M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

Identifikace derivačních vztahů ve španělštině

JÁN FARYAD

ÚFAL Technical Report
TR-2019-63

ISSN 1214-5521

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czechia

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

Identifikace derivačních vztahů ve španělštině

Ján Faryad

Ústav formální a aplikované lingvistiky,
Matematicko-fyzikální fakulta, Univerzita Karlova

13. září 2019

Abstrakt

V posledních letech se v rámci počítačové lingvistiky věnuje stále větší pozornost derivační morfologii, která zkoumá proces odvozování slov v jazyce (např. *znak* → *značka* → *značkový*). Za tímto účelem vznikají lexikální databáze odvozených slov v různých jazycích.

Předložená zpráva popisuje projekt, během něhož vznikla derivační databáze pro španělštinu DeriNet.ES. Podle vzoru české databáze DeriNet reprezentuje španělská databáze slova sdílející týž slovní kořen jako zakořeněný orientovaný strom. Derivační vztahy byly vytvářeny na základě substitučních pravidel. Pomocí metrik precision a recall je představena úspěšnost vytvořených derivačních párů.

Jedním z přínosů práce je prezentovaná metoda určování derivačních vztahů, zejména pak použitá substituční pravidla. Vzniklá databáze obsahuje 151 173 lemmat a 36 935 derivačních vztahů a je dostupná na webových stránkách Ústavu formální a aplikované lingvistiky.

1 Úvod

Ve srovnání s flektivní morfologií, jíž se počítačová lingvistika už dlouho zabývá (etablovanou úlohou je např. morfologická analýza - Hajič 2004, Straková et al. 2014), se derivační morfologii začala věnovat větší pozornost teprve v posledních letech. Tato lingvistická disciplína zkoumá proces odvozování slov, zvláště pomocí předpon a přípon (např. *učit* → *naučit, učit* → *učitel*). Za tímto účelem vznikají lexikální databáze, které se snaží tento jazykový jev náležitě modelovat. Na Ústavu formální a aplikované lingvistiky (ÚFAL) MFF UK byla práce na tvorbě těchto zdrojů orientována zpočátku pouze na češtinu. Lexikální databáze českých derivátů DeriNet (Ševčíková et al., 2016) je zde budována od roku 2012. Současná verze databáze DeriNet 2.0 (Vidra et al., 2019) obsahuje přes milion lemmat propojených více než 800 000 derivačními vztahy.

Za účelem zkoumání lingvistických jevů napříč jazyky vznikají jazykové zdroje anotované podle stejných pravidel – v oblasti syntaxe např. anotační

projekt Universal Dependencies, na jehož vývoji se ÚFAL podílí (Nivre, 2019). Podobným způsobem se uvažuje o harmonizaci anotačních stylů různých jazykových zdrojů modelujících derivace (Kyjánek et al., 2019). Na ÚFAL vznikly pilotní derivační databáze také pro další jazyky včetně španělštiny (Lango et al., 2018).

Struktura španělské databáze byla inspirována českým DeriNetem. Lemmata v ní jsou modelována jako uzly a vztah derivace mezi nimi je reprezentován hranou vedoucí od derivovaného lemmatu k základovému. Slova sdílející tentýž lexikální kořen pak tvoří v databázi zakořeněný orientovaný strom. Pilotní španělská databáze Spanish Word-Formation Network (WFN) obsahovala 162 751 lemmat a 18 441 derivačních vztahů mezi nimi, řada z nich však byla chybná.

Předkládaná zpráva popisuje projekt, jehož původním cílem bylo upravit a rozšířit španělskou derivační databázi Spanish WFN co do sady lemmat i derivačních relací. Vzhledem k množství chybných derivačních relací však bylo rozhodnuto, že vznikne nová databáze, DeriNet.ES, do níž bude převzato schéma modelování závislostí a také většina lemmat z původní sady, nicméně síť derivačních vztahů bude vybudována nově, nezávisle na derivacích v původní databázi.

