Lexical datasets across languages

Zdeněk Žabokrtský

November 27, 2024

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

- my **previous lecture**
  - about datasets with features capturing **characteristics of a language as a whole**

- **this lecture**
  - about lexically oriented datasets, with features that capture **characteristics of individual words**

Outline

A terminological note: “lexical”

- a lexical data resource - a collection of information related to the vocabulary (lexicon) of a language... simply about words
- Etymology of ”lexical”: From Latin lexis, from Ancient Greek (léxis, “word”) + -al, from (légō, “to speak”), ultimately from Proto-Indo-European *leǵ- (“to gather, collect”).
- Related terms: lexeme, lexis, lexicon, lexicology, lexicography (the difference between the last two?)

A terminological note: the multi-/cross- distinction

a working definition, not necessarily accepted by all:

- a **multilingual** data resource – contains information about multiple languages, but often does not explicitly facilitate interactions or tasks across those languages
  in brief, multilingual $\approx N\times$ monolingual
- a **crosslingual** data resource – specifically designed to support interactions or tasks that involve bridging or connecting between different languages

A terminological note: the multi-/cross- distinction, cont.

- obviously a fuzzy boundary: e.g. annotation categories used in a harmonized scheme used in a multilingual data resource can already be considered as a point of cross-lingual connection (similarly informal translation glosses)
- rather a gradual scale (as always in linguistics) between multilingual and crosslingual resources

Multilingual lexical data resources

UniMorph

- goal: represent world languages' morphology in a common scheme
- a collaborative effort, started around 2016 at JHU
- originally specialized at inflection, but derivation and morphological segmentation later added too
- 169 languages, with highly different data sizes
- https://unimorph.github.io/
- simple tsv files used for storage, separate files for
  - inflection
  - derivation (only some languages)
  - morpheme segmentation (only some languages)

UniMorph – examples from inflectional files (eng and deu)

```
eat     eats        V;PRS;3;SG
eat     eating      V;V.PTCP;PRS
eat     ate         V;PST
eat     eaten       V;V.PTCP;PST
eat     eats        N;PL
rastrography  rastrographies N;PL
magnetencephalography magnetencephalographies N;PL

Bild  Bild      N;NOM;NEUT;SG
Bild  Bilder    N;NOM;NEUT;PL
Bild  Bildes    N;GEN;NEUT;SG
Bild  Bilder    N;GEN;NEUT;PL
Bild  Bilde     N;DAT;NEUT;SG
Bild  Bildern   N;DAT;NEUT;PL
Müllmann Müllmänner N;NOM;MASC;PL
```

UniMorph – examples from derivational files (eng and deu)

abandon	abandoned	N:ADJ	-ed
abandoned	abandonedly	ADJ:ADV	-ly
abandon	abandonee	N:N	-ee
abandon	abandoner	V:N	-er
abandon	abandonment	N:N	-ment

bestellen	Bestellung	V:N	-ung
Vietnam	Vietnamese	N:N	-ese
lang	langsam	ADJ:ADJ	-sam
Wissen	Wissenschaft	N:N	-schaft
England	englisch	N:ADJ	-isch
Engel	englisch	N:ADJ	-isch
rauben	Räuber	V:N	-er

UniMorph – examples from segmentation files (eng and deu)

