Harmonizace jazykových zdrojů zachycujících slovotvorbu různých jazyků

Lukáš Kyjánek

Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova

Obhajoba diplomové práce, 23. června 2020

Slovotvorba zachycená v jazykových zdrojích

- slovotvorba jako způsob tvoření nových lexémů,
  viz. Dokulil 1962, Horecký 1989, Furdík 2004, Štekauer 2012

  - derivace = přidáním/odebráním/změnou lexikálního afixu u existujícího lexému
    1. der. svazek = lexémy derivované ze stejného základového lexému
       list → *líst-ek*
       → *list-oví*
       → *list-natý*
    2. der. řada = postupná derivace lexémů
       list → *líst-ek* → *lístk-ový* → *lístkov-itý* → *lístkovit-ost*
    3. der. čeleď/hnízdo/rodina = všechna derivačně příbuzná slova;
       rekurzivní kombinace svazků a řad

  - skládání = spojením dvou a více lexémů, např. *modrá* + *bílá* → *modr-o-bílý*

Existující jazykové zdroje obsahující slovotvorbu

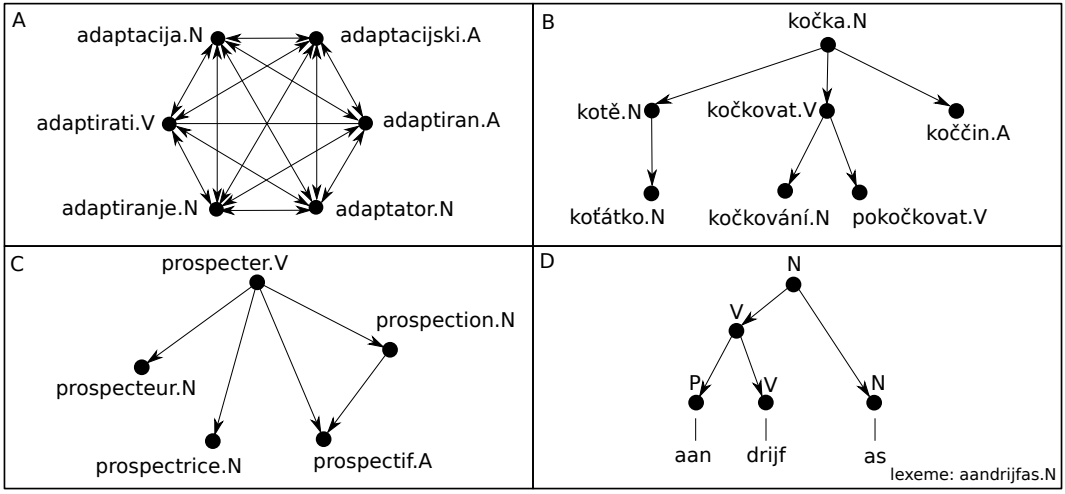
CatVar
DerIvaTario
EstWordNet
GermaNet
OpenWordNet-PT
FinnWordNet
CELEX (+3)
CroDeriV
NOMLEX
Morphonette
NomLex-PT
Prague Dep. Tr.

DerivBase
WFL
Etym. WordNet (+150)
BulNet
PlWordNet
E-Lex
The Polish WFN
The M-S Database
ADJADV
VerbAction
Unimorph
Russian National Corpus

DerivBase.Ru
Démonette
DeriNet.ES
CroWordNet
RoWordNet
E-dictionary
The Spanish WFN
Framorpho-FR
NOMADV
DErivCELEX
MorphoLex-fr

DerivBase.Hr
DeriNet
DeriNet.FA
Czech WordNet
SrpWordNet
Sloleks
WiktiWF (+5)
Nomage
NOMLEXPlus
Morphological Treebank
MorphoLex-en

