Lexikální databáze DeriNet jako formální model tvoření slov v češtině

Magda Ševčíková a Zdeněk Žabokrtský

23. 11. 2021

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

BY NC SA
unless otherwise stated

Osnova

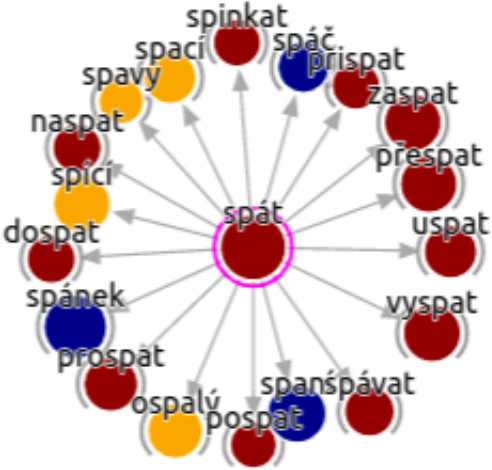
O DeriNetu

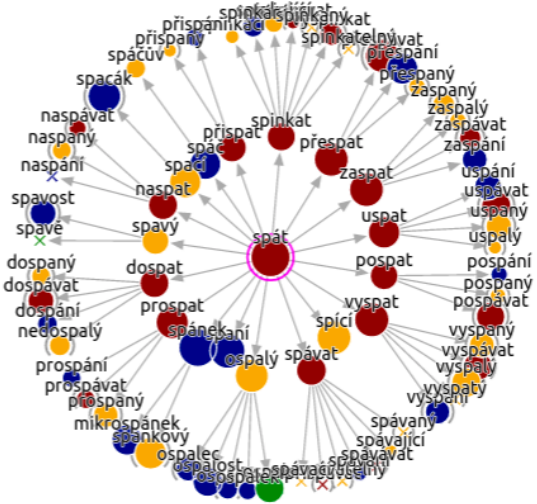
O morfémech

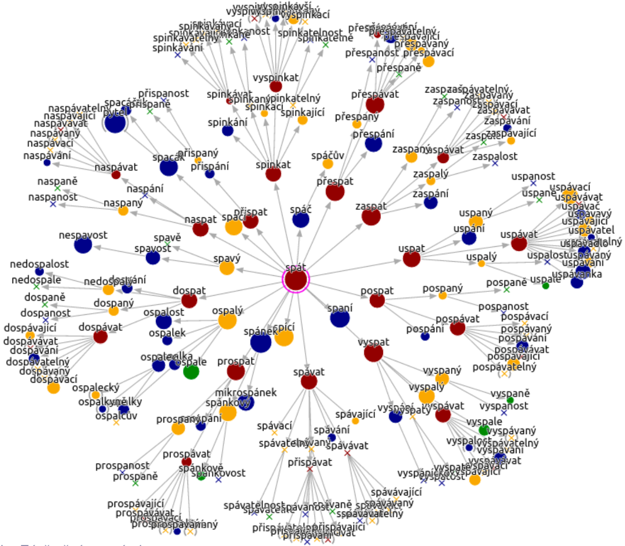
O paradigmatech

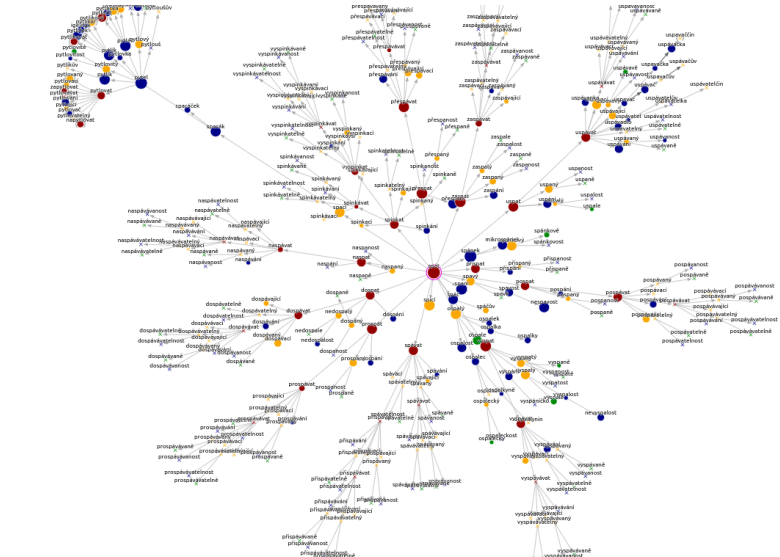
Závěrečné poznámky

O DeriNetu


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


spinkat
přispánikat
přispaný
spinkatelný
spáčův
přespání
spacák
přespaný
zaspaný
zaspalý
naspávat
zaspávat
naspaný
přispat
přespat
zaspání
naspání
spáč
zaspat
uspání
naspat
uspávat
spavost
uspat
uspaný
spavé
spavý
uspalý
spát
dospaný
pospat
pospání
dospat
dospávat
pospaný
prospat
vyspat
pospávat
dospání
spící
nedospalý
spánekání
vyspaný
vyspávat
ospalý
spávat
prospání
vyspalý
prospávat
vyspatý
prospaný
vyspání
mikrospánek
spávaný
spankový
spávající
ospalec
spávávat
ospalost
spávávající
ospaté


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



DeriNet jako formální model slovotvorby

- v pojmech teorie grafů:
  - lemma = uzel
  - derivace = hrana
  - slovotvorné hnízdo = strom
  - všechny derivační vztahy v jazyce = les

Odkud se v DeriNetu berou uzly?

