ČASOPIS PRO MODERNÍ FILOLOGII 99, 2017, Č. 1, S. 40–54

OPEN ACCESS

Nástroj na tvaroslovnou analýzu staré angličtiny

Ondřej Tichý (Praha)

MORPHOLOGICAL ANALYSER OF OLD ENGLISH

The paper describes the construction and testing of an electronic application for semi-automatic morphological analysis of Old English.

It introduces the state of the art in the field of electronic analysis of Old English, provides a brief overview of Old English morphology and discusses the reasoning behind our theoretical framework. An account of the chosen methodology is offered and a specific description of its implementation is provided: from the acquisition and preparation of the lexical input data, through the programming of the forms generator to the testing of the results by analysing Old English text. The resulting recall of 95% is a success; however, the paper also hints at how it may be improved. It also discusses further use and development of the analyser, especially the disambiguation of its results.

The paper makes a future semi-automatic morphological tagging of Old English texts a real possibility.

KEYWORDS

Old English, historical linguistics, computational linguistics, automatic morphological analysis, morphology, forms generator

KLÍČOVÁ SLOVA

stará angličtina, historická lingvistika, komputační lingvistika, automatická morfologická analýza, morfologie, generátor tvarů

1. ÚVOD[1]

Současný diachronní výzkum jazyka se stále častěji jako k zásadnímu zdroji nových poznatků i pro ověření poznatků starších obrací k jazykovým korpusům. Práce s korpusy staršího jazyka má však svá významná specifika a úskalí. Diachronní korpusy jsou z pohledu synchronního výzkumu stále málo rozsáhlé (běžné jsou korpusy do několika milionů pozic/tokenů) a především bývají jen velmi spoře označkovány. U synchronních korpusů (obzvláště moderní angličtiny, ale i češtiny) dnes běžně očekáváme např. lemmatizaci nebo přímo morfologické a syntaktické značkování. U diachronních korpusů se s ním setkáme zřídka. To však zásadně komplikuje využití korpusů pro analýzu historického jazyka, neboť např. bez lemmatizace je prakticky nemožné provádět kolokační analýzu a většina frekvenčních analýz (základní metoda korpusové analýzy) je na nelemmatizovaném materiálu mnohem náročnější.

1 Článek je shrnutím nepublikované stejnojmenné disertační práce obhájené v roce 2014 na FF UK.

Jednou z hlavních příčin tohoto stavu je, že značkování (a lemmatizaci budeme pro tyto účely pokládat za druh morfologického značkování) historického jazyka je výrazně obtížnější než značkování jazyka současného. To souvisí především s možnostmi automatizace značkovacího procesu. Pro moderní angličtinu již existují poměrně spolehlivé nástroje pro morfologické i syntaktické značkování, nakonec lemmatizovat moderní anglický text je např. z pohledu moderní češtiny s jistou nadsázkou téměř banální. Ovšem i pro nesporně komplexnější českou gramatickou morfologii existují kvalitní nástroje automatické analýzy. Problém však nespočívá jen v morfologické komplexitě, ale především ve standardizaci.

Každému, kdo kdy pracoval s historickým jazykem či s nestandardizovanou jazykovou varietou je jasné, že pravopis, formy a pravidla z příruček často přesně neodpovídají jazyku dokladů. Jako jsou jazykové příručky moderních standardních variet (např. spisovné češtiny) abstrakcí nad komplexnější jazykovou realitou (langue/parole), jsou příručky týkající se historických variet jazyka ještě navíc abstrakcí nad komplexitou nářečí, písařských tradic a idiosynkrasií případně diachronních rozdílů v rámci popisovaných období.

Motivací k vytvoření nástroje na automatickou tvaroslovnou analýzu staré angličtiny je tedy především usnadnit či v některých případech vůbec umožnit korpusový výzkum tohoto období anglického jazyka, přičemž jak z výše uvedeného vyplývá, tento nástroj si musí umět poradit jak s morfologickou komplexitou staré angličtiny, tak s nízkou standardizací jejích dokladů.