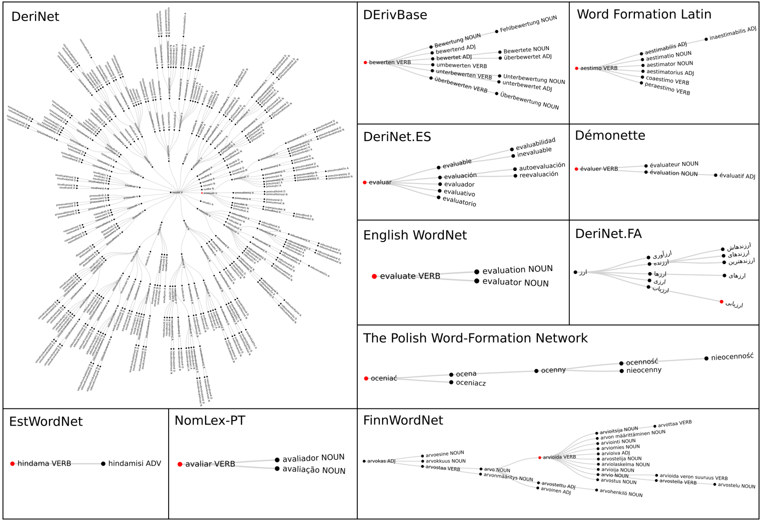
```
hammerhead     hammerheads     N|PL          hammerhead|s
sluicegate     sluicegates     N|PL          sluicegate|s
paraffinize    paraffinizes    V|PRS;3;SG    paraffinize|s
paraffinize    paraffinizing   V|V.PTCP;PRS  paraffinize|ing
paraffinize    paraffinized    V|PST         paraffinize|ed
kiloampere     kiloamperes     N|PL          kiloampere|s
fictionalizer  fictionalizers  N|PL          fictionalizer|s

abzählen       abzählend   V;V.PTCP;PRS     ab-|zähl|-end
abzählen       abgezählt   V;V.PTCP;PST     ab-|ge-|zähl|-t
Zahnarzt       Zahnärzte   N;MASC|NOM;PL    Zahnarzt|e
Zahnarzt       Zahnarztes  N;MASC|GEN;SG    Zahnarzt|es
Zahnarzt       Zahnarzts   N;MASC|GEN;SG    Zahnarzt|s
Zahnarzt       Zahnärzte   N;MASC|GEN;PL    Zahnarzt|e
Zahnarzt       Zahnarzte   N;MASC|DAT;SG    Zahnarzt|e
Zahnarzt       Zahnärzten  N;MASC|DAT;PL    Zahnarzt|en
```

Time for a demo

Universal Derivations

- created at ÚFAL
- UDer 1.1 available at Lindat
  https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3247
- or downloaded in the ÚFAL file system `/net/data/universal-derivations/`
- 21 languages in v. 1.1

Universal Derivations – a sample of derivational trees



Universal Derivations – a sample from the Spanish part

```
33059.0 acentuar#         acentuar
33059.1 acentuable#       acentuable       33059.0 Type=Derivation
33059.2 acentuación#      acentuación      33059.0 Type=Derivation
33059.3 desacentuación#   desacentuación   33059.2 Type=Derivation
33059.4 inacentuación#    inacentuación    33059.2 Type=Derivation
33059.5 preacentuación#   preacentuación   33059.2 Type=Derivation
33059.6 acentuado#        acentuado        33059.0 Type=Derivation
33059.7 acentuadamente#   acentuadamente   33059.6 Type=Derivation
33059.8 inacentuado#      inacentuado      33059.6 Type=Derivation
33059.9 acentuador#       acentuador       33059.0 Type=Derivation
33059.10 acentuamiento#   acentuamiento    33059.0 Type=Derivation
```

UniSegments

- created at ÚFAL
- UniSegments 1.0 https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4629
- or in the ÚFAL file system `/net/data/universal-segmentations/`
- publicly distributable and UFAL-internal parts distinguished because of license limitations of the original resources
- public version: 38 data resources for 30 languages harmonized into the same scheme

UniSegments – samples (eng, deu)