- v prvních verzích DeriNetu (r. 2013): cca 200k korpusově nejčastějších lemmat
- od r. 2015: 1M lemmat z MorfFlex.CZ (Hajič & Hlaváčová 2013)
- množinu lemmat vždy synchronizujeme s novými verzemi MorfFlexu
- příklady nedávných změn: jiné pojetí homonym, přidaná vlastní jména, změny v technických příponách lemmat atd.

Odkud se v DeriNetu berou hrany?

- hlavní metodologický přístup: oportunní predace
  - kombinace nejrůznějších datových datových zdrojů, pravidlových metod, strojového učení, vlastních anotací apod.
  - desítky komponent
- technické řešení: po aktualizaci množiny lemmat lze spustit všechny komponenty pro generování hran "od nuly"
- aktuálně cca 840k hran

Ještě k velikosti

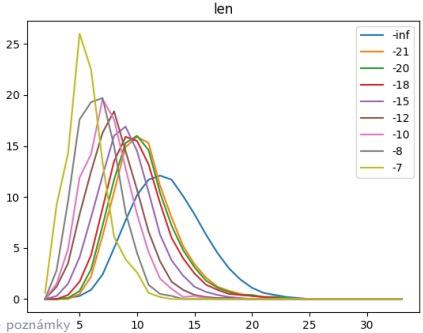
- obvyklé reakce z publika na 1M lemmat a 840k derivací:
  - nemluvčí češtiny: *Wow!*
  - mluvčí češtiny: *Bože, co to tam máte za divná slova?!?*
- relativní četnost 1 ku 2^{20}: *detailizování*, *pětasedmdesátinásobek*, *rojený*, *sprosťácky*, *anekdotářský*, *uzounce*, *zpatlání*, *igelitově*, *rozfrázovat*, *štokovací*
- rychlé jasné snadné řešení: minimální práh na počet korpusových výskytů
- problém specifický pro modelování slovotvorby: vynechání periferie vede k rozpojení i v centrálnější části slovní zásoby

Korpusové frekvence v DeriNetu

- nutnosti hranice mezi centrem a periferií se úplně vyhnout neumíme
- ale můžeme tu hranici vytlačit co nejvíce do periferie
- a uživateli dát k dispozici korpusové frekvence
- v DeriNet 2.1 korpusové frekvence pro lemma+POS z SYN4
- (nový problém: nitroslovnědruhová homonymie)
- korpusově nedoložená lemmata z MorfFlex.CZ ponecháváme též

Příklady použití korpusových frekvencí

- váhový koeficient při evaluaci
- sampling dávek k ruční anotaci (pro trénování i evaluaci)
  - např. uniformně přes exponenciálně odstupňovaná frekvenční pásma
- zkoumání vlivu frekvence na ...
- příklad: pravděpodobnostní distribuce (svislá osa) délky slova (vodorovná) v jednotlivých frekvenčních pásmech (barva)

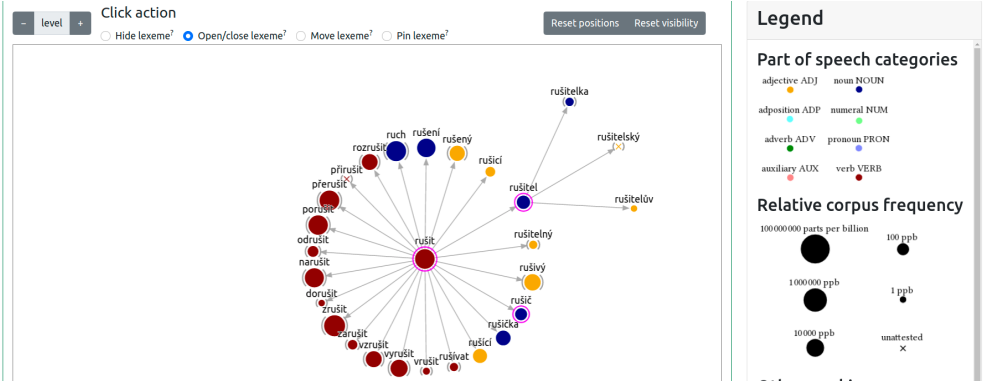


Derivace jako strom?

- kladivový zákon
- strom – spojitý acyklický graf
  - derivační vztahy – opravdu spojitý graf?
  - derivační vztahy – opravdu acyklický graf?
- podobné problémy jako u závislotní syntaxe, podobná řešení:
  - umělé fiktivní lexémy (např. *lidnit*, *bízet*, *běhnout*)
  - konvence pro volbu kostry u alternativních cest tvořících cyklus (např. pořadí prefix/suffix u *odčerpání*)
- (po začlenění kompozice je strom samozřejmě neudržitelný)

Krásný nový vyhledávač DeriSearch 2.0

- dotazovací jazyk podobný CQL
- on-line řešení s vyhledávačem na straně serveru a webovým GUI (Vidra & Žabokrtský, 2020)
- ukázka dotazu: `[pos="VERB"] ([lemma=".*ič"], [lemma=".*tel"])`



Z Guinessovy knihy derivačních stromů

- nejhlubší derivační strom:
  - nejhlubší korpusově doložený uzel: *nezpochybnitelnost* v 10. patře
  - nejhlubší nedoložený: *zpochybňovatelčin* v 11. patře
- největší derivační strom
  - největší strom: 1874
  - nejvíce doložených lemmat: 685 (sloveso *stát*)
  - nejvíce nedoložených lemmat: 1291 (*téci*)
- nejvyšší stupeň uzlu (počet přímých derivátů)
  - 41 *řezat*

