ČASOPIS PRO MODERNÍ FILOLOGII 98, 2016, Č. 1, S. 62–76

OPEN ACCESS

Lexikální síť DeriNet: elektronický zdroj pro výzkum derivace v češtině[1]

Magda Ševčíková (Praha) – Zdeněk Žabokrtský (Praha) – Jonáš Vidra (Praha) – Milan Straka (Praha)

THE DERINET LEXICAL NETWORK: A LANGUAGE DATA RESOURCE FOR RESEARCH INTO DERIVATION IN CZECH

The paper introduces the DeriNet lexical database, which includes more than 969,000 Czech words interconnected by 718,000 links corresponding to derivational relations (relations between a base word and a word derived from it). Derivational relations were identified by semi-automatic procedures and manual annotation. As the DeriNet network is fully compatible with a large inflectional dictionary of Czech (MorfFlex CZ), it can be used as a resource for an integrating approach to derivational and inflectional morphology of Czech both in linguistic research and in natural language processing.

KEYWORDS

derivation, inflection, lexical network, vowel and consonant alternations, homonymy

KLÍČOVÁ SLOVA

derivace, flexe, lexikální síť, hláskové alternace, homonymie

1. ÚVOD

Čeština jako morfologicky bohatý jazyk disponuje komplexním systémem jak formantů tvarotvorných, tak formantů odvozovacích. I když spojování popisu flexe s popisem odvozování v rámci morfologických výkladů nemá v teoretickém popisu češtiny tradici (na rozdíl od jazyků s méně bohatou morfologií, jako je např. angličtina), dostupné studie a projekty jdoucí tímto směrem (např. Bednaříková, 2009; Osolsobě, 2014; Pala — Šmerk, 2015) ukazují, že z propojování morfologie flektivní a derivační může mít prospěch lingvistický popis i automatické zpracování češtiny (srov. odd. 2).

V příspěvku představíme lexikální síť DeriNet, která obsahuje 969 tisíc českých slov, mezi nimiž jsou identifikovány dvojice slov základových a odvozených, ty jsou spojeny derivačním vztahem. DeriNet je — pro češtinu, ale i v kontextu jiných jazyků[2] — jed-

1 Tento příspěvek vznikl za podpory projektu GAČR 16–18177S. Při práci byly využívány jazykové zdroje vyvinuté, uložené nebo distribuované v rámci projektu LINDAT/CLARIN MŠMT ČR (projekt LM2015071).

2 Srov. CELEX pro angličtinu, němčinu a holandštinu (Baayen, 1995), DerivBase pro němčinu (Zeller a kol., 2013), DerivBase.Hr pro chorvatštinu (Šnajder, 2014), Démonette pro francouzštinu (Hathout — Namer, 2014) nebo jazykově nezávislý přístup (Baranes — Sagot, 2014).

ním z mála velkých, pro vědecké účely volně dostupných jazykových zdrojů specializovaných na derivační morfologii. Díky úzkému provázání s velkým slovníkem české flektivní morfologie (MorfFlex CZ; Hajič — Hlaváčová, 2013) umožňuje propojovat výzkum derivačních aspektů české morfologie s aspekty flektivními.

Výchozí lingvistická rozhodnutí o architektuře sítě DeriNet jsou prezentována v odd. 3, odd. 4 popisuje jednotlivé fáze její výstavby: nejprve byla se základovými slovy spojena frekventovaná, pravidelně utvořená slova, méně početné skupiny derivátů včetně derivátů s hláskovými alternacemi byly zpracovávány v dalších krocích. Data sítě DeriNet jsou k dispozici v několika verzích, pro prohlížení dat a vyhledávání v nich byla vyvinuta dvě webová rozhraní (odd. 5). Problémy, které zpracování velkého množství jazykových dat v souladu s široce akceptovaným dokulilovským přístupem k české derivaci přináší, jsou diskutovány v odd. 6.

2. DERIVACE V TEORETICKÉM POPISU A V AUTOMATICKÉM ZPRACOVÁNÍ ČEŠTINY

2.1 DERIVACE V ČEŠTINĚ

Derivace je v češtině nejčastějším a nejvíce produktivním způsobem tvoření slov; dalšími slovotvornými procesy (konverzí, skládáním, reflexivizací sloves; Dokulil a kol., 1986) vznikla menší část lexikonu češtiny. Slovo odvozené je se slovem základovým spojeno po stránce formální (vztahem fundace) i významové (vztahem motivace; Dokulil, 1962, s. 11–14).

Čeština disponuje velkým množstvím derivačních formantů (převážně přípon), mnohé z nich jsou homonymní (např. přípona *-ka* se uplatňuje při přechylování, zdrobňování a tvoření názvů prostředků činnosti nebo dějových substantiv; *učitel >učitelka*,[3] *skříň > skříňka, mýt > myčka, donášet > donáška*) a zároveň řada slovotvorných významů je vyjadřována hned několika formanty (srov. přípony uplatňující se při přechylování: *učitel > učitelka, ministr > ministryně, jeřábník > jeřábnice, král > královna, princ > princezna* ad.). Pro český systém odvozovacích formantů je také příznačná komplikovaná kombinatorika, řada formantů se spojuje se základovými slovy hned několika slovních druhů, srov. *houba > houbař, tesat > tesař; ryba > rybina, pevný > pevnina, třetí > třetina, vidět > vidina*. Odvozování je navíc doprovázeno hláskovými alternacemi, vkládáním či vypouštěním hlásek, v psané podobě pak také změnou velkého počátečního písmena na malé (př. *sníh > sněžit, list > lístek, žít > žití, léčba > léčebna, Vimperk > vimperský*).

2.2 POPIS DERIVACE V LINGVISTICKÝCH PRACÍCH

Komplexní teoretický popis české derivace podal Miloš Dokulil (1962). Podle strukturace mimojazykového obsahu v jazyce vymezil Dokulil čtyři onomaziologické kategorie, a to kategorii substanční (typicky odpovídající slovnímu druhu substan-

3 Znakem > reprezentujeme vztah odvození (derivace) mezi slovem základovým (nalevo od znaku) a slovem odvozeným (napravo).

tiv), kategorii vlastností (odpovídající adjektivům), kategorii příznaků konstantních (korespondující s adverbii) a příznaků dynamických (odpovídající slovesům). Uvnitř těchto kategorií a přes jejich hranice dochází ke třem typům posunů: (i) transpozice je chápána jako změna onomaziologické kategorie, která není doprovázena změnou ve významu (př. z adjektiva *líný* je transpozicí odvozeno substantivum *lenost*), (ii) modifikace je definována jako přidání příznaku, a to v rámci dané onomaziologické kategorie (př. *židle* > *židlička*), (iii) mutace je chápána jako změna významu při přechodu do jiné onomaziologické kategorie nebo v jejím rámci (př. *bloudit* > *bludiště*, *galerie* > *galerista*).