2. TEORETICKÝ RÁMEC

Při bližším zkoumání dosavadních výsledků na poli automatického zpracování staroanglické morfologie se ukázalo, že se nejedná o problém zcela nový, byť jde bezesporu o problém zatím uspokojivě nevyřešený. Část dosavadních pokusů o automatizaci tvaroslovné analýzy se omezovala jen na část jazykové struktury. Např. *Verbix* (Lindberg, 1995), *The Nerthus Project* (Martín Arista, 2004) nebo *Morphological Analyzer for Old English Verbs* (Adams, 2007) se zaměřily pouze na slovesnou morfologii. Část se zase zabývala pouze úzkým výběrem materiálu určeného k analýze — např. *Morphological Analyser of Old English Texts* (Calle Martín et al., 1997–2001). Dva dosud nedokončené projekty pak přímo navazují na naši datovou základnu (resp. stejný zdroj lexikálních dat): jedná se o lemmatizaci staroanglické básně *Daniel* (Kleinman, 2012) a lemmatizační projekt korpusu *YCOE*.[2]

Společným rysem všech těchto projektů je teoretický rámec, který určuje jako základ nástroje různým způsobem generovaný slovník, se kterým jsou následně porovnávány doklady, tedy samotné formy vyskytující se ve staroanglických textech. Způsob porovnávání a následné analýzy se však liší. V nejjednodušším případě (Kleinman) jde o pouhé porovnávání řetězců znaků s využitím tzv. Levenštejnovy vzdálenosti (míra podobnosti textových řetězců), v rafinovanějších nástrojích se položky

2 *The York-Toronto-Helsinki Parsed Corpus of Old English Prose.* Informace o projektu Ans Van Kemenade a Erwina Komena byly získány v roce 2013 soukromou korespondencí.

slovníku (lemmata) upravují na základě morfologických pravidel a pak se teprve porovnávají s doklady.

Podobným způsobem jsme se rozhodli postupovat i v našem případě, jelikož však jsou výše uvedené nástroje, resp. jejich popis, buď nedostupné, nebo neúplné, rozhodli jsme se v obecných rysech inspirovat implementačně ověřeným řešením morfologické analýzy moderní češtiny (Osolsobě, 1996; Sedláček, 1999; Sedláček — Smrž, 2001).

Takto stanovený teoretický rámec předpokládá tři základní fáze řešení:

1. **Přípravu dat**, která budou tvořit **základní slovník** nástroje. Základní slovník by měl být co do pokrytí staroanglického lexika pokud možno vyčerpávající, měl by kromě samotného lexika obsahovat i základní morfologické informace o jednotlivých položkách.
2. Tvorbu **generátoru**, který z položek základního slovníku na základě morfologických **pravidel** vygeneruje všechny gramatické formy staroanglických slov a ke každé formě přiřadí i příslušné morfologické kategorie, tedy **slovník forem**.
3. Vytvoření aplikace, která bude srovnávat doklady staroanglického textu s položkami slovníku forem, tedy samotného **analyzátoru** v užším slova smyslu. V případě shody bude doklad označen jak odkazem na příslušnou položku základního slovníku (lemmatizace), tak na morfologické kategorie patřící k shodné položce slovníku forem (morfologická analýza).

K výsledné lemmatizaci a morfologickému značkování využitelné badateli však bude třeba výsledky automatické analýzy ještě disambiguovat (ať již ručně, nebo z části strojově), jelikož nelze předpokládat, že by si nástroj pouze na základě morfologie poradil např. s výběrem z homonymních tvarů. Proces disambiguace však již není součástí tohoto nástroje na tvaroslovnou analýzu.

Jak bylo předesláno v úvodu, konkrétní implementace všech tří fází nutně vychází ze specifik staroanglického tvarosloví a musí zároveň odpovídat skutečnému stavu analyzovaných dokladů.