Modelování slovotvorby v existujících jazykových zdrojích

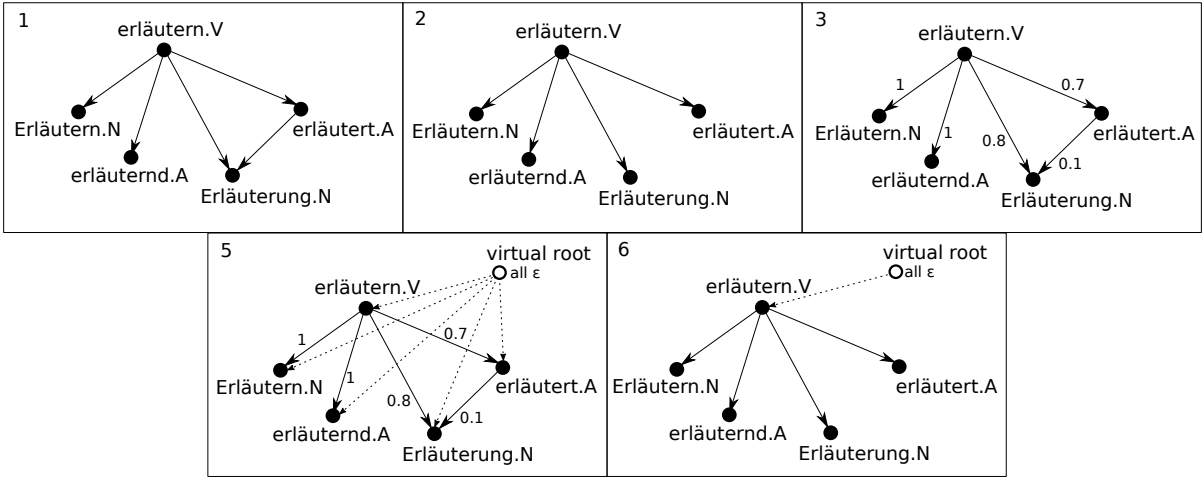


Harmonizační proces

- cílová datová struktura: (používaná v DeriNet 2.0) — zakořeněný strom (derivace), resp. slabě souvislý graf (kompozice)

- zásadní harmonizační rozhodnutí
  1. nepřidávat ani neodebírat ze vstupních zdrojů žádné derivačně příbuzné lexémy
  2. nepřidávat nové druhy anotací (např. slovně-druhové zařazení)
  3. zachovat všechny původní slovotvorné vztahy, ale restrukturovat je tak, aby bylo možné uložit je do cílové datové reprezentace

Postup harmonizace

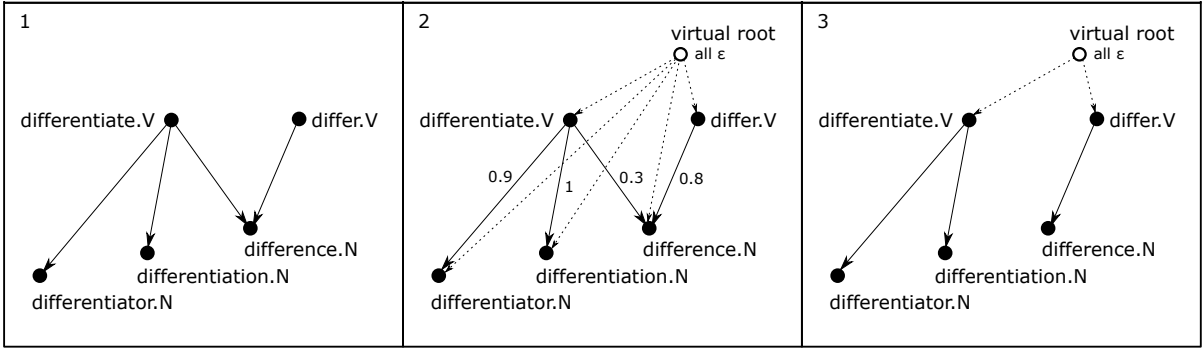


Postup harmonizace: (2) manuální anotace

- snaha unifikovat reprezentaci některých vztahů napříč zdroji, např. **negace**, varianty, výpůjčky

Postup harmonizace: (4) identifikace zakořeněných stromů

- Maximum Spanning Tree (Chu-Liu-Edmondsův algoritmus)
- v některých grafech bylo identifikováno více než jeden zakořeněný strom
- zaveden *virtual root*
  1. zabraňuje selhání běhu MST algoritmu
  2. vyhlazuje výsledné stromy