Onomaziologické kategorie jsou dále děleny na slovotvorné kategorie zahrnující slova utvořená od téhož slovního druhu a mající stejný slovotvorný význam; srov. početnou kategorii deadjektivních substantiv s významem kvality. Slovotvorné kategorie se podle použitého formantu rozpadají na jednotlivé slovotvorné typy, např. slovotvorný typ deadjektivních substantiv s významem kvality končících na *-ost* vs. obdobný slovotvorný typ s příponou *-ství*.

Dokulilův přístup se stal široce respektovaným a v zásadě jediným východiskem popisu odvozování v češtině v následujících desetiletích, popis odvozovacích formantů je ve specializovaných publikacích i v souhrnných mluvnických pracích většinou organizován právě podle slovotvorných typů (např. Daneš a kol., 1967; Šmilauer, 1971; Hauser, 1986; Dokulil a kol., 1986; Grepl a kol., 2000; Čermák, 2012). Pokud jde o vztah odvozování a flektivní morfologie, je odvozování v mluvnicích češtiny tradičně popisováno v rámci slovotvorných výkladů, které jsou s popisem flektivní morfologie propojeny jen sporadicky (kromě citovaných mluvnic srov. i další příručky z posledních dvou dekád, např. Čechová a kol., 1996; Cvrček a kol., 2010; Štícha a kol., 2013).

Kořeny této separace sleduje Bednaříková (2009, s. 24) a konstatuje, že v akademické Mluvnici češtiny (Dokulil a kol., 1986, Komárek a kol., 1986) je ve slovotvorbě „spatřována jakási přechodová oblast mezi morfologií a slovníkem" a „[s]lovotvorné prostředky jsou považovány za prostředky nemorfologické, stejně jako např. prostředky lexikální, slovosledné, intonační a jiné". Kromě Bednaříkové (2009) nabízí integrující pohled na českou derivační a flektivní morfologii např. Osolsobě (2014).

2.3 DERIVAČNÍ MORFOLOGIE V AUTOMATICKÉM ZPRACOVÁNÍ ČEŠTINY

Separace flektivní a derivační morfologie, jak se s ní setkáváme v mluvnických pracích, byla překonána také v některých projektech v oblasti automatického zpracování češtiny. Nejkomplexnějším — vedle zde prezentované sítě DeriNet — je projekt Derivancze, který je modifikací derivační verze morfologického analyzátoru Ajka (Sedláček — Smrž, 2001; Pala — Šmerk, 2015). Derivancze je webový nástroj, který pro slovo zadané do webového formuláře uvede slovo základové a slovo bezprostředně odvozené (popř. více bezprostředních derivátů); kompletní slovotvornou čeleď je možné získat postupným dotazováním po jednotlivých dvojicích. V datech, se kterými tento nástroj pracuje, je zachyceno více než 255 tisíc derivačních vztahů. Vztahy jsou sémanticky anotovány sadou celkem 17 značek (v síti DeriNet sémantické značky dosud chybí, srov. odd. 6).

Ve slovníku morfologického analyzátoru Ajka je navíc možné vyhledávat dvojice (nebo n-tice) slov odvozených a základových pomocí nástroje Deriv (Osolsobě, 2009), a to na základě pravidel formulovaných pomocí regulárních výrazů spíše než grama-

tických rysů. Obdobným nástrojem, ovšem pracujícím nad daty Českého národního korpusu, je nástroj Morfio (Cvrček — Vondřička, 2013), který vyhledává páry (n-tice) slov se shodným základovým řetězcem a různými formanty. Tento webový nástroj umožňuje zahrnout do dotazů nejběžnější hláskové alternace, ty ovšem nelze podmínit hláskovým okolím, což může vést k vysokému podílu nerelevantních výsledků.

Dílčí derivační informace je k dispozici také ve velkém slovníku české flektivní morfologie MorfFlex CZ (Hajič — Hlaváčová, 2013): pro pravidelně utvořené, vysoce frekventované typy derivátů je zde uvedena informace o jejich základovém slově, a to jako součást morfologického lemmatu (jako jedna z jeho tzv. technických přípon, Hajič, 2004).[4]

Derivační vztahy jsou zahrnuty také v dalších zdrojích, většinou ovšem pouze okrajově. Např. v datech lexikální databáze Czech WordNet byla implementována sada 14 derivačních vztahů (Pala — Smrž, 2004; Pala — Hlaváčková, 2007). V Pražském závislostním korpusu 2.0 (Hajič a kol., 2006) byly v rámci hloubkově syntaktické anotace vybrané typy derivátů převáděny na jejich základové slovo (Razímová — Žabokrtský, 2006).

3. DESIGN DERIVAČNÍ SÍTĚ — VÝCHOZÍ TEORETICKÁ ROZHODNUTÍ

Síť DeriNet byla navržena jako elektronická databáze českých plnovýznamových slov (substantiv, adjektiv, adverbií a sloves),[5] která jsou spojována odkazy odpovídajícími slovotvornému vztahu derivace.

Vztahy mezi slovem základovým a slovy odvozenými (vztahy fundačně-motivační) jsou modelovány jako orientovaný graf. Uzly grafu odpovídají slovům (lemmatům), hrany grafu jednotlivým krokům odvozovacího procesu. Orientace hran odráží směr odvozování: hrany odkazují od slova základového ke slovu odvozenému. Každý derivát může být spojen nanejvýš s jedním slovem základovým. Slova spojená derivačními vztahy tak vytvářejí stromový graf, který — v ideálním případě — reprezentuje celou slovotvornou čeleď a jehož kořenem je základové slovo nemotivované.

Budování sítě DeriNet bylo zahájeno výběrem sady lemmat. Původní sada lemmat byla extrahována z dat velkých korpusů češtiny (srov. odd. 4.1), čímž bylo zajištěno, že síť obsahovala pouze lemmata doložená v autentických českých textech. Zvažované alternativní řešení poloautomatického generování lemmat by oproti tomu vedlo k zařazení množství lemmat, která jsou sice správně utvořená, nejsou ovšem součástí úzu.

