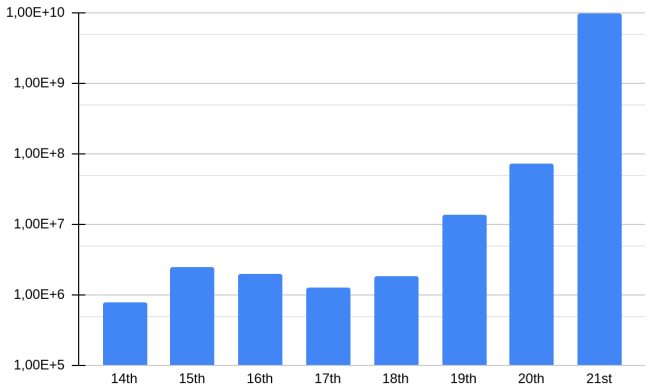
- example: MPQA Corpus

Highly multi-lingual corpora

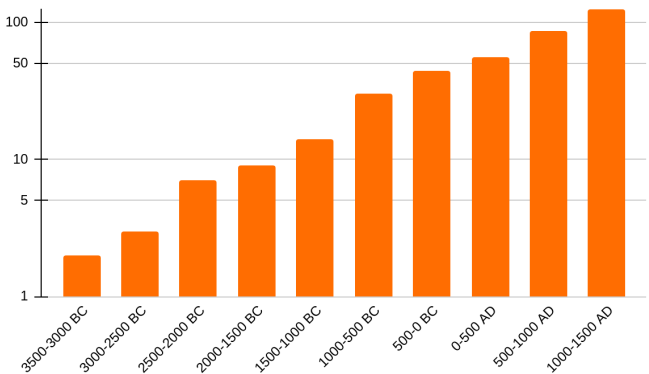
- specific feature: as many languages as possible
- examples:
  - W2C - at least 1MW for more than 100 languages
  - The Bible Corpus - translations of the Bible into 900 languages
  - The OPUS Corpus - the open parallel corpus – 60 languages
    `https://opus.nlpl.eu/`