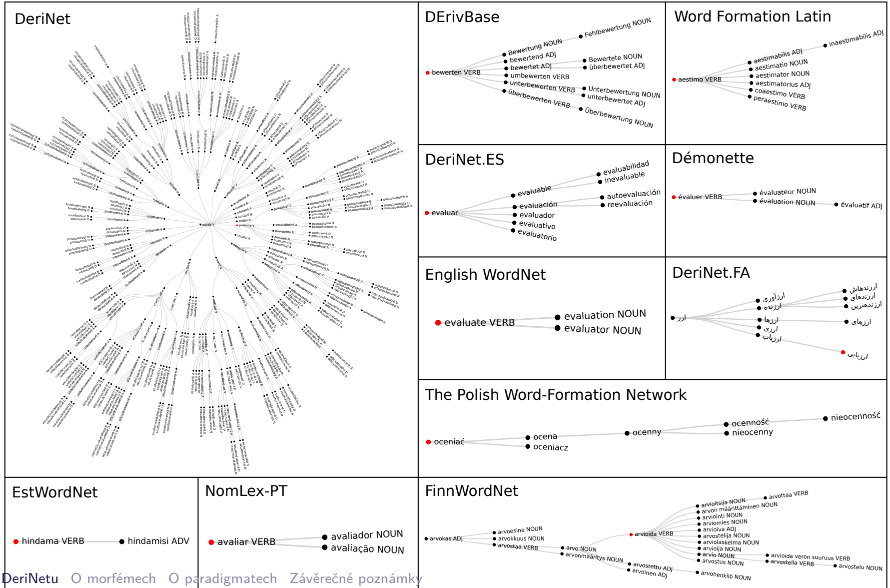
Vybrané technikálie

- souborový formát inspirovaný CoNLL-U
- pythonové API `https://github.com/vidraj/derinet/`
- distribuce dat DeriNetu přes repozitář LINDAT/CLARIAH-CZ
- licence CC BY-NC-SA 3.0

Universal Derivations

- https://ufal.mff.cuni.cz/universal-derivations
- UDer 1.1: 31 harmonizovaných derivačních datových zdrojů
  http://hdl.handle.net/11234/1-3247
  – 21 jazyků
  – rozdílná velikost i kvalita

Universal Derivations



Vyhledávání v Universal Derivations: opět DeriSearch

https://quest.ms.mff.cuni.cz/derisearch2/v2/databases/

Search

DeriSearch database list

	Language	Original resource	Harmonized resource	License	Authors of the original resource
Search	Catalan	Etymological WordNet 2013	UDer-1.1 (README)	CC BY-SA 3.0	Melo, Gerard de; Weikum, Gerhard
Search	Croatian	CroDeriv 0.9	UDer-1.1 (README)	CC BY-NC-SA 3.0	Šojat, Krešimir; Filko, Matea; Štefanec, Vanja
Search	Croatian	DerivBase.HR 1.0	UDer-1.1 (README)	CC BY-SA 3.0	Šnajder, Jan
Search	Czech	DeriNet 2.1	UDer-1.1 (README)	CC BY-NC-SA 3.0	Vidra, Jonáš; Žabokrtský, Zdeněk; Ševčíková, Magda; Kalužová, Adéla; Kyjánek, Lukáš; Dohnalová, Šárka; Svoboda, Emil; Bodnár, Jan; Hudeček, Vojtěch; Straka, Milan; Mediankin, Nikita
Search	Czech	Etymological WordNet 2013	UDer-1.1 (README)	CC BY-SA 3.0	Melo, Gerard de; Weikum, Gerhard
Search	English	CatVar 2.1	UDer-1.1 (README)	OSL-1.1	Habash, Nizar; Dorr, Bonnie
Search	English	The Morpho-Semantic Database 1.0	UDer-1.1 (README)	CC BY-NC-SA 3.0	Fellbaum, Christiane; Osherson, Anne; Clark, Peter E.
Search	Estonian	EstWordNet 2.4	UDer-1.1 (README)	CC BY-SA 3.0	Kahusk, Neeme; Vider, Kadri; Kerner, Kadri; Orav, Heili; Parm, Sirli
Search	Finnish	FinnWordNet 2.0	UDer-1.1 (README)	CC BY-SA 4.0	Lindén, Krister; Niemi, Jyrki; Hyvärinen, Mirka
Search	French	Démonette 1.2	UDer-1.1 (README)	CC BY-NC-SA 3.0	Hathout, Nabil; Namer, Fiammetta

O morfémech

Základní pojmy

- morfém - nejmenší část slova, která nese význam
- morf - konkrétní forma morfému
- alomorfy - různé morfy k témuž morfému

Jak souvisí DeriNet a morfosegmentace?

- derivace a segmentace na morfy: dvě strany téže mince
- prototypicky: jeden derivační krok = jeden přidaný morfém
- ovšem četné asymetrie:
  - některé jiné slovotvorné procesy (kompozice)
  - cirkumfixace
  - konverze
  - flexe

Pro inspiraci několik segmentačních příkladů

Pro inspiraci několik segmentačních příkladů

- *pro inspir-ac-i ně-kolik segment-ač-n-ích pří-klad-ů*

Pro inspiraci několik segmentačních příkladů

- *pro inspir-ac-i ně-kolik segment-ač-n-ích pří-klad-ů*
- nebo *pro in-spir-ac-i ně-kolik seg-ment-ač-n-ích pří-klad-ů* ?