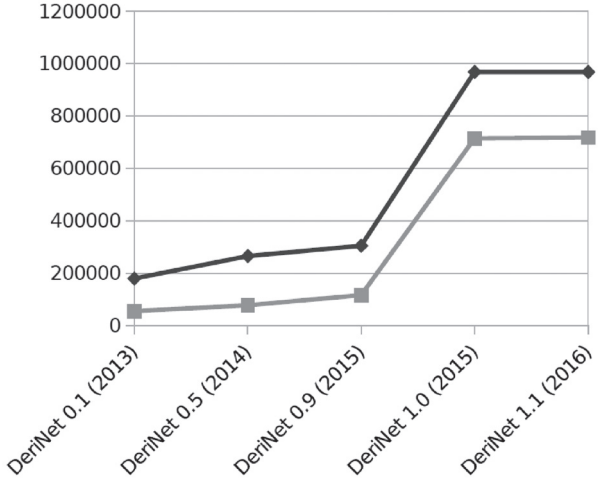
4 Jako technická přípona lemmatu je v kontextu automatické morfologické analýzy češtiny označován řetězec, ve kterém je formálně zaznamenána derivační nebo jiná informace vztahující se k danému slovu. Tento řetězec se speciálním znakem (podtržítkem) připojuje k základnímu tvaru slova. Srov. lemma *vaření_^(*3it)*, z jehož technické přípony je možné stanovit základové slovo tohoto lemmatu (tj. odstraněním tří posledních znaků a připojením řetězce *it* dospějeme ke slovesu *vařit*; viz zde odd. 4.1).

5 Stupňování adjektiv a adverbií není v síti DeriNet zachycováno jako proces slovotvorný, je chápáno jako součást morfologie flektivní. Slova odvozená od komparativu (př. *zlepšit*) tak v síti DeriNet mohou být spojena s příslušným adjektivem reprezentovaným základním tvarem pozitivu.

V této sadě slov jsme chtěli identifikovat co největší množství derivačních vztahů. Usilovali jsme přitom o to, aby zachycení vztahů bylo lingvisticky adekvátní, konzistentní a také ekonomické. Při vytváření derivačních vztahů proto byly využity dostupné zdroje jazykových dat (z nichž nejvýznamnější byl slovník MorfFlex CZ) a proces byl automatizován. Protože jsme přesnost zachycovaných vztahů (jejich lingvistickou správnost) považovali za důležitější aspekt než jejich pokrytí (množství vztahů), automaticky navržené derivační vztahy před zanesením do dat vždy procházely manuální kontrolou.

4. BUDOVÁNÍ DERIVAČNÍ SÍTĚ DERINET

V této sekci představíme postupné změny, kterými lexikální síť DeriNet prošla od svého založení v roce 2013 do současnosti: síť DeriNet se vyvinula ze zdroje zaměřeného na specifické typy deadjektivních derivátů v aktuálně největší existující (i když v mnoha aspektech stále ještě rozpracovanou) veřejně dostupnou databázi derivačních vztahů v češtině; viz obr. 1. Přesná chronologie budování sítě není pro výklad zásadní, použijeme ji jen jako rámec pro představení různých metod, kterými byla síť postupně naplňována.



OBRÁZEK 1. Vývoj sítě DeriNet: horní křivka zachycuje počet lemmat v síti, spodní křivka odpovídá počtu derivačních vztahů

4.1 ZALOŽENÍ SÍTĚ, AUTOMATICKY VYVOZENÁ PRAVIDLA PRO IDENTIFIKACI DERIVAČNÍCH VZTAHŮ (DERINET VERZE 0.1, 2013)

První, pilotní verze sítě DeriNet vznikla v souvislosti s experimenty s deadjektivními substantivy odvozenými příponou *-ost*. Základem sítě se stala lemmata čtyř plnovýznamových slovních druhů, která se vyskytovala v české části paralelního korpusu

CzEng 1.0 (Bojar a kol., 2012). Tato množina byla rozšířena o substantivní lemmata končící řetězcem *-ost*, která byla nalezena v korpusu SYN2005 (Čermák a kol., 2005).

Derivační vztahy mezi základovými adjektivy a substantivy odvozenými touto příponou byly identifikovány pomocí heuristických pravidel, která byla ručně sestavena jako regulární výrazy nahrazující příponu lemmatu a doplněna o seznam výjimek; srov. pravidlo *A-ý > N-ost*[6] identifikující dvojici adjektiva končícího na *-ý* a substantiva s totožnou hláskovou skladbou, ovšem končícího místo *-ý* řetězcem *-ost* (př. *krutý > krutost, závislý > závislost, dokonalý > dokonalost*). Pouze několik obdobných pravidel dostačovalo pro spojení deadjektivních adverbií s jejich základovými adjektivy (*A-ý >D-ě: krutý > krutě*; *A-ý >D-y: přátelský > přátelsky*; *A-í >D-ě: revoluční > revolučně* ad.).

Následující rozhodnutí zásadně rozšířit repertoár zachycovaných derivačních vztahů (s ambicí zachytit co nejvíce těchto vztahů) bylo spojeno s prvním pokusem o výraznější automatizaci procesu identifikace dvojic slov základových a odvozených. Pro dvojice lemmat s dostatečně dlouhým shodným počátečním podřetězcem proto byly extrahovány koncové podřetězce, kterými se tyto dvojice lišily (zčásti odpovídaly příponám, zčásti se jednalo o řetězce delší). Ze 400 nejčastějších takto získaných dvojic bylo ručně vybráno 18, které odpovídaly frekventovaným derivačním vztahům v české slovní zásobě, u těchto dvojic byl navíc označen směr derivace (ve většině případů nicméně platilo, že koncový podřetězec základového lemmatu je kratší než u lemmatu odvozeného); srov. *N-a >N-ka, A-ý >N-ec, V-t >N-č, V- >N-el*. Extrahovaná pravidla byla aplikována na všechna relevantní lemmata v síti, chybné dvojice byly vyloučeny ručně. Tímto postupem bylo nalezeno zhruba 11 tisíc derivačních vztahů.