Size of now-available historical Czech texts

- the DIAKON corpus, total sizes of texts collected for individual centuries

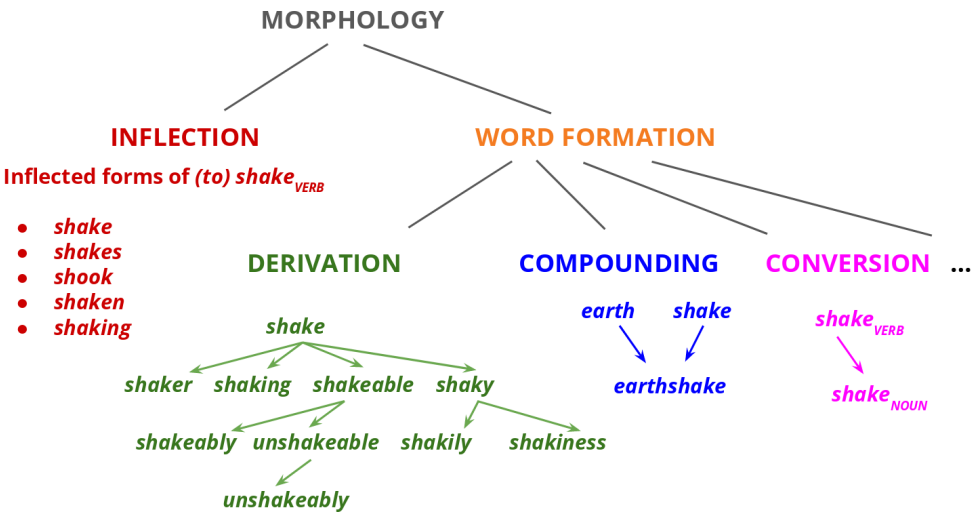


A larger context: the number of languages with written records along time



Lexicon-like Data Resources

Morphology under a minute



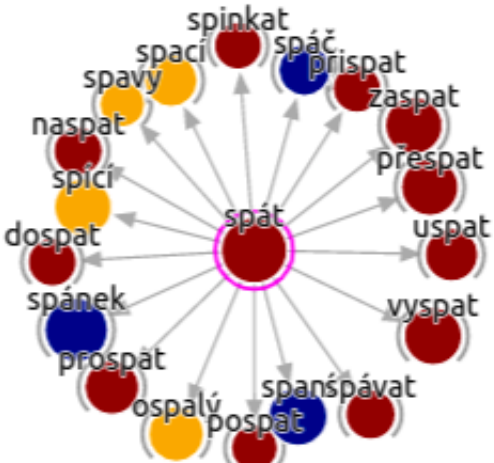
Inflectional lexicons

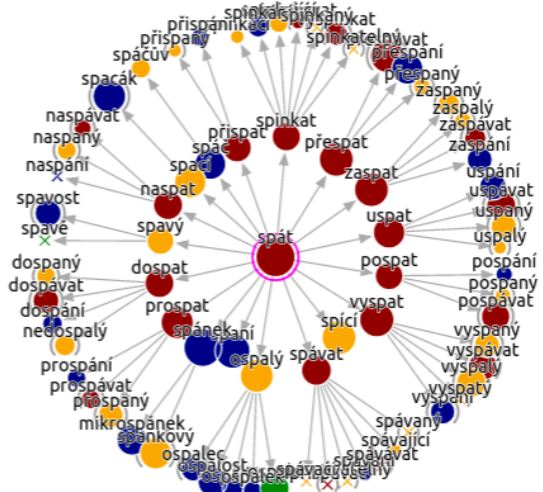
- specific feature: capturing the relation between a lemma and inflected word forms, ideally in both directions
- example: MorfFlex CZ, around 120M word forms associated with 1M lemmas

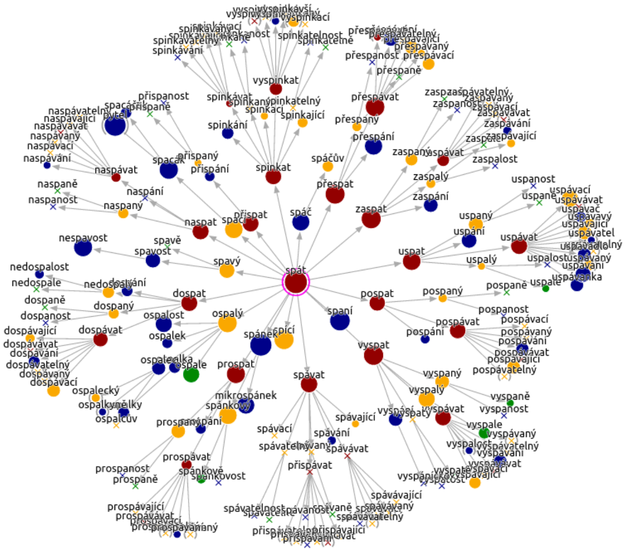
```
podle-1_^(*3ý-1) Dg-------3N---6 nejnepodlejc
podle-1_^(*3ý-1) Dg-------3N---- nejnepodleji
podle-1_^(*3ý-1) Dg-------3A---6 nejpodlejc
podle-1_^(*3ý-1) Dg-------3A---- nejpodleji
podle-1_^(*3ý-1) Dg-------1N---- nepodle
podle-1_^(*3ý-1) Dg-------2N---6 nepodlejc
podle-1_^(*3ý-1) Dg-------2N---- nepodleji
podle-1_^(*3ý-1) Dg-------1A---- podle
podle-1_^(*3ý-1) Dg-------2A---6 podlejc
podle-1_^(*3ý-1) Dg-------2A---- podleji
podle-2 RR--2---------- podle
```