Předchozí výzkum segmentace pro češtinu

- Eleonora Slavíčková: *Retrográdní morfematický slovník češtiny*, 1975.
- Zbyněk Šiška: *Bázový morfematický slovník češtiny*, 1998.
- Jiří Lebeda: *Morf-em-at-ic-k-á strukt-ur-a sou-čas-n-é češ-t-in-y*, 2010.
- nicméně: pro češtinu neexistuje moderní datový zdroj s dobrým pokrytím a svobodnou licencí

Vymezení morfému vs. etymologie

- rychlá snadná jasná odpověď: nutno zůstat v synchronním pohledu
- ve skutečnosti přechodná škála bez jasné hranice
  - *leden* – průzračné
  - *pondělek*, *říjen*, *nemoc*, *zločinecký*, *pořadatel* – hraniční
  - *přítel*, *manžel*, *žlutý-zlatý*, *úterý-vteřina* – určitě mimo synchronní vnímání (alespoň u většiny mluvčích)

Vymezení morfému vs. etymologie (2)

- další problémy:
  - přejímky obecně (vč. novoklasického tvoření) – *privilegium*, *architekt*, *definovat*, *archeologie*
  - pojmenované entity – *Božanov*, *Zdislava*
  - doložená etymologie jde dokonce někdy proti analogii (Panocová 2019)
  - nebezpečí lidové etymologie – *září*, *hřbitov*, *Libeň*, *protežovat*
- závěr:
  - vymezení morfému nelze dokonale objektivizovat
  - těžko oddělitelné od individuální znalosti dalších jazyků a světa vůbec
  - s Wittgensteinem: Hranice mých morfémů jsou hranice mého světa

První průzkum bojem: vlastní anotace

- v tuto chvíli cca 5000 ručně segmentovaných lemmat
- cca 2000 ručně segmentovaných forem

Segmentace a homonymie

- segmentace závislá na zvoleném čtení
  - napříč slovními druhy: *bá-l* vs. *bál*
  - i uvnitř slovního druhu: *od-rol-ov-a-t* vs. *o-drol-ov-a-t*
- homonymie mezi afixy – velmi časté
  - *mír-n-ě – kin-ě*
- homonymie afix vs. kořen – méně časté
  - *ech – kin-ech*
- dokonce homonymie prefix vs. suffix vs. interfix:
  - *o-tří-t, kin-o, svět-o-ná-zor*
  - *kin-a , a-morál-n-í, mlék-a-chti-v-ý*

Allomorfie

- některé varianty průzračné: krácení/dloužení
- ale obecně naivní, že by šlo spoléhat na jednoduchou řetězcovou podobnost
- příklady:
  - hláskové změny v kořeni (extrém: *vy-ň-a-v-še vy-j-mou-t*)
  - allomorfie v důsledku pravopisu: *u-po-zor-ň-ova-t* vs. *u-po-zor-n-i-t*
- příklady allomorfních skupin:
  - *hlas hlás hláš hlaš hles*
  - *jís jíd jed níd nís něd jez něz*
  - *péc peč pík pék pek píč pec*
  - *svít svit svíc svěc svět svic*
  - *hn žen hán hon háň han hůn hůň*

Překrývání segmentů?

- *český*
- *obléci*

Segmentace a kompozice

- *mil-o-srd-n-ě*
- nelze očekávat, že segmentace kompozita bude přímočaře odpovídat segmentaci vstupních lemma: *jednohrbý, blanokřídlý*

Míry pro měření kvality segmentace

- P/R/F1 na morfech
- P/R/F1 na hranicích mezi morfy (morfematický šev)
- procentuální úspěšnost v kompletně správném rozsegmentování slova
- ALE: čím to vážit? asi nutno použít frekvence
- jinak hrozí značné zkreslení (dlouhých slov je více, jsou méně častá, ale pravidelnější a proto na segmentaci snadnější)
- jak přesně? přímo relativní četností? nebo zploštit logaritmem?
- otevřená otázka

Automatická segmentace

- baseline: populární morfosegmentační nástroj Morfessor (Creutz & Lagus 2005)
- s výhodou lze ale využít znalost derivačního stromu (Bodnár 2020):
  - projdu strom od listů nahoru z uzlů do kořene, cestou zpátky zpropaguji nalezené hranice
  - nutno kombinovat s další optimalizací (DL klasifikátor přítomnosti morfematického švu, implementace v TensorFlow)

Method	Morphemes			Boundaries			Words
	Precision	Recall	F1	Precision	Recall	F1	Accuracy
Our method	58.94%	62.62%	63.74%	96.11%	91.63%	93.81%	72.15%
EMmh,i1	64.82%	66.54%	65.67%	91.38%	87.42%	89.36%	35.45%
FlatCat unsup.	37.19%	21.57%	27.31%	97.98%	64.92%	78.10%	1.17%
FlatCat sup.	66.20%	57.72%	61.67%	92.59%	81.15%	86.49%	31.10%

Využití segmentace při hledání v ČNK

- ve spolupráci s Michalem Křenem a Václavem Cvrčkem
- prototyp: kopie syn2015 s přidaným atributem `morfeman` zachycujícím morfematické hranice a reprezentativní allomorf
- příklad hodnoty `morfeman`: `|sou:sou:P|hlás:hlas:R|k:k:S|a:a:S|`
- dotaz: `[morfeman=".+:hlas:R.+"]`

Směrem k "Universal Segmentations"

- pokus o vytvoření zharmonizované kolekce morfosegmentačních dat pro různé jazyky
- hackaton v listopadu 2021: Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, Jonáš Vidra, Zdeněk Žabokrtský
- v tuto chvíli částečně zharmonizováno cca 20 zdrojů, celkem 150 jazyků, extrémně různá kvalita i velikost dat
- příklady:
  - Ab + bau + ge + recht + ig + keit, Ab + be + ruf + ung
  - advant + age + ous + ly, advers + ari + es
  - a + journe + ment, combatt + ant