Příspěvek je rozdělen do tří částí. V první části (oddíl 2) je představen postup úpravy a doplnění původní sady lemmat, druhá část (oddíl 3) podrobně popisuje postup vytváření nových derivačních relací mezi lemmaty. Vyhodnocení správnosti derivačních vztahů v nové databázi DeriNet.ES a její srovnání se Spanish WFN je provedeno ve třetí části (oddíl 4).

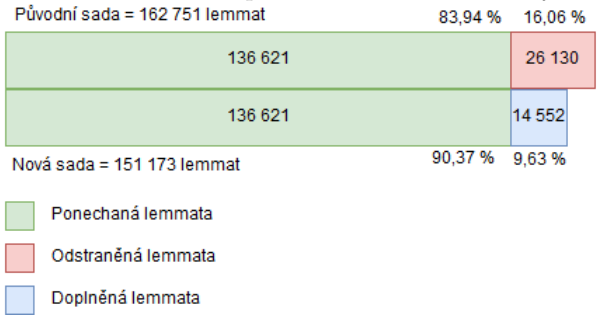
2 Úprava sady lemmat

Prvním cílem práce bylo upravit sadu lemmat. Úprava proběhla ve dvou krocích: doplnění nových lemmat z jiných zdrojů a odstranění lemmat, která z různých důvodů nejsou vhodná pro zařazení do databáze.

Za účelem doplnění lemmat byly použity dva další zdroje. Šlo zejména o 22. vydání slovníku Španělské královské akademie (Ruiz, 2010). Slovník sice obsahuje 87 899 lemmat, nicméně velká část jich už byla v původní databázi zahrnuta. Dalším použitým zdrojem byl seznam archaických výrazů (Flores, 2018). Tento zdroj však obsahuje pouze 3 632 lemmat, z nichž se část nacházela už v původní databázi a většina byla shodná s lemmaty z prvního zdroje. Ve výsledku bylo do databáze přidáno 14 552 nových lemmat, převážně z prvního zdroje.

V lemmasetu Spanish WFN se nacházela také nevhodná lemmata, která bylo zapotřebí odstranit. Jednalo se např. o neslovní řetězce: interpunkci, zkratky nebo unigramy a bigramy bez významu ve španělštině (např. *q, pu, my, oz*). Tato lemmata bylo možné snadno identifikovat a odstranit buď automaticky, nebo ručně. Jinou skupinu nežádoucích lemmat tvořila slova, která nepatří do běžné slovní zásoby španělštiny. Mezi nimi vynikalo nepřiměřeně velké množství francouzských toponym, konkrétně názvů obcí. Jejich odstranění bylo obtížným

Obrázek 1: Porovnání původní a revidované sady lemmat



úkolem, neboť mnoho z nich není snadno odlišitelných od jiných vlastních jmen. U španělských či mezinárodních vlastních jmén je vhodné, aby byla v databázi zahrnuta, protože jsou přirozenou součástí jazyka a vstupují do slovotvorných procesů. Proto byla odebrána pouze taková francouzská toponyma, která bylo možné spolehlivě identifikovat, neboť obsahovala větší počet pomlček (např. *Saint-Pierre-du-Val*) nebo typické koncové řetězce (*-ville* a *-court*). Celkově bylo z původní sady odstraněno 26 130 lemmat.

Lemmaset databáze DeriNet.ES má 151 173 lemmat, což je sice méně, než měla Spanish WFN, zato jsou však revidovaná lemmata reprezentativnější než původní sada lemmat. Na obrázku 1 je znázorněn přehled počtu a podílu odstraněných a nově přidaných lemmat v obou sadách.

3 Vytváření derivačních vztahů

Druhým úkolem projektu bylo nalézt mezi lemmaty derivační vztahy. Řešení spočívalo v navržení substitučních pravidel, podle kterých byly afixy (tj. prefixy i sufixy) derivátů nahrazeny afixem předpokládaného základového slova. Pokud bylo takto vytvořené lemma přítomné v databázi, byl mezi ním a derivátem vyznačen vztah derivace. Pro dosažení lepšího pokrytí se nebrala v úvahu skutečná podoba lemmat, nýbrž podoba normalizovaná: bez diakritiky a s konverzí na malá písmena.