Derivational lexicons

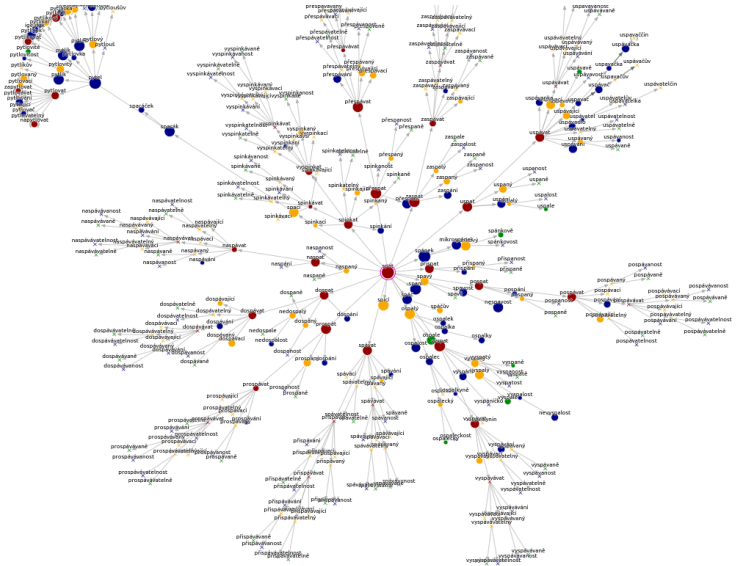
- specific feature: capturing the relation between a base word and a derived word (typically by prefixing and/or suffixing)
- example: DeriNet, 1M lemmas, 700k derivation links


spinkat
spáč
spací
přispat
spavý
zaspat
naspat
přespat
spící
spát
dospat
uspat
spánek
vyspat
prospat
spanšpávat
ospalý
pospat


spinkat
přispat
přespat
spát
spací
spánek
spalý
spávat
spící
zaspat
uspat
vyspat
dospat
prospat
dospalý
dospávat
dospání
nedospalý
prospání
prospávat
prospalý
mikrospánek
spánkový
ospalec
ospalost
naspávat
naspalý
naspání
spavost
spavé
spavý
naspat
spacák
spáček
přispalý
přispání
přispávat
přispalý
spáčův
přespání
přespávat
přespalý
zaspalý
zaspávat
zaspání
uspání
uspávat
uspalý
uspalý
pospat
pospání
pospalý
pospávat
vyspalý
vyspávat
vyspalý
vyspatý
vyspání
spávaný
spávající
spávavat
spávání



spát
spaní
spánek
spící
spavý
spavost
spavě
nespavost
spáč
spáčův
spací
přispat
naspat
dospat
ospalý
ospalost
ospalek
ospale
prospat
spávat
vyspat
pospat
uspat
zaspat
přespat
spinkat
spinkání
vyspinkat
spinkávat
spinkávání
spinkatelnost
spinkatelně
spinkací
spinkající
vyspinkací
přespávat
přespání
přespaný
přespanost
přespaně
přespávací
přespávaný
naspávat
naspaný
naspání
naspaně
naspanost
naspávání
naspávací
naspávaný
naspávající
naspávavat
spacák
přispaný
přispání
přispanost
nedospalost
nedospale
dospaný
dospaně
dospanost
dospávat
dospávající
dospávání
dospávatelný
dospávaný
dospávací
ospalecký
ospalcův
mikrospánek
spánkový
prospávat
prospanost
prospaně
prospávající
prospávavat
spánkově
spánkovost
přispávat
spávající
spávací
spávání
spávávat
spávávající
spávatelnost
vyspávat
vyspaný
vyspalý
vyspaně
vyspanost
vyspale
vyspávaný
vyspávatelný
vyspávání
pospávat
pospaný
pospaně
pospanost
pospání
pospávací
pospávaný
pospávání
pospávavat
pospávatelný
uspaný
uspání
uspalý
uspávat
uspale
uspanost
uspaně
uspávací
uspávavat
uspávač
uspávající
uspávatel
uspávanka
zaspaný
zaspávat
zaspání
zaspalý
zaspalost
zaspávání
zaspávající