3. MORFOLOGIE STARÉ ANGLIČTINY

Stará angličtina, tedy angličtina přibližně z let 700–1100, je na rozdíl od angličtiny moderní z typologického hlediska jazyk ještě převážně flektivní, byť s flexí, oproti některým indoevropským jazykům, jakým je třeba čeština, značně omezenou. Ve staré angličtině se tak setkáme s podobnou deklinací podstatných jmen, přídavných jmen a zájmen, jakou známe např. z moderní němčiny a stejně jako v moderní němčině je ve staré angličtině zvláště rozvinutá slovesná konjugace.

Rámec tohoto článku neumožňuje podrobný popis staroanglického tvarosloví, omezíme se jen na několik obecných poznámek vztahujících se konkrétně k našemu nástroji.

Základní rozdíl mezi tradičními filologickými popisy staroanglického tvarosloví (např. Wright — Wright, 1914), případně moderními popisy určenými studentům (např. Baker, 2012) a popisem, který je výsledkem analýzy staroanglického tvarosloví

za účelem strojového zpracování, je především v úrovni detailu, který je věnován různým jazykovým strukturám. Tradiční popisy staroanglické morfologie sledují kromě popisu samotného i další cíle, např. důraz na historickou kontinuitu struktur a jevů, a někdy tak v popisu rozlišují i struktury, které se již v daném období formálně neliší (např. paradigmata navazující na paradigmata původně odlišná ve své rekonstruované podobě, ve staroanglickém materiálu však již formálně nerozlišená). Náš popis byl v tomto ohledu naopak veskrze pragmatický, a tedy např. nerozlišoval ani takové struktury, jejichž formální odlišnost dle tradičního popisu byla nižší než faktické formální kolísání těchto struktur v textu, ať již jde o kolísání písařské, nářeční nebo diachronní.[3]

Náš popis je však v některých ohledech naopak detailnější, a to jak vzhledem k výrazně okleštěným popisům moderním určeným studentům, tak často i vzhledem k tradičnímu popisu — musí totiž postihnout všechny struktury, které obvyklé popisy zahrnují pod výjimky, či vzhledem k jejich okrajovosti zanedbávají. Oproti tradičním sedmi silným, třem slabým a několika málo préteritoprézentním a nepravidelným slovesným paradigmatům zavádí náš popis pro účely strojového zpracování 132 slovesných tříd.

Dalším rozdílem, který odlišuje náš popis od tradičních, je naše orientace na slovník, a tedy na lemma jakožto základní tvar flektovaných tvarů i v případech, kdy tradiční a filologicky smysluplnější analýza vychází např. z kmene či kořene. Důvod je opět pragmatický — připravit základní slovník lemmat je výrazně jednodušší než připravit slovník kořenů či tvarotvorných kmenů. Zásadou našeho popisu tedy je, že pomocí námi stanovených morfologických pravidel musí být možné z jakéhokoliv lemmatu odvodit všechny flektované formy daného slova. U deklinačních typů lze zpravidla všechny flektované tvary vytvořit skutečně jen na základě lemmatu a koncovek, u slovesných konjugací je situace složitější a bude popsána níže v části týkající se implementace.

Je však nutné zdůraznit, že v jádru náš popis vychází z popisů tradičních, a pokud k tomu není dobrý důvod, snaží se jich držet. Předpokládáme, že i díky tomu bude výsledná morfologická analýza pro uživatele seznámené s tradičními popisy dobře srozumitelná a průhledná.

4. IMPLEMENTACE

Jak jsme naznačili v části popisující teoretický rámec našeho řešení, implementace je rozdělena na tři základní fáze.