Každé pravidlo má tvar *[derivační afix]* → *[základový afix]*, např. *able* → *ar*. Při tvorbě pravidel se dodržovala podmínka, aby afix derivátu nebyl kratší než afix základového slova.

Lemmata mají v databázi podobu řetězce znaků a nelze poznat, kde je hranice mezi kořenem a afixem. Proto byla provedena analýza frekvence stejných podřetězců na začátku a na konci lemmat, což pomohlo odhadnout, které podřetězce představují afixy. Pro tento rozbor byl vytvořen nástroj, který seřadí počáteční a koncové podřetězce podle četnosti jejich výskytu. Dále pro každý podřetězec prozkoumá, u kolika lemmat jej lze nahradit jiným podřetězcem, a tyto možné substituce opět seřadí podle frekvence (výstup nástroje je znázorněn na obrázku 2). Takto vytvořený seznam posloužil jako jeden z podkladů pro

Obrázek 2: Ukázka výstupu nástroje: počet výskytů sufixu a jeho možných substitucí i s náhodnými příklady

```
1123 able
    776 ar      arrug- compar- honor- contrat-
    502 ador    polariz- limpi- respet- aconsej-
    354 ado     conden- guard- enchuf- identific-
    233 o       alter- lament- tach- ecu-
    217 ante    magnetiz- coagul- ferment- son-
    188 a       deriv- visit- vapor- conjetur-
     88 al      lacrim- emocion- interes- proces-
     82 e       grav- moj- pesc- confort-
     77 ada     entreg- fum- escal- j-
     70 oso     vituper- honr- agasaj- dud-
```

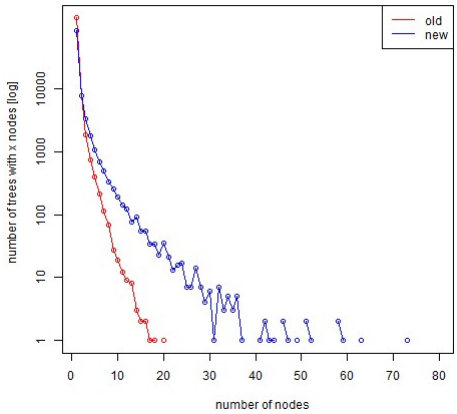
návrhy substitučních pravidel. Několik dalších pravidel bylo navrženo na základě vlastní jazykové intuice.

Uvedeným způsobem bylo pro prefixy vytvořeno 43 pravidel, přičemž základový afix je vždy nulový – jde tedy o pouhé připojení předpony na začátek základového slova. Jedná se převážně o předpony odvozené ze španělských předložek (např. *tras-, con-, sub-*) nebo z řeckých či latinských slov (např. *hipo-, super-, micro-, multi-*).

U sufixů je situace složitější: základové sufixy jsou většinou nenulové a navíc je velmi časté, že jedem derivační sufix může být pomocí různých pravidel nahrazen různými základovými sufixy. Mnohdy bývá derivační sufix substituován typickými jmennými (*-o, -a, -e*) nebo slovesnými (*-ar, -er, -ir*) sufixy, přičemž v různých případech je třeba použít jiné z těchto pravidel pro nalezení existujícího základového slova. Pravidla jsou uvedena v sestupném pořadí podle jejich pravděpodobnosti. Pokud je v datech přítomno více slov vzniklých použitím různých pravidel náležících jednomu derivačnímu sufixu, použije se první z těchto pravidel, tedy to nejpravděpodobnější. Derivačních sufixů je 110 a jsou nahrazovány základovými sufixy pomocí 227 sufixových pravidel.