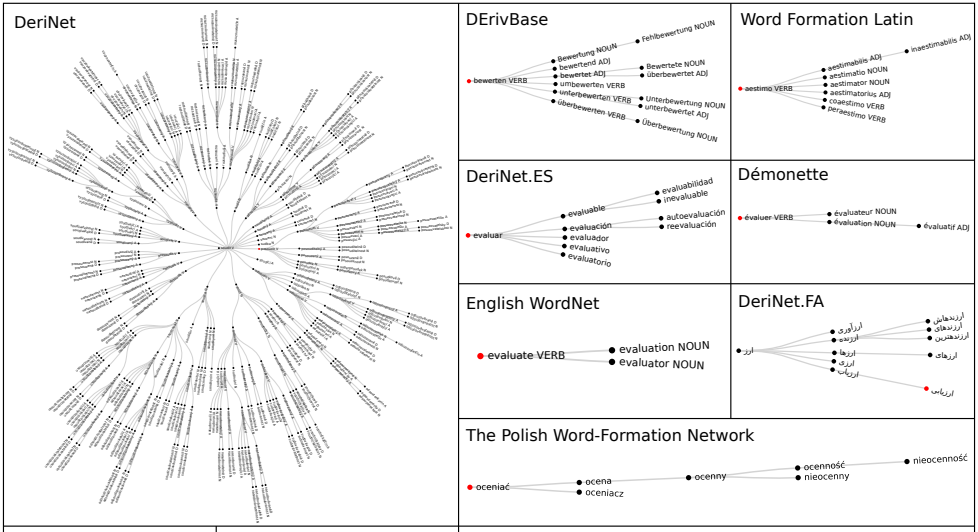
Evaluace (Baseline / ML model)

- baseline jakožto jednoduchý pravděpodobnostní model derivačních vztahů (V-N, V-A, N-V, N-A, . . .) získaný na základě trénovacích dat skóroval vztahy ve validačních a holdout datasetech
- porovnána úspěšnost baseline / ML model jak při skórování vztahů, tak při identifikaci stromů pomocí MST algoritmu (F-score)

	Skórování vztahů		**Identifikace stromů**	
Jaz. zdroj	VALIDATION	HOLDOUT	VALIDATION	HOLDOUT
CatVar	44.6 / 82.4	44.9 / 80.7	51.6 / 83.1	53.3 / **81.0**
D-CELEX	47.2 / 81.1	47.7 / 77.1	54.2 / 81.1	53.0 / **79.5**
DerIvaTario	47.7 / 77.5	47.5 / 76.0	48.7 / 78.1	50.0 / **75.1**
DErivBase	24.9 / 88.6	25.4 / 85.8	75.1 / 93.4	78.9 / **92.1**
DerivBase.Hr	45.2 / 77.2	45.4 / 80.7	56.4 / 81.1	58.3 / **81.0**
DerivBase.Ru	35.1 / 83.0	34.1 / 83.1	49.3 / 84.4	45.0 / **85.5**
E-CELEX	47.1 / 74.0	47.1 / 74.0	59.7 / 74.9	59.4 / **73.8**
FinnWordNet	38.2 / 74.0	37.8 / 70.1	62.0 / 80.2	62.9 / **76.9**
G-CELEX	45.8 / 75.6	46.1 / 76.8	57.5 / 79.5	57.5 / **77.4**