Další metodou použitou pro určování derivačních vztahů bylo využití technických přípon lemmat dostupných ve slovníku MorfFlex CZ; např. z technické přípony lemmatu *vaření_^(*3it)* lze odstraněním tří posledních znaků a připojením řetězce *it* stanovit základové sloveso (*vařit*; popis technických přípon viz Hana a kol., 2005). I když je derivační informace ve slovníku MorfFlex CZ dostupná jen pro část lemmat (slovník je zaměřen na flektivní analýzu), přínos tohoto zdroje pro vytváření vztahů v síti DeriNet byl zásadní.

Ve verzi 0.1 tak síť DeriNet obsahovala 180 tisíc lemmat (uzlů) propojených 56 tisíci derivačními vztahy (hranami). Data byla zveřejněna v roce 2013.

4.2 SUBSTITUČNÍ PRAVIDLA SESTAVENÁ NA ZÁKLADĚ MLUVNICKÉHO POPISU, ZAHRNUTÍ FREKVENTOVANÝCH HLÁSKOVÝCH ZMĚN (DERINET 0.5, 2014; DERINET 0.9, 2015)

V následující verzi dat (DeriNet 0.5) byla sada lemmat rozšířena o všechna lemmata čtyř autosémantických slovních druhů, která se v korpusu SYN (Křen a kol., 2014) vyskytovala alespoň dvakrát, zároveň neobsahovala číslici ani interpunkční znaménko

6 Uvedené pravidlo se interpretuje tak, že z adjektiva končícího na *-ý* (*A-ý*) vzniká náhradou koncového *-ý* za *-ost* substantivum končící řetězcem *-ost* (*N-ost*). V dalších pravidlech pracujeme také se značkami *D* (zastupuje adverbium) a *V* (zastupuje sloveso). Pokud za spojovníkem nenásleduje žádný řetězec (př. *V-* v pravidle *V- >N-el*), při derivaci nedochází k náhradě koncových řetězců, ale k základovému slovu jako celku se připojuje nový řetězec.

a skládala se nejméně ze dvou písmen, z nichž aspoň jedno bylo malé (s cílem vyhnout se zkratkám). Z výše zmíněných 400 automaticky nalezených pravidel pro záměny sufixů bylo vybráno a doplněno o seznamy výjimek dalších 17 spolehlivých pravidel.

Významným zdrojem informací pro generování derivačních relací byl také seznam substitučních pravidel sestavený na základě *Příruční mluvnice češtiny* (Grepl a kol., 2000). Po odstranění duplicitních pravidel, která se do seznamu dostala proto, že formanty jsou v mluvnici uspořádány podle slovotvorných typů a formálně totožná dvojice základového a odvozeného slova zde může být uvedena hned u několika typů (srov. pravidlo *V-t > N-č* postihující tvoření činitelských jmen typu *holič* i jmen prostředků činnosti jako *vypínač*; Grepl a kol., 2000, s. 141 a 144), jsme získali zhruba 450 pravidel. Ve srovnání s automaticky extrahovanými pravidly, se kterými jsme pracovali v předchozí fázi, tato pravidla postihovala méně frekventované slovotvorné typy (např. *V-it > N-ba: léčit > léčba*; *A-ý > N-oba: starý > staroba*), do některých pravidel jsme navíc zahrnuli i vysoce frekventované hláskové změny, například vkládání hlásek, jejich vypouštění, krácení nebo dloužení a palatalizaci (vedle pravidla *N- > N-ka: učitel > učitelka* jsme zařadili i pravidlo *N-c > N-čka: herec > herečka*; vedle *N- > A-ový: achát > achátový* také *N-ec > A-cový: konec > koncový*). Pravidla byla využita obdobně jako automaticky extrahovaná pravidla.

Vedle nově formulovaných pravidel zahrnujících hláskové změny jsme v této fázi opakovaně aplikovali i pravidla z fáze předchozí, při jejich aplikaci jsme ovšem vyhledávali i dvojice s hláskovými změnami; celkem jsme povolili 18 typů samohláskových změn (př. *í — ě* jako při derivaci *sníh > sněžný*, *ů — o* v *dům > domek*) a 11 typů souhláskových změn (*ch — š, h — ž, g — ž* ad.). Tato procedura vedla — podobně jako v případě automaticky extrahovaných pravidel — ke generování nadměrného množství dvojic slov základových a odvozených, i v tomto případě tedy byly z výstupu ručně vyloučeny nesprávné dvojice. Další dvojice slov základových a odvozených byly nalezeny pomocí pravidel pro odvozování slov předponami latinského nebo řeckého původu.

Kromě přidávání lemmat a derivačních vztahů byla data DeriNet 0.5 první verzí, kde dostatečná hustota derivačních vztahů již umožňovala systematičtější testování celých derivačních shluků (celých derivačních stromů odpovídajících slovotvorným čeledím nebo jejich částí). Byla například použita heuristika, že pokud dva různé derivační shluky obsahují dvě množiny lemmat, které jsou stejně velké a současně jsou totožné v tom, jak se jednotlivá lemmata liší od kořene daného shluku, potom by oba derivační shluky měly mít stejný tvar stromu (tímto postupem bylo restrukturováno 760 shluků).

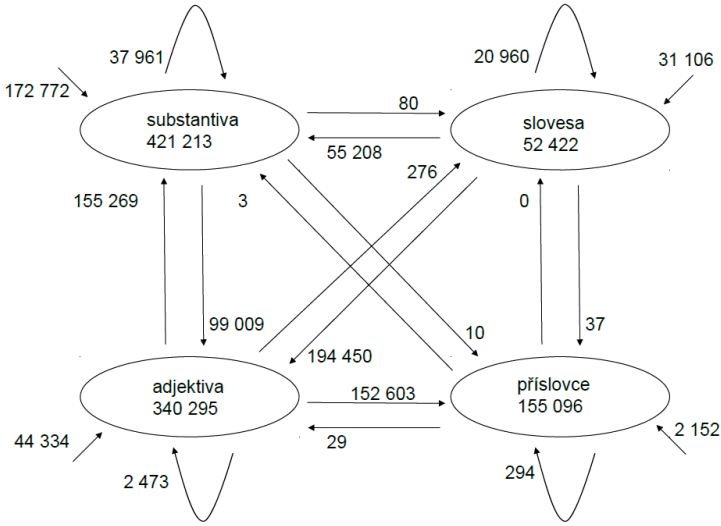
Verze DeriNet 0.5 byla zveřejněna v květnu 2014 a obsahovala 266 tisíc lemmat a 79 tisíc derivačních vztahů (Ševčíková — Žabokrtský, 2014).