Jiným důležitým rozhodnutím bylo určit, zda se mají nejprve uplatňovat prefixová, či sufixová pravidla (např. slovo *desfragmentación* by mohlo být odvozeno prefixovou substitucí *des* $\rightarrow$ *0* od *fragmentación* nebo sufixovou substitucí *ación* $\rightarrow$ *ar* od *desfragmentar*). Na základě rozboru vzorku příkladů byla dána přednost sufixovým pravidlům. Pro každé pravidlo bylo náhodně vybráno nejvýše deset vytvořených derivačních párů, které dostali ke kontrole dva anotátoři. U každého příkladu měli vyznačit, zda je základové slovo pro daný derivát navrženo správně. Pokud byly správné alespoň dvě třetiny vztahů vytvořených jedním navrženým pravidlem, pravidlo bylo ponecháno, jinak bylo ze seznamu odstraněno. Takto přísně nastavené kritérium sice může vést k nižší hodnotě recall (pokrytí) hledaných derivačních relací, je však zachována vysoká hodnota precision (přesnost), na níž byl kladen zvláštní důraz.

Obrázek 3: Srovnání velikosti stromů v původní a v nové databázi



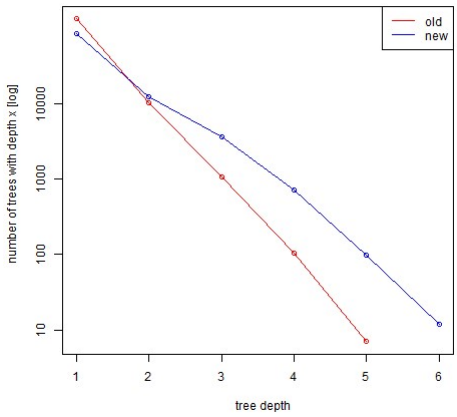
4 Evaluace výsledků

Oproti 18 441 derivačním relacím ve Spanish WFN jich DeriNet.ES obsahuje 36 935. Počet derivačních stromů tudíž klesl, a naopak narostla jejich velikost (počet lemmat v derivačním stromu; viz obrázek 3) a hloubka (počet úrovní v derivačním sdtromu; viz obrázek 4). Rovněž mají lemmata v nově vytvořené databázi v průměru více derivátů (viz obrázek 5).

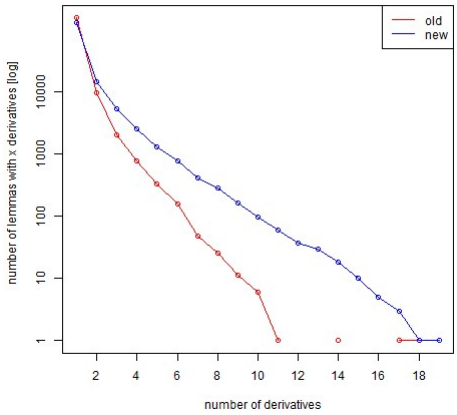
Závěrečné vyhodnocení úspěšnosti rozpoznávání derivací proběhlo opět na základě manuální anotace. Z dat v nové databázi DeriNet.ES bylo náhodně vybráno 300 lemmat i s jejich základovým slovem, pokud u nich bylo uvedeno. Anotátoři pro každé lemma vyznačili, zda je uvedení či neuvedení základového slova správně, a v případě chyby navrhli opravu. Proti týmž příkladům byly vyhodnoceny vztahy ve Spanish WFN, aby bylo možné provést srovnání a vyčíslit zlepšení. Vybrána byla sice slova přítomná v obou databázích, přesto se však se jimi zastupované lemmasety liší (viz oddíl 2). Změny v sadě lemmat však nebyly tak rozsáhlé, aby výrazně snížily porovnatelnost naměřených hodnot mezi oběma databázemi.

Porovnání počtů správných a nesprávných vztahů včetně typu chyby je uvedeno v tabulce 1. Pro evaluaci nelze přímočaře použít tradiční metriky precision a recall. Ty se obvykle používají pro úlohy, v nichž je třeba z dané množiny vzorků vybrat podmnožinu vzorků splňujících určitou vlastnost. Testované příklady se rozdělí do čtyř skupin: *true positives* (TP) zahrnuje příklady, které danou vlastnost splňují a byly správně vybrány, *true negatives* (TN) naopak ty, které vlastnost nesplňují a vybrány nebyly, *false positive* (FP) obsahuje vzorky, které vlastnost nesplňují, ale přesto byly nesprávně vybrány a nakonec ve skupině *false negative* (FN) se nacházejí příklady, která vybrány nebyly, ačkoli požadovanou vlastnost splňují. Skupiny FP a FN představují u těchto úloh dva