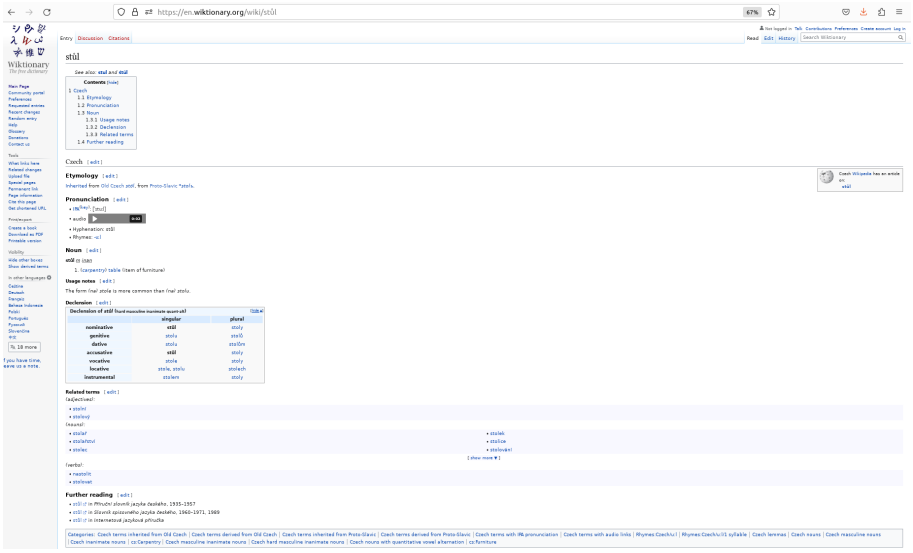
amplification	amplification	NOUN	ampl + ifi + cation
amplified	amplified	VERB	ampl + ifi + ed
amplifier	amplifier	NOUN	ampl + ifi + er
amplifiers	amplifiers	NOUN	ampl + ifi + er + s
amplifies	amplifies	VERB	ampl + ifi + es
amplify	amplify	VERB	ampl + ify
amplifying	amplifying	VERB	ampl + ify + ing
amplitude	amplitude	NOUN	ampl + itude

entschlussfähig	entschlussfähig	ADJ	ent + schluss + fähig
entschlusslos	entschlusslos	ADJ	ent + schluss + los
entschlüpfen	entschlüpfen	VERB	ent + schlüpf + en
entschlüsseln	entschlüsseln	VERB	ent + schlüss + el + n
entschrotten	entschrotten	VERB	ent + schrott + en
entschuldbar	entschuldbar	ADJ	ent + schuld + bar
entschulden	entschulden	VERB	ent + schuld + en

Wiktionary

- (wiktionary = blend of ‘wiki’ + ‘dictionary’)
- https://www.wiktionary.org/
- a collaborative wiki-based (browser-editable) free-content multilingual dictionary
- November 2023: 192 languages, 37 M articles, 6k creators
- an entry = a wikipage about a word

Wiktionary – an example of an entry: Czech noun ‘stůl’ (a table)



stůl

See also: stul and stül

Contents [hide]

1 Czech
- 1.1 Etymology
- 1.2 Pronunciation
- 1.3 Noun
  - 1.3.1 Usage notes
  - 1.3.2 Declension
  - 1.3.3 Related terms
- 1.4 Further reading

Czech [edit]

Etymology [edit]

Inherited from Old Czech stol, from Proto-Slavic *stolъ.

Pronunciation [edit]

- (file): [stuːl]
- audio 0:02
- Hyphenation: stůl
- Rhymes: -uːl

Noun [edit]

stůl m (inan)

1. (carpentry) table (item of furniture)

Usage notes [edit]

The form /nal stole is more common than /nal stolu.

Declension [edit]

Declension of stůl (hard masculine inanimate accent-a) [hide]

	singular	plural
nominative	stůl	stoly
genitive	stolu	stolů
dative	stolu	stolům
accusative	stůl	stoly
vocative	stole	stoly
locative	stole, stolu	stolech
instrumental	stolem	stoly

Related terms [edit]

(adjectival):

- stolní
- stolový

(nouns):

- stolař
- stolařství
- stolec
- stolek
- stolice
- stolování

[show more ▼]

(verbs):

- nastolit
- stolovat

Further reading [edit]

- stůl in Příruční slovník jazyka českého, 1935–1957
- stůl in Slovník spisovného jazyka českého, 1960–1971, 1989
- stůl in Internetová jazyková příručka