Zlepšení v datech následující verze (DeriNet 0.9, 2015) nevyplývala z použití nových lingvisticky relevantních metod, ale spíše z revize stávajících modulů a řady drobnějších technických oprav. Verze DeriNet 0.9 obsahovala 306 tisíc lemmat a 117 tisíc derivačních vztahů.

4.3 ROZŠÍŘENÍ SADY LEMMAT (DERINET 1.0, 2015)

Ve verzi 1.0 prošel DeriNet radikální změnou: volba repertoáru lemmat již nebyla dána kritériem korpusové četnosti, ale do sítě byla zařazena všechna lemmata obsažená ve slovníku MorfFlex CZ, který v této době již byl k dispozici pod otevřenou

 To vedlo ke zhruba ztrojnásobení sítě co do počtu lemmat a ještě většímu růstu, pokud jde o počet derivačních vztahů.



OBRÁZEK 2. Základní kvantitativní vlastnosti sítě DeriNet ve verzi 1.1

Toto výrazné zvětšení sítě DeriNet si — na první pohled paradoxně — nevyžádalo zásadní objem ruční práce, většina nových derivačních vztahů byla rozpoznána existujícími automatickými metodami (Vidra, 2015). Průběžně prováděná měření kvality dat ukázala, že průměrná správnost derivačních hran dokonce vzrostla. Důvod by bylo možné hledat v principu jazykové ekonomie: s klesající četností slova v korpusu roste pravděpodobnost, že jde o derivát (pro mluvčí flektivního jazyka zřejmě není ekonomické pamatovat si velké množství kořenů) a že tento derivát je navíc utvořen pravidelně (nutnost pamatovat si nepravidelně utvořené deriváty by se v případě řídkých slov jevila také jako neekonomická).

Síť DeriNet ve verzi 1.0 byla zveřejněna v říjnu 2015 a obsahovala 969 tisíc lemmat propojených 716 tisíci derivačních vztahů.

4.4 ZPRACOVÁNÍ HLÁSKOVÝCH ZMĚN NA ZÁKLADĚ JEJICH IDENTIFIKACE VE FLEKTIVNÍCH PARADIGMATECH (DERINET 1.1, 2016)

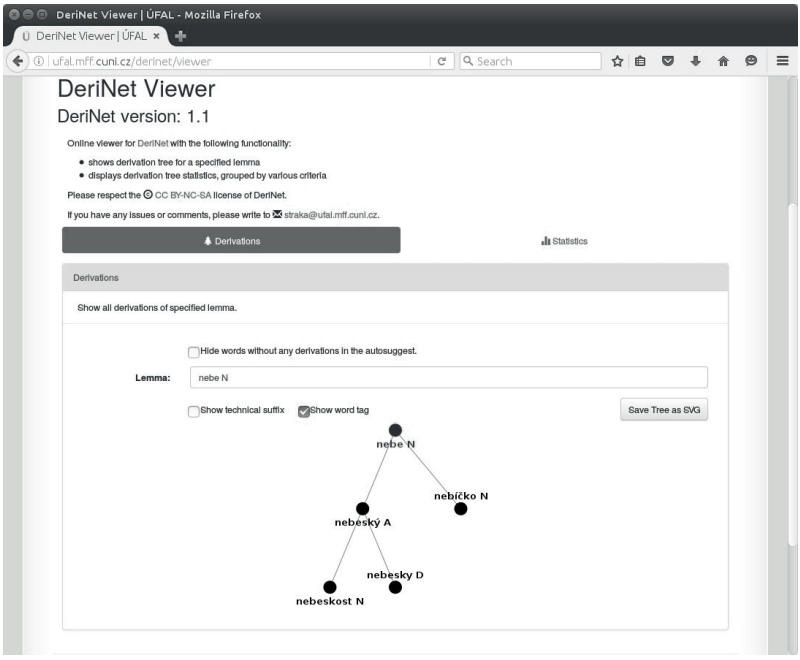
V některých derivátech nastávají oproti základovému lemmatu hláskové změny, které je velmi obtížné postihnout obecněji, neboť neprobíhají jenom na švu kořene a derivačního formantu, ale zasahují hlouběji do kořene. Při další práci na datech sítě DeriNet jsme využili pozorování, že tyto změny v některých případech korelují s hláskovými změnami probíhajícími při skloňování nebo časování základového slova. Například změna samohlásky *ů* v *o* a změna souhlásky *h* v *ž* v jednotlivých tvarech substantiva *bůh* (*boha, bože* atd.) se opakuje při adjektivní derivaci (*bůh* > *boží*). Sub-

stituční pravidla vyvinutá v předcházejících verzích byla tedy aplikována i na takto alterovaná vstupní lemmata, po ručním vyloučení chyb byly nalezeny zhruba dva tisíce derivací, které se předchozími postupy nepodařilo identifikovat.

Aktuální verze dat, DeriNet 1.1, obsahuje 969 tisíc lemmat a 718 tisíc hran. Podrobnější kvantitativní vlastnosti verze 1.1 jsou zachyceny v diagramu na obrázku 2. V uzlech jsou uvedeny absolutní četnosti lemmat daného slovního druhu, u šipek mezi uzly jsou uvedeny počty dvojice se základovým a odvozeným lemmatem v daných slovních druzích. Šipky směřující do téhož slovního druhu, odkud vycházejí, znázorňují derivaci v rámci daného slovního druhu (tedy např. substantiva odvozená od substantiv). U šipek vedoucích z vně diagramu je uveden počet lemmata, která jsou v síti DeriNet 1.1 zachycena jako kořeny derivačních stromů, jako základová slova nemotivovaná.

4.5 PLÁNY NA ROZŠIŘOVÁNÍ SÍTĚ DERINET V NEJBLIŽŠÍ BUDOUCNOSTI

Pokud jde o budoucí růst sítě DeriNet, množství pokrytých lemmat považujeme již za dostačující, jakkoli i v oblasti repertoáru lemmat zbývá lépe zpracovat dvě lingvisticky i technicky obtížná témata, a sice homonymii a pravopisné varianty (srov. odd. 6). Většinu úsilí tak věnujeme hledání nových derivačních vztahů mezi lemmaty, popřípadě opravám vztahů stávajících. V nejbližší době se jako vhodná a zvládnutelná témata nabízí propojení vidových dvojic, například s využitím valenčního slovníku VALLEX (Lopatková a kol., 2016), a zachycení prefixace, zejména slovesné. Zatím nerozpoznané derivační vztahy se také chystáme vyhledávat s využitím metod strojového učení.