4.1 PŘÍPRAVA DAT

Na základě vyčerpávajícího přehledu lexikografických zdrojů mapujících staroanglický materiál (Tichý, 2007) jsme se rozhodli použít jako zdroj lemmat pro základní

3 Jde tedy v podstatě o poměr signálu a šumu. Pokud např. tradiční popis rozlišuje struktury na základě formálního rozdílu *-on* vs. *-an*, avšak skutečný výskyt těchto forem je vzhledem k tradičně popsaným strukturám statisticky nahodilý, v našem popisu tento rozdíl zanedbáváme.

slovník *An Anglo-Saxon Dictionary* (Bosworth — Toller, 1921, dále jen BT). Jeho hlavní výhodou je šíře pokrytí, neb se jedná o dosud jediný slovník alespoň teoreticky pokrývající celou šíři staroanglického lexika. Při přípravě dat bylo navíc možné vyjít z předchozího digitalizačního projektu (Tichý, 2007), přesto však bylo nutné v elektronické verzi slovníku provést celou řadu úprav (značkování morfologických informací, opravy v makrostruktuře slovníku atp.), než bylo možné exportovat seznam lemmat se základními morfologickými informacemi a identifikátory umožňujícími odkazovat zpět na původní slovníková hesla.

Data základního slovníku jsou dále obohacena informacemi z *Old English Grammar* (Wright — Wright, 1914), resp. obohacena o morfologické informace jsou lemmata, která se vyskytují v obou zdrojích (viz *tabulka 1*). Mluvnice manželů Wrightových (dále jen mluvnice) byla využita jednak kvůli své struktuře — všechna příkladová slova jsou obsažena v indexu ve formě lemmat a jasně odkazují na příslušné morfologické informace, jednak proto, že jde o podrobnou, tradiční a dosud hojně využívanou gramatiku. Oproti podrobnější a modernější mluvnici Campbelově (1983) je navíc licenčně nezatížená a vytěžené informace tak bylo možné využít i ve volně přístupné online verzi Anglosaského slovníku (viz bosworthtoller.com).

Nakonec byl základní slovník obohacen o ručně sestavená morfologická data týkající se příkladových slovesných paradigmat, která generátoru popsanému v další fázi umožňují zohledňovat ve flektovaných tvarech i infixaci (viz *tabulka 2*).

4.2 GENERÁTOR

Generátor forem nejprve načítá data základního slovníku a provádí jejich základní standardizaci. Jedná se o standardizaci, která buď nemá dopad na samotnou analýzu ani její výsledky (např. sjednocení alografů *thorn* a *edh*) nebo dává vzniknout alternativním formám lemmat, která budou hrát v dalším procesu různou úlohu (např. značení délky samohlásek).[4] Generátor při načítání dat také automaticky dopočítává některé informace o jednotlivých položkách, které se následně využívají pro generování a přiřazování forem (např. počet slabik, rozlišení délky kmene atp.).

Následují dvě etapy, v nichž generátor nejprve přiřazuje položkám základního slovníku vzorová paradigmata[5] a následně na základě paradigmatu a konkrétní položky generuje jednotlivé flektované slovní tvary.

Přiřazování paradigmat probíhá dle následujícího algoritmu:

1. Nejprve jsou jakožto textové řetězce porovnány položky základního slovníku s lemmaty vzorových paradigmat, přičemž se bere ohled i na doplňkové mor-

4 Značení délky sice odráží reálný fonologický rozdíl v kvantitě, v původních textech se ale nevyskytuje a je až příspěvkem moderních editorů (např. autorů BT či gramatiky). Generátor i analyzátor jej tedy zanedbávají, ale alternativní forma lemmatu s diakritikou převzatou z původních zdrojů základního slovníku je uchována a je tedy k dispozici uživatelům, kteří si např. přejí analyzovat neautentický moderně editovaný text, případně kteří chtějí využít generovaný materiál k výuce a upřednostňují tak text s diakritikou.