Categories: Czech terms inherited from Old Czech | Czech terms derived from Old Czech | Czech terms inherited from Proto-Slavic | Czech terms derived from Proto-Slavic | Czech terms with IPA pronunciation | Czech terms with audio links | Rhymes:Czech/uːl | Rhymes:Czech/uːl/1 syllable | Czech lemmas | Czech nouns | Czech masculine nouns | Czech inanimate nouns | cs:Carpentry | Czech masculine inanimate nouns | Czech hard masculine inanimate nouns | Czech nouns with quantitative vowel alternation | cs:Furniture

Crosslingual lexical resources

Keep in mind:

First, "crosslingual" implies pairs (or tuples) of languages, which implies $\mathcal{O}(L^2)$, with the number of languages or language varieties $L \approx 10^3$ to 10^4

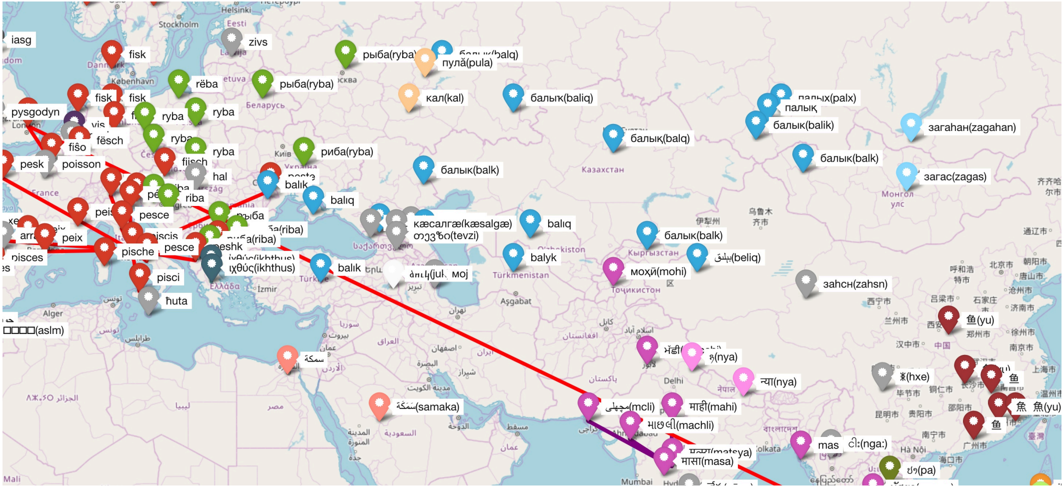
Second, translation equivalents are hardly ever 1:1, hence the size of the space of translation equivalents is $\mathcal{O}(W^2)$, with the number of dictionary words per language $W \approx 10^5$ (or even 10^7 if we consider inflected forms)

The question of an appropriate representation is not only a big $\mathcal{O}$ problem. It is a BIG problem indeed.

CogNet

- cognates
  - words that have the same origin, typically similar forms and similar meanings at the same time
  - slightly more formally: sets of words inherited from an etymological ancestor in a common parent language
  - in theory, distinguished from loanwords that have been borrowed "horizontally"
- an example: night (English), nui (French), noche (Spanish), Nacht (German) ...
- CogNet – a cognate database for 338 languages
  - 8.1 M cognates
  - clustered in 91k concepts (based on Princeton WordNet concepts)
  - 38 writing systems
- data available at https://github.com/kbatsuren/CogNet
- or downloaded in the ÚFAL Linux filesystem `/net/data/CogNet/`