Obrázek 4: Srovnání hloubky stromů v původní a v nové databázi



Obrázek 5: Srovnání rozvětvenosti derivačních stromů v původní a v nové databázi



Tabulka 1: Srovnání správnosti derivačních vztahů v původních a v nové databázi

		Spanish WFN	**DeriNet.ES**
správně	(1) shodně bez rodiče	159	157
	(2) shodný rodič	21	73
chybně	(3) chybějící rodič	113	47
	(4) nadbytečný rodič	7	9
	(5) odlišný rodič	0	14
celkem		300	300

Tabulka 2: Matice konfuze derivačních vztahů v databázi

		zlaté vztahy od anotátorů	
		pozitivní	negativní
vztahy v databázi Spanish WFN	poz.	TP = 21	FP = 7
	neg.	FN = 113	TN = 159
vztahy v databázi DeriNet.ES	poz.	TP = 73	FP = 23
	neg.	FN = 47	TN = 157

typy chyb, k nimž může dojít. V případě určování derivačních vztahů však mohou nastat tři chyby: u lemmatu chybělo určení základového slova (v tabulce typ 3), u lemmatu bylo uvedeno základové slovo, přestože uvedeno být nemělo (typ 4), anebo bylo lemma správně vybráno jako derivované, ale s uvedením chybného základového slova (typ 5).

Jelikož jsou však precision a recall přehledné metriky, je vhodné se pokusit jejich pomocí popsat výsledky i za cenu mírné nepřesnosti. Pokud do *false positives* zahrneme obě kategorie s nesprávným základovým slovem, tedy s nadbytečným (typ 4) a odlišným od správného (typ 5), a ve *false negatives* ponecháme případy s chybějícím základovým slovem, dostaneme hodnoty uvedené v tabulce 2.

Z nich vycházejí hodnoty precision, recall a jejich harmonický průměr F_1-score prezentované v tabulce 3. Z těchto výsledků je patrné, že se podařilo výrazně zlepšit pokrytí derivačních vztahů v databázi při zachování podobné přesnosti.

Tabulka 3: Přibližné hodnoty precision, recall a F_1-score

	Spanish WFN	**DeriNet.ES**
precision	75,00 %	76,04 %
recall	15,67 %	60,83 %
F_1-score	25,92 %	67,59 %

Obrázek 6: Ukázka vyhledávacího dotazu a nalezených výsledků

5 Dostupnost dat

Vytvořená databáze DeriNet.ES byla v květnu 2019 zveřejněna na webových stránkách ÚFAL (http://ufal.mff.cuni.cz/derinet) pod licencí CC-BY-ND. Obsahuje 151 173 lemmat a 36 935 derivačních vztahů mezi nimi. Celou databázi je možné stáhnout v jednoduchém formátu *tsv*. K dispozici je také prohledávání databáze pomocí dotazů v jazyce DCQL (https://ufal.mff.cuni.cz/derinet/derinet-search). Na obrázku 6 je ukázka dotazu na slova končící příponou *-able* a několika výsledků v podobě derivačních stromů obsahujících nalezená slova.

6 Závěr

Zpráva představila postup tvorby nové španělské derivační databáze DeriNet.ES. Proces spočíval ve dvou krocích: v úpravě sady lemmat z původní databáze Spanish WFN a ve vytváření nových derivačních vztahů mezi lemmaty. V prvním kroce se podařilo odstranit řadu nežádoucích výrazů (např. interpunkci, zkratky, francouzská toponyma). Doplnění řady lemmat z jiných zdrojů však nestačilo tento úbytek vyvážit, takže výsledná sada je menší než původní. V budoucnu bude vhodné sadu dále doplňovat o lemmata z dosud nepoužitých zdrojů.