Dílčí experimenty související se segmentací

- Jan Bodnár - alomorfie a embeddingy
  - hypotéza: vektorové reprezentace alomorfů by si měly být blízké
- Niyati Bafna - transfer embeddingů z hindštiny do márátštiny bez paralelních dat
  - hypotéza: EM algoritmus pro indukci bilinguálních embeddingů by mohl fungovat lépe pro morfémy než pro slova
- Ebrahim Ansari - zavedení atomických podřetězců ve fastTextu
  - hypotéza: podřetězce odpovídající morfémům je ve fastTextu efektivnější nedělit na ještě kratší podřetězce

Sci-fi okénko: morfémy a evoluce přirozeného jazyka

- vzdálenější výhled: evoluce jazyka jako migrace morfémů v čase a prostoru
- podobnosti s biologickou evolucí
  - dva předpoklady: rozrůzněnost a selekční tlak
  - konkurence o ekologické niky (například přirozená nutnost vyjádření konatele)
  - evoluci jazyka nelze vidět, jen zpětně rekonstruovat
  - evoluce nemá směr, nelze ji predikovat
- odlišnosti
  - evoluce jazyka – mnohem mnohem rychlejší (homo sapiens cca 10k generací, čeština cca 40 generací mluvčích)
  - "fosilní záznam" vývoje jazyka jen za nepatrné časové období
  - "horizontální přenos morfémů" četnější než genů

Morfém jako jediný formalizovatelný mem?

- mem
  - jako kulturní analogie genu (Dawkins 1976), replikující se kulturní informace
  - těžko uchopitelná jednotka, memetika některými považovaná za pseudovědu
- morfém - narozdíl od ostatních "memů" relativně diskrétní a atomická jednotka

O paradigmatech

Paradigma v morfologii flektivní a derivační

- flektivní / tvaroslovné paradigma
  - = "soubor tvarů ohebného slova představující systém jeho gramatických kategorií" (Osolsobě, `czechency.org`, heslo Morfologické paradigma)
- derivační / odvozovací paradigma
  - = lexém a jeho přímé deriváty
  - – Furdík (2004: 74nn): "usporiadané súbory komotivátov tvoria akoby sústavu 'slovotvorných pádov', ktorú však nemožno pokladať za takú ustálenú, akou je morfologická paradigma"
  - – viz Dokulilův (1962: 12nn) termín *slovotvorný svazek*

*škol-**a***	:	*škol-**y***.gen-sg	*škol(a)*	:	*škol-**k(a)***.DIMINUTIVE
		*škol-**e***.dat-sg			*škol-**n(í)***.RELATION
		*škol-**u***.acc-sg			*škol-**sk(ý)***.RELATION2
		*škol-**o***.voc-sg			*škol-**ník***.ACTOR/AGENT/PERSON
		*škol-**e***.loc-sg			*škol-**ák***.?
		*škol-**ou***.instr-sg			

Paradigma v morfologii flektivní a derivační

- flektivní / tvaroslovné paradigma
  - = "soubor tvarů ohebného slova představující systém jeho gramatických kategorií" (Osolsobě, `czechency.org`, heslo Morfologické paradigma)
- derivační / odvozovací paradigma
  - = lexém a jeho přímé deriváty
  - – Furdík (2004: 74nn): "usporiadané súbory komotivátov tvoria akoby sústavu 'slovotvorných pádov', ktorú však nemožno pokladať za takú ustálenú, akou je morfologická paradigma"
  - – viz Dokulilův (1962: 12nn) termín *slovotvorný svazek*

*škol-**a***	:	*škol-**y***.gen-sg	*škol(a)*	:	*škol-**k(a)***.DIMINUTIVE
		*škol-**e***.dat-sg			*škol-**n(í)***.RELATION
		*škol-**u***.acc-sg			*škol-**sk(ý)***.RELATION2
		*škol-**o***.voc-sg			*škol-**ník***.ACTOR/AGENT/PERSON
		*škol-**e***.loc-sg			*škol-**ák***.?
		*škol-**ou***.instr-sg			

Paradigma v morfologii flektivní a derivační

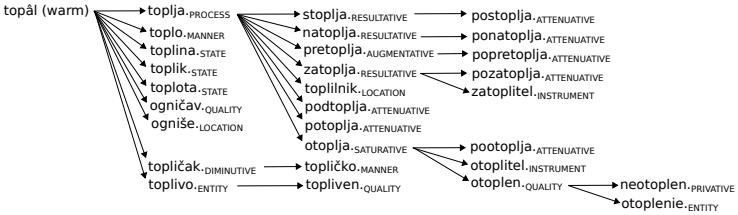
- flektivní / tvaroslovné paradigma
  - = "soubor tvarů ohebného slova představující systém jeho gramatických kategorií" (Osolsobě, `czechency.org`, heslo Morfologické paradigma)
- derivační / odvozovací paradigma
  - = lexém a jeho přímé deriváty
  - – Furdík (2004: 74nn): "usporiadané súbory komotivátov tvoria akoby sústavu 'slovotvorných pádov', ktorú však nemožno pokladať za takú ustálenú, akou je morfologická paradigma"
  - – viz Dokulilův (1962: 12nn) termín *slovotvorný svazek*

*škol-**a***	:	*škol-**y***.gen-sg	*škol(a)*	:	*škol-**k(a)***.DIMINUTIVE
		*škol-**e***.dat-sg			*škol-**n(í)***.RELATION
		*škol-**u***.acc-sg			*škol-**sk(ý)***.RELATION2
		*škol-**o***.voc-sg			*škol-**ník***.ACTOR/AGENT/PERSON
		*škol-**e***.loc-sg			*škol-**ák***.?
		*škol-**ou***.instr-sg			

Derivační paradigma jako jedna dimenze derivační sítě

Körtvélyessy et al.: *Derivational Networks across Languages*. De Gruyter 2020.