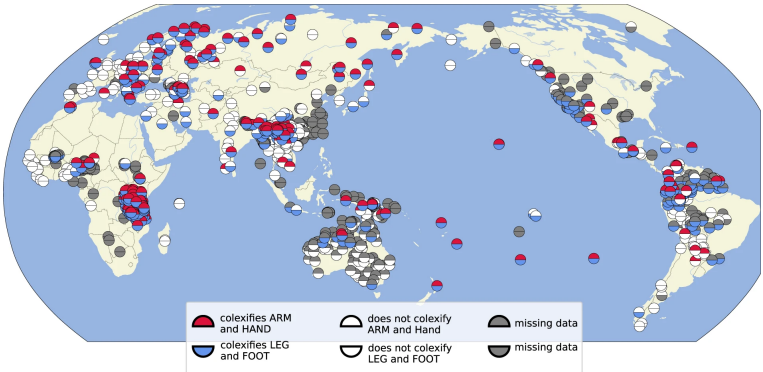
CogNet


iasg
fisk
zivs
рыба(ryba)
пулă(pula)
rëba
рыба(ryba)
кал(kal)
балык(baliq)
палык
балык(balik)
заrahан(zagahan)
pysgodyn
fisk
fisk
ryba
ryba
fësch
fiŝo
ryba
ryba
балык(balq)
pesk
poisson
риба(ryba)
hal
балык(balk)
балык(balk)
загас(zagas)
riba
balık
baliq
pesce
peix
кæсалгæ(kæsalgæ)
თევზი(tevzi)
baliq
балык(balk)
پیلق(beliq)
pische
pesce
ιχθύς(ikhthus)
balık
балык
моҳӣ(mohi)
pisces
pisci
zahsn(zahsn)
ħuta
鱼(yu)
(aslm)
سمكة
ह(nya)
ञा(nya)
སྙ(nya)
(hxe)
鱼
سَمَكَة(samaka)
مچھلی(mcli)
माही(mahi)
મ્છલી(machli)
魚 魚(yu)
鱼
मासा(masa)
mas
ငါး(nga:)
ປາ(pa)
Stockholm
Helsinki
København
Беларусь
Казахстан
Türkmenistan
Aşgabat
Delhi
Mumbai
Kolkata
Alger
Mali
Monгол улс

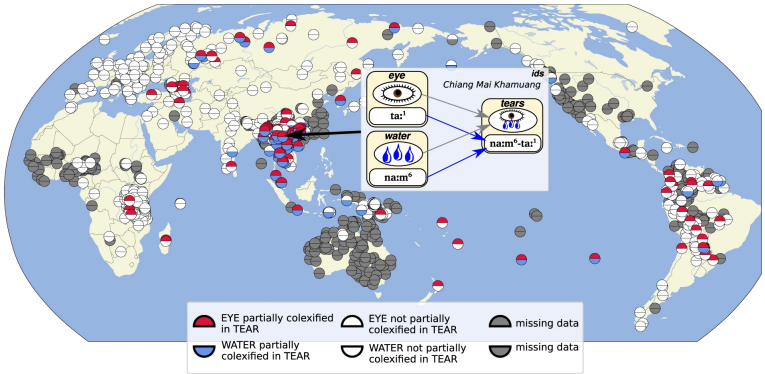
Lexibank

- a collection of standardized wordlists
- List, JM., Forkel, R., Greenhill, S.J. et al. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. Sci Data 9, 316 (2022). DOI: 10.1038/s41597-022-01432-0
- data available at https://zenodo.org/records/7836668
- or downloaded in the ÚFAL file system `/net/data/lexibank/lexibank/`
- 4,000+ wordlists for 2,400+ language varieties
- standardization efforts on already existing lexical datasets
- colexification - different meanings expressed by the same word form (co-lexification)
  - hand vs. arm in Czech ('ruka')
  - people vs. village in Spanish ('pueblo')
- partial colexification – two word forms expressing two different concepts are not identical, but share a common substring