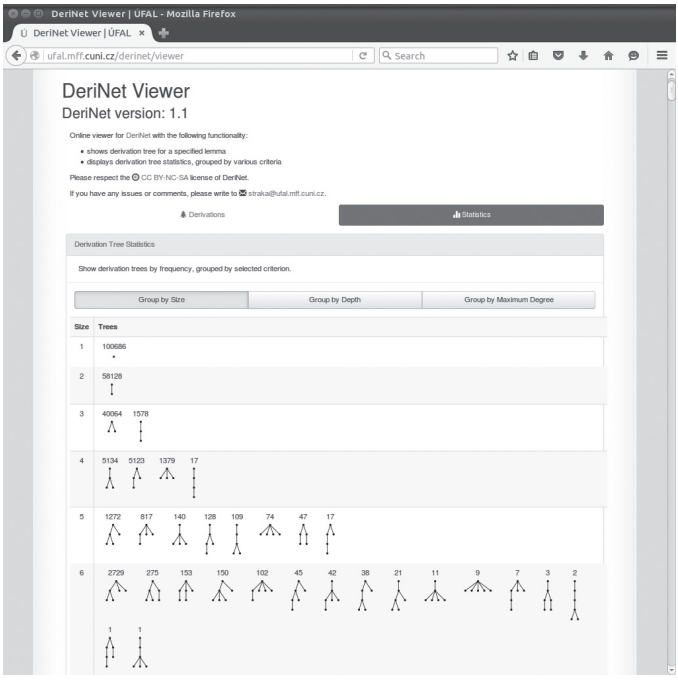


OBRÁZEK 3. DeriNet Viewer — zobrazení derivačního stromu substantiva *nebe*

5 DOSTUPNOST DAT SÍTĚ DERINET, UŽIVATELSKÁ ROZHRANÍ

5.1 DOSTUPNOST DAT

Jednotlivé verze sítě DeriNet jsou dostupné z webového rozcestníku projektu na adrese http://ufal.mff.cuni.cz/derinet. Významné verze sítě jsou a budou publikovány prostřednictvím repozitáře Lindat/Clarin (zatím jde jen o verzi 1.0 dostupnou na adrese http://hdl.handle.net/11234/1-1520; Vidra a kol., 2015). Pro nekomerční účely jsou všechny dosavadní verze dat k dispozici zdarma, podmínky užívání jsou upravené licencí Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License (CC-BY-NC-SA). Data jednotlivých verzí jsou ke stažení vždy ve dvou jednoduchých souborových formátech (tsv a xml). Data v této podobě se dají snadno zpracovávat běžnými programátorskými prostředky, nejsou ovšem vhodná pro lingvistickou práci (srov. odd. 5.2).



OBRÁZEK 4: DeriNet Viewer — četnost derivačních stromů podle jejich tvaru

5.2 UŽIVATELSKÁ ROZHRANÍ DERINET VIEWER A DERINET SEARCH

Pro běžné vyhledávání v datech sítě DeriNet byla vyvinuta dvě grafická uživatelská rozhraní realizovaná formou webové aplikace. Jednodušší a historicky starší je prohlížeč DeriNet Viewer (autora Milana Straky; http://ufal.mff.cuni.cz/derinet/viewer). Základní funkcí je zobrazení derivačního stromu pro zadané lemma (obr. 3)

s možností exportovat obrázek do formátu SVG. Vedle toho tento nástroj umožňuje také shlukovat derivační stromy podle podobnosti jejich tvaru (obr. 4) a zobrazit některé další statistické vlastnosti sítě.

Druhým nástrojem je DeriNet Search (vytvořený Jonášem Vidrou), dostupný na adrese http://ufal.mff.cuni.cz/derinet/search. Uživatel zde má k dispozici jednoduchý dotazovací jazyk, kterým může specifikovat vlastnosti hledaného uzlu (např. jeho slovní druh nebo podřetězec lemmatu) a pomocí jednoduché závorkové notace vyjádřit strukturní omezení pro tvar derivačního stromu. Např. dotazem *[pos="A"] ([pos="N" lemma="tví$"], [pos="N" lemma="ismus$"])* se vyhledají derivační stromy, ve kterých je od adjektiva přímo odvozeno jak substantivum končící řetězcem *-ství*, tak substantivum na *-ismus*; srov. obr. 5.

6. OTEVŘENÉ OTÁZKY

Od počátku budování sítě DeriNet je zřejmé, že lingvisticky adekvátní reprezentace derivačních vztahů se neobejde bez konzistentního přístupu k případům asymetrie formy a významu ve slovní zásobě, zvláště k polysémii a homonymii. Polysémní lemmata (lexémy) reprezentujeme v síti jediným uzlem, všechny jejich deriváty jsou pak shromážděny v jednom derivačním stromě. Oproti tomu homonyma pocházejí z různých slov základových a mívají odlišnou sadu derivátů, v síti je proto chceme reprezentovat různými uzly.