Thesaurus

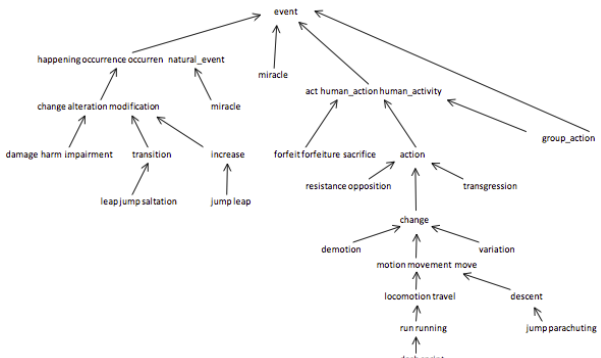
- specific feature: capturing semantic relations between words, such as synonymy and antonymy
- example:

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

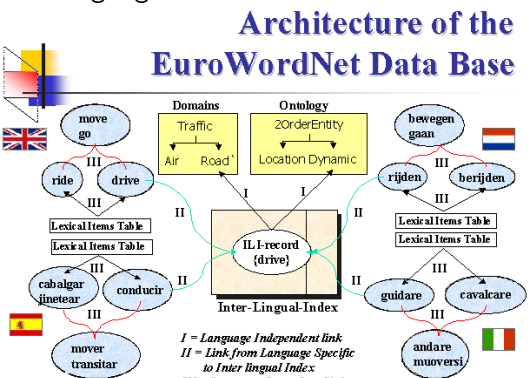
Wordnets

- specific feature: hyponymy (hyperonymy) forest composed of synsets (sets of synonymous words)
- example: Princeton Wordnet
  `http://wordnetweb.princeton.edu/perl/webwn`



EuroWordNet

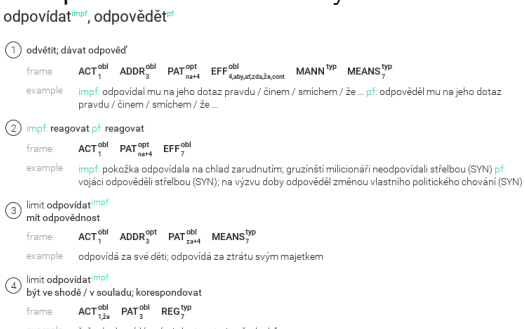
- specific feature: wordnets of several languages interconnected through English as the hub language



(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



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 uninteded as corpora

- Web as a corpus
- Wikipedia as a corpus (e.g. for named entity linking by "wikification")
- 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

Troubles with choosing an annotation scheme: a case study on problematic corpus/treebank design decisions

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)

CC coordinating conjunction (*and*)
CD cardinal number (*1, third*)
DT determiner (*the*)
EX existential there (*there is*)
FW foreign word (*d'hoevre*)
IN preposition/subordinating conjunction (*in, of, like*)
JJ adjective (*green*)
JJR adjective, comparative (*greener*)
JJS adjective, superlative (*greenest*)
LS list marker (*1)*)
MD modal (*could, will*)
NN noun, singular or mass (*table*)
NNS noun plural (*tables*)
NNP proper noun, singular (*John*)
NNPS proper noun, plural (*Vikings*)
PDT predeterminer (*jiż both/iż the boys*)
POS possessive ending (*friend's*)
PRP personal pronoun (*I, he, it*)

PRP$ possessive pronoun (*my, his*)
RB adverb (*however, usually, naturally, here, good*)
RBR adverb, comparative (*better*)
RBS adverb, superlative (*best*)
RP particle (*give up*)
TO to (*to go, to him*)
UH interjection (*uhhuhhuhh*)
VB verb, base form (*take*)
VBD verb, past tense (*took*)
VBG verb, gerund/present participle (*taking*)
VBN verb, past participle (*taken*)
VBP verb, sing. present, non-3d (*take*)
VBZ verb, 3rd person sing. present (*takes*)
WDT wh-determiner (*which*)
WP wh-pronoun (*who, what*)
WP$ possessive wh-pronoun (*whose*)
WRB wh-adverb (*where, when*)

Variability of PoS tag sets, cont.