Example of colexification in Lexibank



Example of partial colexification in Lexibank



WOLD – The World Loanword Database

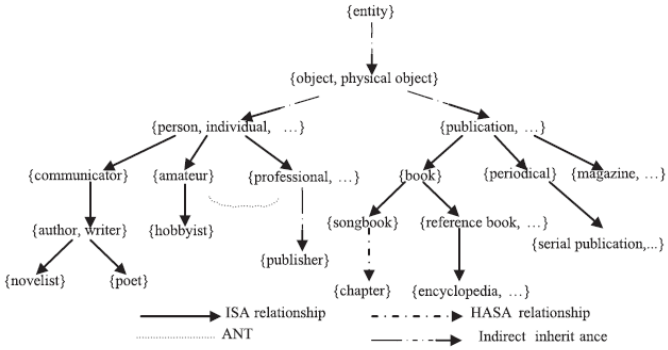
- search interface and data download at https://wold.clld.org/
- downloaded in the ÚFAL file system `/net/data/WOLD/`
- mini-dictionaries of about 1000-2000 entries
- 41 languages
- information on the loanword status of each word (source language and source word given for loanwords)

PanLex

- https://panlex.org/
- a collection of thousands of translation dictionaries
- 5,700 languages
- in total 25 M words, 1.3 G translation pairs

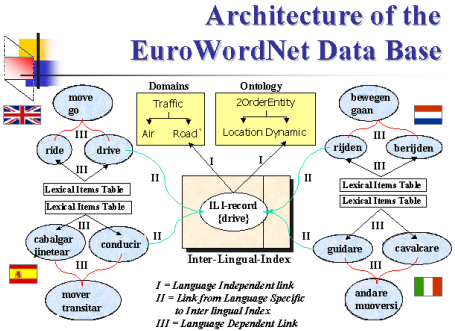
WordNets

- a network (= a graph, in terms of graph theory)
  - nodes – words, or rather "synsets" – sets of synonyms
  - (directed) edges – semantic relations, especially hyponymy and hyporonymy
  - edges constitute a directed acyclic graph
- Princeton WordNet for going back to 1985 – a monolingual version (for English)



WordNets for multiple languages

- wordnets exist for 200+ languages, varying size
- multilingual wordnets: collections of wordnets for individual languages, plus added cross-lingual correspondce links: EuroWordNet, Open Multilingual WordNet, MultiWordNet...
- instead of $\mathcal{O}(N^2)$ language pairs, English is sometimes used as the hub language (e.g. in EuroWordNet)



Swadesh list

- a list of "universal concepts", compiled by Morris Swadesh
- gradual development from 1950's to 1970's, various changes in size
- a version from 1972: 100 terms
- perhaps more popular version from 1952: 207 terms
- (but shorter versions exist too)
- examples in English: they, eye, walk, black, water, hear, all, father, eat, bark, tree, flesh, one, big, not
- various criticism of the concept universality: e.g. Navajo does not have a standalone word for water (drinking water distinguished from rain water), Finnish does not have a standalone for not...

Swadesh list, 100-word version

I	dog	nose	die	smoke
you	louse	mouth	kill	fire
we	tree	tooth	swim	ash
this	seed	tongue	fly	burn
that	leaf	claw	walk	path
who	root	foot	come	mountain
what	bark	knee	lie	red
not	skin	hand	sit	green
all	flesh	belly	stand	yellow
many	blood	neck	give	white
one	bone	breasts	say	black
two	grease	heart	sun	night
big	egg	liver	moon	hot
long	horn	drink	star	cold
small	tail	eat	water	full
woman	feather	bite	rain	new
man	hair	see	stone	good
person	head	hear	sand	round
fish	ear	know	earth	dry
bird	eye	sleep	cloud	name

Etymological resources

Lexical borrowing

- borrowing – the process by which a word from one language is adapted for use in another
- an extremely important factor for studying words' origin
- e.g. there are more words in Modern English that have been gradually borrowed from French, Latin, and Greek, than words inherited from the ancestors of English
- almost every etymological resource is a cross-lingual resource by its nature
- sometimes connecting a living language with extinct languages

Sample from Etymological Dictionary of Czech (J. Rejzek, 2001)

playback z *play* 'hrát' a *back* 'zpět'. Srov. →plejboj, →bek.