Univeral Derivations 1.0

						Vlastnosti rodin [prům./max]			Slovní druhy [%]				
Jaz. zdroj	**Jazyk**	**Lexémy**	**Der. vztahy**	**Rodiny**	**Singletony**	**Velikost**	**Hloubka**	**Šířka**	**N**	**A**	**V**	**D**	**–**
CatVar	Angličtina	82,675	24,873	57,802	45,954	1.4 / 18	0.3 / 7	0.3 / 10	60	24	11	5	0
D-CELEX	Nizozemština	125,611	13,435	112,176	107,112	1.1 / 301	0.1 / 11	0.1 / 73	63	8	8	1	21
Démonette	Francouzština	21,290	13,808	7,482	69	2.8 / 12	1.1 / 4	1.8 / 8	63	2	34	0	0
DeriNet	Čeština	1,027,665	809,282	218,383	96,208	4.7 / 1638	0.8 / 10	1.1 / 40	44	35	5	16	0
DeriNet.ES	Španělština	151,173	36,935	114,238	98,325	1.3 / 35	0.2 / 5	0.3 / 14	0	0	0	0	0
DeriNet.FA	Perština	43,357	35,745	7,612	0	5.7 / 180	1.5 / 6	3.3 / 114	0	0	0	0	0
DerIvaTario	Italština	8,267	1,787	6,480	5,255	1.3 / 13	0.2 / 5	0.2 / 6	51	26	14	9	0
DErivBase	Němčina	280,775	43,368	237,407	216,982	1.2 / 46	0.1 / 5	0.1 / 13	86	10	5	0	0
DerivBase.Hr	Chorvatština	99,606	35,289	64,317	50,100	1.5 / 945	0.3 / 21	0.4 / 863	59	30	12	0	0
DerivBase.Ru	Ruština	270,473	133,759	136,714	116,037	2.0 / 1142	0.3 / 13	0.4 / 36	62	18	17	3	0
E-CELEX	Angličtina	53,103	9,826	43,277	37,951	1.2 / 51	0.2 / 8	0.2 / 33	47	15	13	7	18
EstWordNet	Estonština	988	507	481	22	2.1 / 3	1.0 / 2	1.0 / 3	16	29	8	47	0
EtymWordNet-cat	Katalánština	7,496	4,568	2,928	8	2.6 / 13	1.1 / 4	1.5 / 13	0	0	0	0	0
EtymWordNet-ces	Čeština	7,633	5,237	2,396	14	3.2 / 48	1.1 / 4	2.0 / 42	0	0	0	0	0
EtymWordNet-gla	Gaelština	7,524	5,013	2,511	15	3.0 / 15	1.1 / 3	1.8 / 13	0	0	0	0	0
EtymWordNet-pol	Polština	27,797	24,876	2,921	19	9.5 / 75	1.1 / 3	8.3 / 66	0	0	0	0	0
EtymWordNet-por	Portugalština	2,797	1,610	1,187	9	2.4 / 57	1.0 / 3	1.3 / 57	0	0	0	0	0
EtymWordNet-rus	Ruština	4,005	3,227	778	15	5.1 / 44	1.0 / 3	4.0 / 44	0	0	0	0	0
EtymWordNet-hbs	Srbochorv.	8,033	6,303	1,730	6	4.6 / 108	1.0 / 3	3.6 / 107	0	0	0	0	0
EtymWordNet-swe	Švédština	7,333	4,423	2,910	3	2.5 / 116	1.0 / 3	1.5 / 116	0	0	0	0	0
EtymWordNet-tur	Turečtina	7,774	5,837	1,937	11	4.0 / 42	1.1 / 4	2.8 / 22	0	0	0	0	0
FinnWordNet	Finština	20,035	11,922	8,113	1,461	2.5 / 20	1.0 / 5	1.3 / 14	55	29	15	0	0
G-CELEX	Němčina	53,282	13,553	39,729	34,156	1.3 / 39	0.2 / 11	0.3 / 35	52	17	17	2	12
Nomlex-PT	Portugalština	7,020	4,201	2,819	17	2.5 / 7	1.0 / 1	1.5 / 7	60	0	40	0	0
The M-S Database	Angličtina	13,813	7,855	5,958	65	2.3 / 6	1.0 / 1	1.3 / 6	57	0	43	0	0
The Polish WFN	Polština	262,887	189,217	73,670	41,332	3.6 / 214	1.0 / 8	1.1 / 38	0	0	0	0	0
Word Formation Latin	Latina	36,417	32,414	4,003	121	9.1 / 524	1.7 / 6	4.3 / 236	46	29	21	0	4


DeriNet
DErivBase
bewerten VERB
Bewertung NOUN
bewertend ADJ
bewertet ADJ
umbewerten VERB
unterbewerten VERB
überbewerten VERB
Fehlbewertung NOUN
Bewertete NOUN
überbewertet ADJ
Unterbewertung NOUN
unterbewertet ADJ
Überbewertung NOUN
Word Formation Latin
aestimo VERB
aestimabilis ADJ
aestimatio NOUN
aestimator NOUN
aestimatorius ADJ
coaestimo VERB
peraestimo VERB
inaestimabilis ADJ
DeriNet.ES
evaluar
evaluable
evaluación
evaluador
evaluativo
evaluatorio
evalualibilidad
inevaluable
autoevaluación
reevaluación
Démonette
évaluer VERB
évaluateur NOUN
évaluation NOUN
évaluatif ADJ
English WordNet
evaluate VERB
evaluation NOUN
evaluator NOUN
DeriNet.FA
ارز
ارزآوری
ارزبر
ارزها
ارزی
ارزیاب
ارزهای
ارزهایی
ارزهایی
ارزیاب
The Polish Word-Formation Network
oceniać
ocena
oceniacz
ocenny
ocenność
nieocenny
nieocenność
EstWordNet
hindama VERB
hindamisi ADV
NomLex-PT
avaliar VERB
avaliador NOUN
avaliação NOUN
FinnWordNet
arvokas ADJ
arvoesine NOUN
arvokkuus NOUN
arvostaa VERB
arvo NOUN
arvonmääritys NOUN
arvostaja NOUN
arvostelu NOUN
arvioinen ADJ
arvioitava NOUN
arvon määrittäminen NOUN
arviointi NOUN
arviointi ADJ
arvioitsija NOUN
arvioitsijatar NOUN
arvioija NOUN
arvo NOUN
arvoton ADJ
arvohenkilö NOUN
arvottaa VERB
arvoton vertoja suoraan VERB
arvoitella VERB
arvoitelu NOUN