Negra Corpus POS tagset (for German)

ADJA Attributives Adjektiv
ADJD Adverbiales oder prdikatives Adjektiv
ADV Adverb
APPR Prposition; Zirkumposition links
APPRART Prposition mit Artikel
APPO 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 Normales Nomen
NE Eigennamen
PDS Substituierendes Demonstrativpronomen
PDAT Attribuierendes Demonstrativpronomen
PIS Substituierendes Indefinitpronomen
PIAT Attribuierendes Indefinitpronomen
PIDAT Attribuierendes Indefinitpronomen mit Determiner
PPER Irreflexives Personalpronomen
PPOSS 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-Erstglied
VVFIN Finites Verb, voll
VVIMP Imperativ, voll
VVINF Infinitiv, 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
$, Komma
$. Satzbeendende Interpunktion
$(Sonstige Satzzeichen; satzintern
NNE Verbindung aus Eigennamen und normalen Nomen

Variability of PoS tag sets, cont.

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

Form	Lemma	Morphological tag
Některé	*některý*	PZFP1----------
kontury	*kontura*	NNFP1-----A----
problému	*problém*	NNIS2-----A----
se	*se_^(zvr._zájmeno/částice)*	P7-X4----------
však	*však*	J^-------------
po	*po-1*	RR--6----------
oživení	*oživení_^(*3it)*	NNNS6-----A----
Havlovým	*Havlův_;S_^(*3el)*	AUIS7M---------
projevem	*projev*	NNIS7-----A----
zdají	*zdát*	VB-P---3P-AA---
být	*být*	Vf--------A----
jasnější	*jasný*	AAFP1----2A----
.	.	Z:-------------

Variability - conclusion

- to some entent, differences in annotation schemes can be attributed to real differences among languages
- however, much more diversity is caused simply by different design decisions of the annotation schemes' authors

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

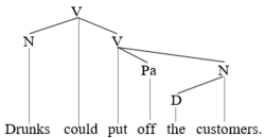
Why trees: Initial thoughts

1. Honestly: trees are irresistibly attractive data structures.
2. We believe sentences can be reasonably represented by discrete units and relations among them.
3. Some relations among sentence components (such as some word groupings) make more sense than others.
4. In other words, we believe there is an latent but identifiable discrete structure hidden in each sentence.
5. 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 ...*).
6. This resembles recursivity. Recursivity reminds us of trees.
7. Let's try to find such trees that make sense linguistically and can be supported by empirical evidence.
8. Let's hope they'll be useful in developing NLP applications such as Machine Translation.

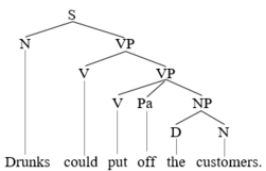
So what kind of trees?

There are two types of trees broadly used:

- constituency (phrase-structure) trees
- dependency trees



– Dependency grammar analysis of constituent structure



– Phrase structure grammar analysis of constituent structure

Credit: Wikipedia

Constituency trees simply don't fit to languages with freer word order, such as Czech. Let's use dependency trees.

BTW how do we know there is a dependency between two words?

- There are various clues manifested, such as
  - word order (juxtapositon): "*...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.

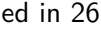
- Sometimes sentence boundaries are unclear (generally in speech, but e.g. in written Arabic too…)
  - **OK, so let's introduce annotation rules for sentence segmentation.**
- Sometimes word boundaries are unclear, (Chinese, "ins" in German, "abych" in Czech).
  - **OK, so let's introduce annotation 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: morhology, surface syntax, deep syntax
  - dependency trees for about 100,000 sentences
- high degree of design freedom and local linguistic tradition bias
- different treebanks $\Longrightarrow$ different annotation styles

An example of a treebank variability cause: the case of coordination

- coordination structures such as "*lazy dogs, cats and rats*" consists of
  - conjuncts
  - conjunctions
  - shared modifiers
  - punctuation tokens
- 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!**