plazit se, plazivý, plaz, plazí, připlazit se, odplazit se, proplazit se. P. *pełzać*, r. *ползать*, *ползти*, s./ch. *plàziti se*, *puzíti*, Psl. **polzti (sę)*, **pelz(a)ti*, **pəlzti* (B8) jsou asi odvozeniny od ie. **pel-ǵ/h-* (A1) od **pel-* 'pohybovat se (sem a tam), téci, plavat aj.' (srov. i sln. *peljáti* 'vézt'). Významově nejblíže je ř. *pélō* 'pohybuji se', *pelázomai* 'přibližuji se'. S jinými formanty sem asi patří →plout, →plachý. Srov. →plž, →oplzlý, →plouhat (se).

plazma 'tekutá složka krve; základní hmota buňky'. V 19. st. utvořeno na základě pozdnělat. *plasma*, ř. *plásma* 'tvoření, obraz, výtvor' od *plássō* 'tvořím, vymýšlím' (srov. →plastický). Původně ve spojení *Plasmacellula*, tedy doslova 'buněčný obraz, buněčné dílo'.

pláž, plážový. Z fr. *plage* tv. z it. *piaggia* 'úbočí, břeh' z pozdnělat. *plagia* tv. a to asi z ř. *plágios* 'příčný, šikmý'. Srov. →plagiát.

plebejec 'příslušník lidových vrstev', plebejský. Přes něm. *Plebejer* z lat. *plēbēius* tv., původně adj. 'lidový', od *plēbs* 'lid, dav, množství', jež souvisí s ř. *plḗ⃝thos* tv. a vzdáleněji i s naším →plný. Srov. →plebiscit.

plebiscit 'hlasování lidu'. Převzato (případně přes něm. *Plebiszit*) z lat. *plēbiscītum* 'usnesení plebejského shromáždění', což je složenina z *plēbs* (gen. *plēbis*) 'lid, dav' (viz →plebejec) a *scītum* 'usnesení, rozhodnutí', což je původem příč. trp. od *sciscere* 'usnášet se; rozhodovat se'.

plec 'část zad nad lopatkou', plecko. Všesl. - p. *plece*, r. *плечó*, s./ch. *plèće*, stsl. *plešti*, Psl. **plet'a* (B3) je příbuzné s lot. *plecs* tv., střír. *leithe* 'rameno, lopatka' a snad i chet. *paltana-* 'plec obětního zvířete', východiskem je nejspíš ie. **plet-* 'plochý a široký' od ie. **pel-* tv. Srov. →plást, →plán aj.

pléd 'velký vlněný šátek'. Z angl. *plaid* tv. ze skot. *plaide* (srov. ir. *plaid* 'přikrývka, oděv') a to od ie. **pel-* 'obléci, přikrýt', o němž viz →plátno.

plédovat kniž. 'přimlouvat se za něco, obhajovat'. Z fr. *plaider* 'obhajovat (u soudu)' od stfr. *plaid* 'úmluva, smlouva' z lat. *placitum* 'zásada, úřední výnos' od *placēre* 'líbit se, být vhodné'.

plech, plíšek, plechový, plechovka, plechovkový, plecháč. Stejně jako p. *blach(a)*, sln. *pleh* přejato ze stfrn. *blech* tv., původně 'to, co se leskne', příbuzné je stfrn. *blicken* (viz →blikat).

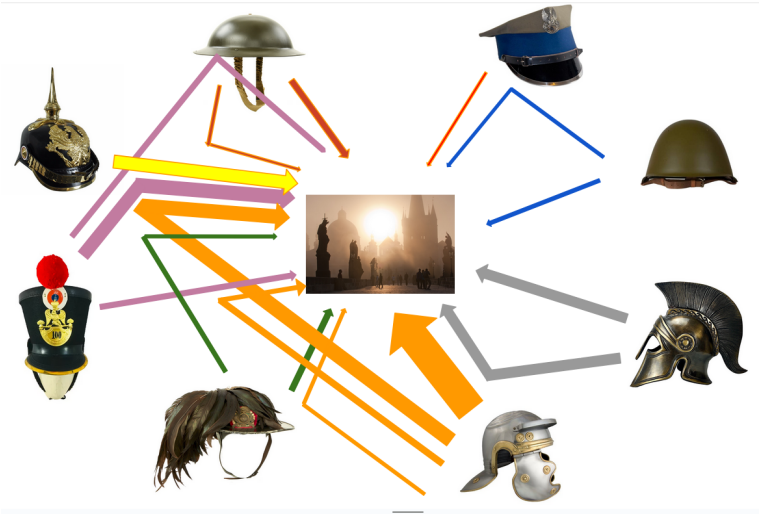
plejáda 'skupina významných osob (umělců, sportovců ap.)'. Podle souhvězdí *Plejády*, nesoucího jméno sedmi dcer obra Atlanta (ř. Pleiades), které byly podle ř. mytologie proměněny v holubice a pak ve hvězdy.