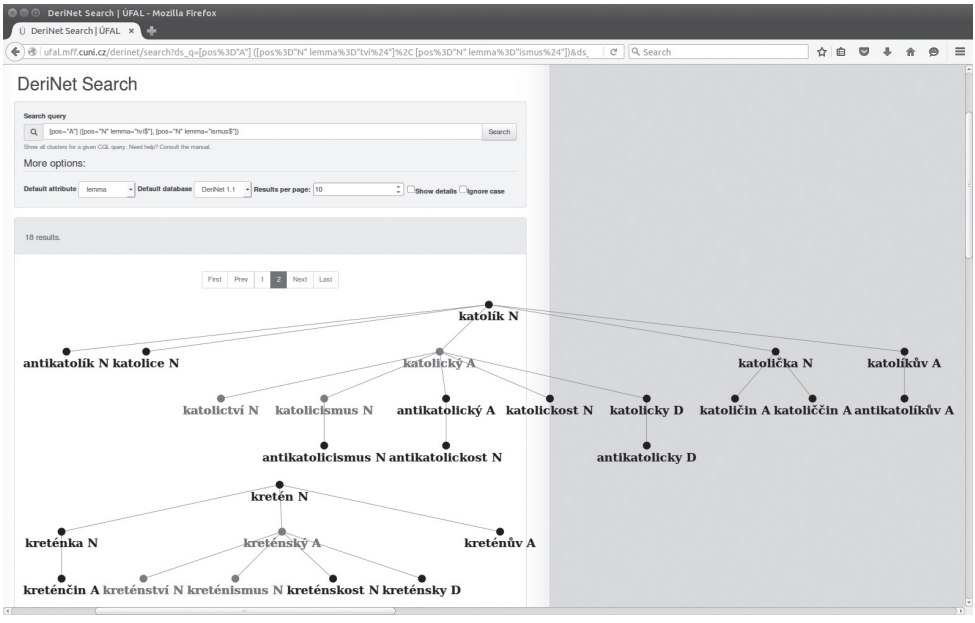
Naším cílem je brát při zpracování homonymie v úvahu pouze aspekty relevantní z hlediska derivace a flexe (v souvislosti s propojením dat sítě s daty slovníku MorfFlex CZ; srov. odd. 4.3). Přístup k homonymům implementovaný ve slovníku MorfFlex CZ se ovšem ukázal jako příliš široký, často bez jasné hranice k polysémii. Abychom zajistili kompatibilitu dat sítě DeriNet a slovníku MorfFlex CZ i z hlediska homonymie, lemmata slovníku prošla nejprve automatickou revizí, následná, právě probíhající ruční revize má za cíl vymezit jádro (maximálně několika set n-tic) homonym lišících se svými flektivními a derivačními vlastnostmi. Mezi tato homonyma budou patřit:

a) formálně identické lexémy tvoření různými morfémy, srov. sloveso *proudit* utvořené buď od substantiva *proud* (*proud-it*), nebo od slovesa *udit* (*pro-udit*),
b) lexémy odvozené od různých základů formanty se širokou sémantikou, srov. adjektivní přípony *-ový* a *-ný* vyjadřující široký vztah k substantivu (př. adjektivum *masový* se vztahuje k substantivu *masa* nebo *maso*, *vinný* k substantivu *vina* nebo *víno*),
c) lexémy odvozené od stejného základu homonymním formantem, např. přípona *-ič* se používá k odvozování činitelských jmen i jmen prostředků činnosti, přípona *-ka* k přechylování i zdrobňování (př. *dlaždič* > *dlaždička*, *dlaždice* > *dlaždička*).

Otázka korespondence mezi uzly sítě a lemmaty zůstává kromě homonymie otevřená také při zachycování pravopisných (grafických) variant nejrůznějšího typu, např. *nakladač/nakládač*, *citron/citrón*, *diskuze/diskuse*. Preferované řešení — shromáždit tyto varianty vždy do jediného uzlu — bude konfrontováno s dalšími možnostmi a výsledná reprezentace bude implementována v nejbližších měsících.

Lingvisticky nejednoznačný je také způsob zachycování skupin derivátů s negačním prefixem, popř. dalšími prefixy. Např. substantivum *nemačkavost* lze zachytit jako derivát substantiva *mačkavost* nebo adjektiva *nemačkavý*, při zachycení obou vztahů by pak vznikl cyklus, který není v derivačním stromě přípustný; obdobně srov. derivační řetězce *biologie* > *biologický* >*biologicky* a *mikrobiologie* > *mikrobiologický* >*mikrobiologicky* — slova druhého řetězce lze také považovat za deriváty jednotlivých členů řetězce prvního. Tyto typy případů je třeba důkladně vymezit a jejich reprezentaci v datech sjednotit.

Zmíníme zde i další aktuálně zpracovávaná témata. Při plánovaném sémantickém značkování derivačních vztahů vycházíme ze základní sady zhruba 10 značek (činitelská jména, přechýlená substantiva, deminutiva, posesiva ad.), značkování bude prováděno z velké části automaticky (s následnou manuální kontrolou), a to na základě ručně vytvořených pravidel využívajících hláskovou podobu lemmat nebo jejich podřetězců (zvl. přípon), jejich slovnědruhové charakteristiky a morfologických charakteristik (zvl. rodu) a také struktury derivačních stromů. Např. substantiva na *-ka* utvořená od maskulin (př. *učitel* > *učitelka*) budou značkována jako přechýlená, substantiva s tímtéž zakončením, ovšem odvozená od feminina (*skříň* > *skříňka*) budou hodnocena jako deminutiva, substantiva na *-ka* odvozená od sloves pak jako jména prostředků činnosti nebo jako dějová substantiva (*brousit* > *bruska*, *donášet* > *donáška*). Možnost prohlížet data sítě DeriNet podle vyznačených sémantických vztahů významně rozšíří jejich využitelnost. Dalším úkolem je pak spojení sítě DeriNet se slovníkem MorfFlex CZ do jediného uživatelského rozhraní.



OBRÁZEK 5: Výsledky dotazu *[pos="A"] ([pos="N" lemma="tví$"], [pos="N" lemma="ismus$"])* ve vyhledávači DeriNet Search

Nejzásadnější zásah do datové reprezentace sítě DeriNet plánujeme v blízké budoucnosti v souvislosti s rozšířením sítě o další slovotvorné způsoby, tedy kompozici a derivačně-kompoziční tvoření slov. Tímto krokem opustíme úvodní omezení sítě DeriNet pouze na vztah derivace, data pak budou podávat adekvátnější obraz české slovotvorby. Omezení týkající se zařazení pouze čtyř plnovýznamových slovních druhů, v jehož důsledku nemohou být v síti se svým základovým slovem spojeny deriváty číslovek, zájmen, předložek a dalších slovních druhů (např. substantiva *dvojče, třetina* nebo spřežková slova jako *zčásti*), tímto krokem odstraněno nebude. Při revizi tohoto omezení je třeba zvážit důsledky zařazení každé další slovnědruhové třídy zvlášť; např. ve slovním druhu číslovek jsou základovými slovy především číslovky označující nižší hodnoty (př. *trojčata, pětina, zdvojit*) a vybrané hodnoty vyšší (př. *tisícina*), z hlediska počtu lemmat se ale jedná o třídu otevřenou.