- 30 nemotivovaných slov a jejich přímé i nepřímé **deriváty** ve 40 evropských jazycích
  - 10 substantiv: *bone*, *eye*, *tooth*, *day*, *dog*, *louse*, *fire*, *stone*, *water*, *name*
  - 10 sloves: *cut*, *dig*, *pull*, *throw*, *give*, *hold*, *sew*, *burn*, *drink*, *know*
  - 10 adjektiv: *bad*, *new*, *black*, *straight*, *warm*, *old*, *long*, *thin*, *thick*, *narrow*
- tři dimenze derivační sítě
  1. deriváty uspořádány do derivačních paradigmat = vertikální dimenze
  2. deriváty uspořádány do derivačních řad = horizontální dimenze
  3. deriváty ohodnoceny sémantickou kategorií (Bagasheva 2017)



http://www.ugr.es/~svalera/Monika/index.html

Paradigma pouze v derivaci? Slovotvorné paradigma

- Štekauer (2014: 369)
  - "It may be assumed that the concept of the derivational paradigm is restricted to affixation in terms of word-formation processes participating in its constitution. ... [T]he idea of a derivational paradigm, constituted by all word-formation processes, is paradigmatically vacuous because it does not lead to a predictable and regularly organized system of complex words."
- Dokulil (1962: 14)
  - *slovotvorné paradigma* = souhrn všech typů úměr napříč slovotvornými hnízdy

    – úměry mezi *slovotvornými řadami*:

list	>	*lístek*	>	*lístkový*	>	*lístkovitý*
květ	>	*kvítek*	>	*kvítkový*	>	*kvítkovitý*

    – úměry mezi *slovotvornými svazky*:

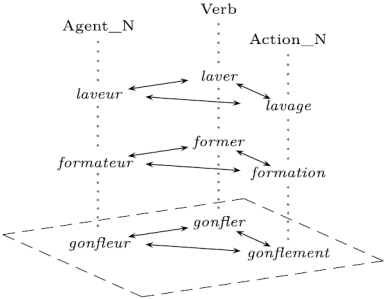
medvěd	> *medvědí*	*slon*	> *sloní*	*velbloud*	> *velbloudí*
	> *medvídě*		> *slůně*		> *velbloudě*
	> *medvědice*		> *slonice*		> *velbloudice*

- Fernández-Domínguez a kol. (2020): paradigmata v kompozici atd.

Slovotvorné paradigma se členy vymezenými sémanticky

- Bauer (2019: 156)
  - *paradigm of functions*: "Sometimes, the precise form which will be found in a paradigm is not predictable, but we can predict the presence of a form to fill that particular function."

- Bonami & Strnadová (2019)
  - *paradigma* = sada lexémů, které v daném hnízdě vyjadřují určité sémantické kategorie
    – slovotvorný proces ani formant nebrány v potaz
    – nezahrnuje všechny členy hnízda
    – každý člen spojen s každým
    – nerozlišován motivující vs. motivovaný člen (hrany nejsou orientované)
  - *paradigmatický systém* = soubor paradigmat napříč hnízdy



propose	*proposal*.ACTION	*proposer*.AGENT
inspect	*inspection*.ACTION	*inspector*.AGENT
protest	*protest*.ACTION	*protester*.AGENT

Cizí slova v češtině

- srovnávací studie čeština vs. angličtina
- dvojice substantiva a slovesa utvořené bez derivačního materiálu
  - v češtině rozdíl ve flektivních morfémech (téma a infinitivní koncovka u slovesa)
    start-ova-t : *start* = konverze / transflexe / derivace (Ševčíková, v tisku)
  - angl. substantivum a sloveso mají formálně totožný slovníkový tvar
    start.v : *start*.n = konverze (Valera, 2015)
- paradigmatický přístup
  - konverzní dvojice uskupeny podle sémantického vztahu mezi subst. a slovesem
  - bez ohledu na směr tvoření (sloveso > subst vs. subst < sloveso)

  – *experiment* > *experimentovat* (= dělat experimenty)
    experiment.n > *experiment*.v (= to make an experiment upon)
  – *revize* > *revidovat* (= konat revizi)
    revise.n (= the action of revising) < *revise*.v

Ševčíková & Hledíková, forth.

Cizí slova v češtině: agentní a dějové substantivum

- 165 dvojic sloves a formálně totožných substantiv z British National Corpus
- jejich přímé protějšky v češtině extrahovány ze SYN2000
  - v češtině a v angl. v jednotlivých dvojicích stejný sémantický vztah
- konverzní dvojice analyzovány spolu s agentním a dějovým substantivem

telefonovat	*telefon*.INSTR	*telefonista*.AGENT	*telefonát*.ACTION
telephone.v	*telephone*.n.INSTR	*telephonist*.AGENT	X
filtrovat	*filtr*.INSTR	X	*filtrace*.ACTION
filter.v	*filter*.n.INSTR	X	*filtration*.ACTION
extrahovat	*extrakt*.RESULT	X	*extrakce*.ACTION
extract.v	*extract*.n.RESULT	*extractor*.AGENT	*extraction*.ACTION
ruinovat	*ruina*.RESULT	X	*ruinace*.ACTION
ruin.v	*ruin*.n.RESULT	X	*ruination*.ACTION