plejboj 'světák, muž žijící jen pro zábavu'. Z angl. *playboy* tv. z *play* 'hrát' (souvisí s něm. *pflegen* 'zabývat se něčím, opatrovat') a *boy* 'chlapec' nejistého původu. Srov. →playback, →kovboj.

plejtvák 'druh velryby'. Obrozenecký výtvor (Presl) od staršího č. *plýtva*, *plejtva* 'ploutev' podle výrazné hřbetní ploutve. Viz →ploutev.

plemeno, plémě, plemenný, plemeník, plemenářský, plemenářství, plemenit (se). Všesl. (kromě luž.) - p. *plemię*, r. *плéмя*, s./ch. *plême*, stsl. *plemę*. Psl. **plemę* (gen. **plemene*) se obvykle vykládá z **pled-men-* (A9), jehož základ souvisí s. →plod, tedy 'to, co se plodí'.

Most intensive borrowing paths into Czech



Etymological wordnet

- http://etym.org/
- information about how words in different languages are etymologically related
- in spite of the name, it contains also various pieces of information about pronunciation, word formation, translation equivalents etc.

Etymological wordnet – example

Conclusions

Disclaimer

- my selection of the presented resources – inevitably subjective
- many other families of multilingual data resources
  - focused on syntactic features, e.g. multilingual valency lexica
  - phonological lexical databases
  - embedding representations
  - ...

Take-home message (1)

- whatever lexical phenomenon you study ...
- ...almost certainly you can find an online resource for it,
- ...and almost certainly you can find a number of similar resources for various languages (sometimes even cross-lingually interlinked resources),
- ...and almost certainly these resources will hugely differ in size, quality, underlying linguistic theory, and in many other aspects :)

Take-home message (2)

- there are fuzzy boundaries among various phenomena related to vocabulary (e.g. inflection – wordformation – etymology), and hence also the resources overlap
- genuinly new data resources are developed relatively rarely nowadays
- it is much more common that new resources result from various (semi)automatic processing making use of the already existing resources

Homework specification

HW2 specification

- Task: design a scoring function that could be used for approximating lexical similarity languages, compute pairwise similarity scores for a set of at least 10 languages, and visualizes the scores.
- You can use any scoring function that makes sense to you, fox example
  - some kind of vocabulary overlap (based on actual wordforms, or lemmas, or cognates ...),
  - a minimal fallback solution: overlap of letter bigrams extracted from languages' 100 most frequent forms
  - the similarity function does not have to be symmetric (e.g. something based on KL divergence)
- As for visualization
  - you can produce a dendrogram (does it resemble a phylogenetic tree), or a heatmap, or a graph (nodes+edges) e.g. with edge thickness corresponding to the similarity score
  - a minimal fallback solution: a 2D table in a spreadsheet
- Write a short report (0.5 A4 page) about your findings.

HW2 submission

- Like with HW1, submission via gitlab (see the instructions for HW1 for details)
  - Create directory 'hw2' and upload (commit+push) your solution, ideally in a form of a Python code executed from a Makefile (don't upload the data, as they should be downloaded by the Makefile) ; upload also the short report (a PDF file)
- Deadline: see this course's main web page