Main family	Prague family (code fP) [14 treebanks]	Moscow family (code fM) [5 treebanks]	Stanford family (code fS) [6 treebanks]
Choice of head			
Head on left (code hL) [10 treebanks]			
Head on right (code hR) [14 treebanks]			
Mixed head (code hM) [1 treebank]	A mixture of hL and hR		
Attachment of shared modifiers			
Shared modifier below the nearest conjunct (code sN) [15 treebanks]			
Shared modifier below head (code sH) [11 treebanks]			
Attachment of coordinating conjunction			
Coordinating conjunction below previous conjunct (code cP) [2 treebanks]	—		
Coordinating conjunction below following conjunct (code cF) [1 treebank]	—		
Coordinating conjunction between two conjuncts (code cB) [8 treebanks]	—		
Coordinating conjunction as the head (code cH) is the only applicable style for the Prague family [14 treebanks]		—	—
Placement of punctuation			
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)			

Btw 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:
  - 70+ languages in UD in 2019
  - 100+ languages in UD in 2024

Conclusion

- one should keep in mind that there's no straightforward "God's truth" when it comes to language data resources
- all resources are heavily influenced by numerous design choices, for which no perfect answers exists
- examples of trade-offs:
  - the bigger data the better, but you can't remove all noise from really big data
  - parallel annotation reduces the amount of annotation errors, but increases costs
  - linguistically-based annotation brings interpretability, but at the same time we risk being trapped in some suboptimal traditions that are possibly not useful beyond a given language family
  - a better quality/coverage is sometimes achievable by integrating more resources focused on a same task, but their licenses might be incompatible

Final remarks

“Let’s consider any language, for example English...”
“Our approach is theory neutral and language independent...”

- Be very careful when you hear that some language data resource (or an annotation scheme, or a probabilistic model, or a technological standard...) is
  - theory neutral,
    - If fact we cannot “measure” language structures *per se*, and thus we always rely on some assumptions/abstractions/conventions etc.
  - or language independent.
    - In fact it is impossible for a human (a linguist or an NLP developer) to take into account all the diversity in morphology/syntax/semantics in all languages.
    - Keep in mind: even the seemingly harmless assumption in such “language independent” approaches that “a sentence is a sequence of words” doesn’t make sense in some languages.

Current trends in language resources ...

trends in the last decade

- sure, as big data as possible, as everywhere else, but apart from that also:
- multi-linguality
- under-resourced languages
- social media analysis
- discourse, dialog and interactivity
- treebanking
- evaluation methodologies

Academic data centers and commercial vendors of language data resources

Before you start compiling your own dataset, have a glimpse into huge existing catalogues such as

- LRE Map https://lremap.elra.info/
- Linguistic Data Consortium https://www.ldc.upenn.edu/
- LINDAT/CLARIN Repository https://lindat.mff.cuni.cz/repository/

Shared tasks in NLP

- as in any other human activity, competing with others is an important source of motivation of NLP researchers
- a typical scenario
  1. an NLP task is selected (such as machine translation for a pair of languages) and defined in detail, including a performance measure specification
  2. training data are published so that participants can start developing their solutions and optimize them with respect to the data
  3. final evaluation dataset is published only for very limited time (such as previously unseen texts to be translated)
  4. participants submit their systems' outputs for the final dataset, their solutions are evaluated and compared

Shared tasks in NLP, cont.

Examples:

- Workshop in Machine Translation's shared task - a yearly competition in Machine Translation
- the CoNLL series, selected tasks:
  - 2019 – Cross-Framework Meaning Representation Parsing
  - 2018 – Universal Morphological Reinflection
  - 2017 – Multilingual Parsing from Raw Text to Universal Dependencies
  - 2014 – Grammatical Error Correction
  - 2006 – Multi-Lingual Dependency Parsing
- the SemEval series, selected tasks:
  - 2018 – Affect and Creative Language in Tweets
  - 2016 – Textual Similarity and Question Answering
  - 2015 – Word Sense Disambiguation and Induction