Cizí slova v češtině: tendence k nominálnímu vyjadřování

- agentní jméno v češtině nedoloženo, i když cizí protějšek k dispozici
- dějové substantivum v češtině k dispozici, i když angl. protějšek chybí

atakovat	*atak*	X
attack.v	*attack*.n	*attacker*.AGENT
transportovat	*transport*	X
transport.v	*transport*.n	*transporter*.AGENT
skórovat	*skóre*	X
score.v	*score*.n	*scorer*.AGENT

telefonovat	*telefon*	*telefonát*.ACTION
telephone.v	*telephone*.n	X
parfémovat	*parfém*	*parfemace*.ACTION
perfume.v	*perfume*.n	X
archivovat	*archiv*	*archivace*.ACTION
archive.v	*archive*.n	X

- v některých hnízdech v češtině sloveso chybí úplně
  - substantivum jako jediný prostředek pro vyjádření dějového významu

X	*kremace*.ACTION
cremate	*cremation*.ACTION
X	*demise*.ACTION
demit	*demission*.ACTION

X	*infuze*.ACTION
infuse	*infusion*.ACTION
X	*incidence*.ACTION
incide	*incidence*.ACTION

Vid u denominativních vs. nemotivovaných (a deverb.) sloves

- většina cizích sloves v češtině kompatibilní s jedinou kmenotvornou příponou
  - vidový protějšek tvořen předponou nebo netvořen vůbec (obouvidovost)
    test-ova-t > *o-test-ova-t*, *balet-i-t*, *explod-ova-t*
  - dvě témata u malé skupiny cizích sloves
    klik-a-t : *klik-nou-t*, *risk-ova-t* : *risk-nou-t*

- u domácích sloves kompatibilita s jediným tématem typická pro denominativa
  - vidový protějšek tvořen předponou
    (*noc* >) *noc-ova-t* > *pře-noc-ova-t*, (*sůl* >) *sol-i-t* > *o-sol-i-t*
- domácí nemotivovaná slovesa (a deverbativní slovesa)
  - tvoření vidového protějšku změnou kmenotvorné přípony
  - motivují bezpříponová substantiva
    skák-a-t : *skoč-i-t* (> *skok*), *od-káz-a-t* : *od-kaz-ova-t* (> *odkaz*)

Deverbativní bezpříponová substantiva vs. slovesa utvořená od bezpříponových substantiv

skoč-i-t	>	*skok*
pot-i-t	<	*pot*
soud-i-t	?	*soud*

- dvojice slovesa a bezpříponového substantiva jako součást paradigmatu
  - vidový protějšek s odlišným tématem jako zvláštní slot
  - rozdíly v doloženosti bezpříponových substnativ

skákat *křičet*	*skočit* *křiknout*	*skok* *křik*	*vyskočit* *vykřiknout*	*vyskakovat* *vykřikovat*	*výskok* *výkřik*
potit	X	*pot*	*zpotit*	X	X
prášit	X	*prach*	*oprášit*	*oprašovat*	X
listovat	X	*list*	*prolistovat*	*prolistovávat*	X
soudit	X	*soud*	*odsoudit*	*odsuzovat*	X

- určení směru tvoření podle paradigmatu

Ševčíková 2021, Ševčíková, v tisku

Konatelská/činitelská jména v češtině

*sjednot-i-**tel***	X	*sjednot-i-t*	X	X
*sjednoc-ova-**tel***	*sjednoc-ova-t*	X	X	X
*model-**ář***	*model-ova-t*	X	*model*	X
*zvon-**ík***	*zvon-i-t*	X	*zvon*	X
*závod/**n/ík***	*závod-i-t*	X	*závod*	*závod-n-í*
*boj-ov/**n/ík***	*boj-ova-t*	X	X	*boj-ov-n-ý*
*střel-**ec***	*stříl-e-t*	*střel-i-t*	*střel-a*	X
*kup-**ec***	*kup-ova-t*	*koup-i-t*	*koup-ě*	X

válečník – válčit – válka – válečný

target_noun_suffix	*-ník/-ík*
root_final	*č*
root_final_cvs	*consonant*
root_final_vertical	*africate*
root_final_horizont	*postalveolar*
number_prefixes	*0*
v1_theme	*i*
v1_aspect	*imp*
v1_conjug	*4*
v1_suf_asp_counterpart	*no*
v2_theme	–
v2_aspect	–
v2_conjug	–
paradigm_type	*NNA-V-*
inanim_noun	*no*
freq_parent_noun	*25,895*
freq_parent_adj	*4,953*
freq_parent_v1	*499*
freq_parent_v2	–
freq_slots	*VAN*

Ševčíková et al. 2021

Predikce agentního formantu podle paradigmatu

- A: kořen a téma motivujícího slovesa

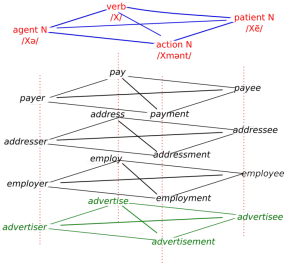
  [root_final, root_final_cvs, root_final_vertical, root_final_horizont, v1_theme, v2_theme]
- B: prefixy, téma, aspekt, konjug. třída motivujícího slovesa