Ve druhém kroku se při zachování podobné přesnosti dosáhlo výrazně lepšího pokrytí derivačních vztahů ve srovnání s původní verzí databáze. Největším problémem stále zůstávají deriváty bez označeného základového slova. Jedná se většinou o příklady málo častých záměn základové předpony za derivační (např. *chasc**a*** → *chasc**udo***) nebo o derivace při nichž dochází ke změně v kořeni (např. *sup**o**ner* → *sup**ue**sto*). Část z nich by se podařilo zachytit, pokud by se při hledání základového slova uplatnila na poslední slabiku kořene některá z hláskových změn typických pro španělštinu (např. *e* ↔ *i*, *e* ↔ *ie*, *u* ↔ *ue*, *o* ↔ *ue*). Výrazný pokrok lze očekávat od natrénování modelu strojového učení, který by byl schopen rozpoznat některé derivační páry i bez explicitně formulovaného substitučního pravidla.

Ve zprávě byla představena úspěšná metoda hledání derivačních vztahů, která může být užitečná také pro nalezení derivací v dalších jazycích. Použitá substituční pravidla i nová databáze DeriNet.ES mohou sloužit jako materiál pro další zkoumání španělské derivační morfologie, např. pro anotaci některých souvisejících příznaků. Už během tvorby databáze totiž vyvstávala potřeba opatřit některá lemmata dodatečnými značkami pro složeniny, cizí slova nebo vlastní jména.

Seznam použité literatury

Flores, I. M. M. (2018). *Palabras en Desuso*. Madrid: Real Academia Española, dostupné na `https://www.cpimario.com/endesuso.htm`, naposledy přistoupeno 17. 5. 2019.

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Linguistic Data Consortium, University of Pennsylvania.

Kyjánek, L., Žabokrtský, Z., Ševčíková, M., and Vidra, J. (2019). Universal derivations kickoff: A collection of harmonized derivational resources for eleven languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, Prague.

Lango, M., Ševčíková, M., and Žabokrtský, Z. (2018). Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1853–1860.

Nivre, J. (2019). *Universal Dependencies 2.4*. Praha: LINDAT/CLARIN digitální knihovna na ÚFAL MFF UK, dostupné na `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2988`, naposledy přistoupeno 27. 8. 2019.

Ruiz, T. (2010). *Lemario Actualizado del Español*. Madrid: Real Academia Española, dostupné na `https://www.teoruiz.com/archivos/-2010/10/17/lemario/-actualizado-espanol`, naposledy přistoupeno 17. 5. 2019.

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Ševčíková, M., Žabokrtský, Z., Vidra, J., and Straka, M. (2016). Lexikální síť DeriNet: elektronický zdroj pro výzkum derivace v češtině. *Časopis pro moderní filologii*, 98:62–76.

Vidra, J., Žabokrtský, Z., Kyjánek, L., Ševčíková, M., and Dohnalová, Š. (2019). DeriNet 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Nguy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová, *Manual for annotation of discourse relations in the Prague Dependency Treebank*

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)*

ÚFAL TR-2013-54 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *From PDT 2.0 to PDT 3.0 (Modifications and Complements)*

ÚFAL TR-2014-55 Rudolf Rosa, *Depfix Manual*

ÚFAL TR-2014-56 Veronika Kolářová, *Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex*

ÚFAL TR-2014-57 Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, Lenka Šíková, *Annotation of coreference in Prague Czech-English Dependency Treebank*

ÚFAL TR-2015-58 Zdeňka Urešová, Eva Fučíková, Jana Šindlerová, *CzEngVallex: Mapping Valency between Languages*

ÚFAL TR-2015-59 Kateřina Rysová, Magdaléna Rysová, Eva Hajičová, *Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description*

ÚFAL TR-2016-60 Kira Droganova, Daniel Zeman, *Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies*

ÚFAL TR-2018-61 Lukáš Kyjánek, *Morphological Resources of Derivational Word-Formation Relations*

ÚFAL TR-2019-62 Zdeňka Urešová, Eva Fučíková, Eva Hajičová, *CzEngClass: Contextually-based Synonymy and Valency of Verbs in a Bilingual Setting (CzEngClass: Kontextová synonymie a valence sloves v bilingvním prostředí)*

ÚFAL TR-2019-63 Ján Faryad,
Identifikace derivačních vztahů ve španělštině