7. ZÁVĚR

V příspěvku jsme představili lexikální síť DeriNet, která je elektronickým zdrojem jazykových dat specializovaným na zachycování derivačních vztahů v české slovní zásobě. I přes stále otevřená, lingvisticky závažná témata, na jejichž adekvátním řešení pracujeme, je DeriNet v současné chvíli největším, pro vědecké účely volně dostupným zdrojem jazykových dat pro výzkum derivace v češtině, a to jak v lingvistickém výzkumu, tak v oblasti automatického zpracování češtiny (např. při strojovém překladu).

Díky úzkému propojení s daty flektivního slovníku MorfFlex CZ otevírá síť DeriNet nový pohled na derivaci jako součást široce pojatého morfologického systému češtiny.

LITERATURA

Baayen, R. H. et al. (1995): *The CELEX lexical database* (release 2). Data/software. Philadelphia, PA: LDC.

Baranes, M. — Sagot, B. (2014): A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavík: ELRA, s. 2793–2799.

Bednaříková, B. (2009): *Slovo a jeho konverze*. Olomouc: UPOL.

Bojar, O. et al. (2012): The Joy of Parallelism with CzEng 1.0. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul: ELRA, s. 3921–3928.

Cvrček, V. a kol. (2010): *Mluvnice současné češtiny*. Praha: Karolinum.

Cvrček, V. — Vondřička. P. (2013): Nástroj pro slovotvornou analýzu jazykového korpusu. In: *Gramatika a korpus 2012*. Hradec Králové: Gaudeamus.

Čechová, M. a kol. (1996): *Čeština — řeč a jazyk*. Praha: ISV nakladatelství.

Čermák, F. (2012): *Morfématika a slovotvorba češtiny*. Praha: NLN.

Čermák, F. a kol. (2005): *SYN2005: žánrově vyvážený korpus psané češtiny*. Praha: ÚČNK FF UK. Dostupný z WWW: http://www.korpus.cz.

Daneš, F. a kol.(1967): *Tvoření slov v češtině 2: Odvozování podstatných jmen*. Praha: Nakladatelství ČSAV.

Dokulil, M. (1962): *Tvoření slov v češtině 1: Teorie odvozování slov*. Praha: Nakladatelství ČSAV.

Dokulil, M. a kol. (1986): *Mluvnice češtiny 1. Fonetika, fonologie, morfonologie a morfematika, tvoření slov*. Praha: Academia.

Grepl, M. a kol. (2000): *Příruční mluvnice češtiny*. Druhé, opravené vydání. Praha: NLN.

Hajič, J. (2004): *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Praha: Karolinum.

Hajič, J. et al. (2006): *Prague Dependency Treebank 2.0*. Data/software. Philadelphia, PA: LDC.

Hajič, J. — Hlaváčová, J. (2013): *MorfFlex CZ. Morfologický slovník češtiny*. LINDAT/CLARIN digital library at ÚFAL MFF UK, http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.

Hana, J. et al. (2005): *Manual for Morphological Annotation, Revision for the Prague Dependency Treebank 2.0*. Praha: ÚFAL MFF UK.

Hathout, N. — Namer, F. (2014): Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11, s. 125–168.

Hauser, P. (1986): *Nauka o slovní zásobě*. Druhé vydání. Praha: SPN.

Komárek, M. a kol. (1986): *Mluvnice češtiny 2. Tvarosloví*. Praha, Academia.

Křen, M. a kol. (2014): *Korpus SYN*, verze 3 z 27. 1. 2014. Praha: ÚČNK FF UK. Dostupný z WWW: http://www.korpus.cz

Lopatková, M. a kol. (2016): *Valenční slovník českých sloves*. Karolinum, Praha.

Osolsobě, K. (2009): Deriv — nástroj pro automatické vyhledávání slovotvorných vztahů. Slovotvorný stroj pro češtinu — sen nebo skutečnost? In: *Přednášky a besedy z XLII. běhu LŠSS*. Brno: FF MU, s. 132–137.

Osolsobě, K. (2014): *Česká morfologie a korpusy*. Praha: Karolinum.

Pala, K. — Hlaváčková, D. (2007): Derivational Relations in Czech WordNet. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*. Prague: ACL, s. 75–81.

Pala, K. — Smrž, P. (2004): Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7, s. 79–88.

Pala, K. — Šmerk, P. (2015): Derivancze — *Deriv*ational *Ana*lyzer of *C*zech. In: *Proceedings of Text, Speech and Dialogue 2015*. Berlin: Springer, s. 515–523.

Razímová, M. — Žabokrtský, Z. (2006): Annotation of Grammatemes in the Prague Dependency Treebank 2.0. In: *Proceedings of the LREC 2006 Workshop on Annotation Science*. Genoa: ELRA, s. 12–19.

Sedláček, R. — Smrž, P. (2001): A New Czech Morphological Analyzer ajka. In: *Proceedings of Text, Speech and Dialogue 2001*. Berlin: Springer, s. 100–107.

Ševčíková, M. — Žabokrtský, Z. (2014): Word-Formation Network for Czech. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavík: ELRA, s. 1087–1093.

Šmilauer, V. (1971): *Novočeské tvoření slov*. Praha: SPN.

Šnajder, J. (2014): DerivBase.Hr: A High-Coverage Derivational Morphology Resource for Croatian. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavík: ELRA, s. 3371–3377.

Štícha, F. a kol. (2013): *Akademická gramatika spisovné češtiny*. Praha: Academia.

Vidra, J. (2015): *Extending the Lexical Network DeriNet*. Bakalářská práce. Praha: MFF UK.

Vidra, J. et al. (2015): *DeriNet 1.0*. LINDAT/CLARIN digital library at ÚFAL MFF UK, http://hdl.handle.net/11234/1-1520.

Zeller, B. et al. (2013): DerivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia: ACL, s. 1201–1211.

Magda Ševčíková | Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Ústav formální a aplikované lingvistiky| Malostranské nám. 25, 118 00 Praha 1
sevcikova@ufal.mff.cuni.cz

Zdeněk Žabokrtský | Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Ústav formální a aplikované lingvistiky| Malostranské nám. 25, 118 00 Praha 1
zabokrtsky@ufal.mff.cuni.cz

Jonáš Vidra | Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Ústav formální a aplikované lingvistiky| Malostranské nám. 25, 118 00 Praha 1
vidra@ufal.mff.cuni.cz

Milan Straka | Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Ústav formální a aplikované lingvistiky| Malostranské nám. 25, 118 00 Praha 1
straka@ufal.mff.cuni.cz