  [number_prefixes, v1_theme, v1_aspect, v1_conjug, v2_theme, v2_aspect, v2_conjug]
- C: paradigma [paradigm_type, v1_suf_asp_counterpart, inanim_noun]
- D: korpusová frekvence [freq_parent_noun, freq_parent_adj, freq_parent_v1, freq_parent_v2, freq_slots]

sufix	subst.	všechny rysy (log.regr./dec.trees)	A	B	C	D
-tel	85	0.69/0.71	0.32/0.40	0.66/0.67	0.63/0.63	0.39/0.45
-č	77	0.71/0.66	0.55/0.51	0.54/0.50	0.10/0.10	0.59/0.29
-ník|*-ík*	21	0.63/0.65	0.21/0.29	0.00/0.09	0.78/0.78	0.72/0.65
-ář|*-ař*	19	0.56/0.53	0.34/0.39	0.29/0.27	0.53/0.61	0.56/0.42
-ce	13	0.38/0.46	0.22/0.36	0.47/0.48	0.44/0.33	0.45/0.48
-ák	10	0.25/0.27	0.17/0.27	0.21/0.21	0.31/0.31	0.00/0.07
-ec	6	0.40/0.62	0.18/0.21	0.15/0.15	0.13/0.10	0.60/0.14
-čí	2	0.25/0.33	0.15/0.17	0.14/0.13	0.00/0.17	0.00/0.00
všechny suf	233	0.63/0.63	0.37/0.41	0.48/0.48	0.42/0.42	0.49/0.39

Slovotvorná paradigmata v DeriNetu?

- ? ParaDeriNet
  - apriorně stanovená paradigmata plnit lexikálním materiálem
  - inspirace ve zdrojích pro jiné jazyky
- spíše: do stávajících dat doplnit sémantické značky



(Namer & Hathout 2020)

- pilotní experiment
  - 5 sémantických kategorií
    - DIMINUTIVE, FEMALE, POSSESSIVE, ITERATIVE, ASPECT
  - poloautomatická procedura
    - rodič ve stromě, slovní druh, gramatické kategorie (rod, vid)

Ševčíková & Kyjánek 2019

Slovotvorná paradigmata v DeriNetu

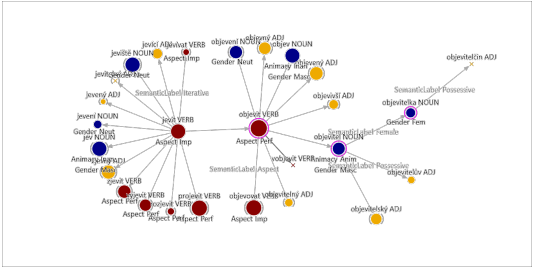
```
[pos=="NOUN"] <SemanticLabel=="Diminutive"> [pos=="NOUN"] <SemanticLabel=="Diminutive"> [pos=="NOUN"]
```

– 258 podstromů (z toho 237 s uzly s nenulovou frekvencí)

bota	*botka*	*botička*
škola	*školka*	*školička*
brouk	*brouček*	*broučínek*
džbán	*džbánek*	*džbáneček*
šál	*šálek*	*šáleček*

```
[pos=="VERB"][lemma="[a-z].*" and feats/Animacy=="Anim" and feats/Gender=="Masc"] <SemanticLabel=="Female"> []
```

– 1789 podstromů (z toho 867 s uzly s nenulovou frekvencí)



Závěrečné poznámky

Další nedávné exprimentální linie kolem DeriNetu

- Emil Svoboda - modelování kompozice
- Kristýna Neumannová - modelování univerbizace
- Lukáš Kyjánek - kompilace derivačních zdrojů pro ruštinu
- Jonáš Vidra - crosslingvální transfer derivačních stromů

Děkujeme!

Reference

- Bauer, L. (2019). Notions of paradigm and their value in word-formation. *Word Structure* 12:153–175.
- Bodnár, J. (2020)
- Bonami, O. & Strnadová, J. (2019). Paradigm structure and predictability in derivational morphology. *Morphology*, 29:167–19.
- Creutz & Lagus (2005)
- Dawkins (1976)
- Dokulil, M. (1962). *Tvoření slov v češtině I*. Academia.
- Fernández-Domíguez, J. et al. (2020). *Paradigmatic Relations in Word Formation*. Brill.
- Furdík, J. (2004). *Slovenská slovotvorba*. Náuka.
- Hajič, J. & Hlaváčová, J. (2013). *MorfFlex CZ*. LINDAT/CLARIN digital library at ÚFAL MFF UK, `http://hdl.handle.net/11858/00-097C-0000-0015-A780-9`
- Namer, F. & Hathout, N. (2020): ParaDis and Démonette. From Theory to Resources for Derivational Paradigms. *The Prague Bulletin of Mathematical Linguistics*, 114:5–33.
- Panocová, R. (2017). Internationalisms with the Suffix *-ácia* and their Adaptation in Slovak. In *Proceedings of DeriMo*. EDUCatt, 61–72.
- Ševčíková, M. (2021):). Action nouns vs. nouns as bases for denominal verbs in Czech. *Word Structure*, 14:97–128, `https://www.euppublishing.com/doi/full/10.3366/word.2021.0181`
- Ševčíková, M. (v tisku): Bezpříponová substantiva a vyjadřování vidového protikladu u příbuzných sloves. *Slovo a slovesnost*.
- Ševčíková, M. & Hledíková, H. (forth.): Paradigms in English and Czech noun/verb conversion: A contrastive study of corresponding lexemes. In *Paradigms in word formation: Theory and applications*. Benjamins.
- Ševčíková, M. & Kyjánek, L. (2019): Introducing Semantic Labels into the DeriNet Network. *Jazykovedný časopis* 70:412–423.
- Ševčíková, M. et al. (2021): Agent noun formation in Czech. Presented at ParadigMo 2021, `http://paradigmo2.eklablog.com/program-c31046450`
- Štekauer, P. (2014). Derivational paradigms. In *The Oxford Handbook of Derivational Morphology*. OUP, 354–369.
- Valera, S. (2015). Conversion. In *Word-Formation. An International Handbook of the Languages of Europe*, Vol 1. De Gruyter, 322–339.