5 Vzorová paradigmata jsou odvozena z mluvnice.

ID	lemma	wright	noun	pron.	adj.	verb	part.	adv.	prep.	conj.	interj.	num.	w. vb.	s. vb.	c. vb.	pp. vb.	a. vb.	un. vb.	m. n.	f. n.	n. n.	u. n.
008031	DRÝ	142;388	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
008032	drý-cræft	NULL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
008033	drý-cræftig	NULL	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
008035	drýfan	NULL	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
008036	drygan	530	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
008040	dryht	390	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
008041	Dryht	NULL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
008042	dryht-bearn	NULL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
008043	dryht-cwén	NULL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
008044	dryht-ealdor	NULL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
008045	dryhten	288;340;563	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
008047	dryhten-bealo	NULL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
008048	dryhten-dóm	NULL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

TABULKA 1. Ukázka dat základního slovníku

ID	verb ID	type	class	subdiv	subc.	§ W	variant	paraID	prefix	PreV	V	PoV	bnd	dent	end
2	sníðan	s	1	0	c	491	0	Inf		sn	í	0	ð		an
2	sníðan	s	1	0	c	491	0	PaInSg1		sn	á	0	ð		0
2	sníðan	s	1	0	c	491	0	PaInPl		sn	i	0	d		on
2	sníðan	s	1	0	c	491	0	Pp		sn	i	0	d		en
74	fremman	w	1	A	b	526	0	Inf		fr	e	mm			an
74	fremman	w	1	A	b	526	0	PsInSg2		fr	e	m			est
74	fremman	w	1	A	b	526	0	PaInSg1		fr	e	m		ed	e
74	fremman	w	1	A	b	526	0	PaInSg1		fr	e	m		od	e
74	fremman	w	1	A	b	526	1	Inf		fr	e	m			an
74	fremman	w	1	A	b	526	2	Inf		fr	e	m	i		an

TABULKA 2. Ukázka struktury slovesných paradigmat (sloves *sníðan* a *fremman*)

fologické informace (slovní druh, typ slovesa, rod substantiva atp.), aby se předešlo chybnému přiřazení homonym. Bere se také ohled na slovotvornou strukturu lemmat — pokud je na začátku lemmatu nalezena některá z předpon, které jsou obsaženy v BT, hledá se nejprve shoda s předponou, poté i bez předpony. V případě dosud nepřiřazených položek se generátor pokusí nalézt shodu alespoň s částí lemmatu již přiřazených položek, čímž dojde k přiřazení řady kompozit, která se zpravidla skloňují podobně jako jejich řídící člen.

2. Dosud nepřiřazené položky jsou následně přiřazeny na základě odkazů k vzorovým paradigmatům v mluvnici. Jelikož tím se okruh přiřazených položek zvýší, opakuje se přiřazování dosud nepřiřazených k těmto již nově přiřazeným položkám pomocí porovnávání textových řetězců dle prvního kroku.

3. Následuje morfonologická analýza[6] dosud nepřiřazených položek a jejich přiřazení paradigmatům na jejím základě. Tímto způsobem lze přiřadit všechna dosud nepřiřazená adjektiva.
4. Zbývající nepřiřazené položky jsou přiřazeny na základě pravděpodobnosti, tedy k nejobvyklejším paradigmatům odpovídajícím jejich morfologii.
   a. Všechna silná slovesa jsou přiřazena ke vzoru *helpan*, všechna ostatní k typu *déman*.
   b. Všechna maskulina a substantiva nejistého rodu jsou přiřazena ke vzoru *stán*, všechna feminina k typu *ár*, všechna neutra s dlouhým kmenem k typu *word* a všechna neutra s krátkým kmenem k typu *hof*.

Generování flektovaných forem probíhá postupně po jednotlivých slovních druzích, přičemž se využívá shod ve flexi mezi slovními druhy, takže např. flektovaná slovesná participia jsou nejprve vygenerována se slovesy v základním tvaru a poté skloňována s adjektivy. Až na slovesa jde většinou o pouhé odstranění případné koncovky lemmatu, změnu na morfematickém švu a připojení koncovky dle daného paradigmatu. U sloves je navíc využita informace o změně samohlásky v kořenové slabice (infix), proto je nejprve analyzována struktura lemmatu, nalezena kořenová samohláska a nahrazena buď samohláskou odpovídající paradigmatu (ablautové samohlásky), nebo samohláskou přehláskovou, která je odvozena dle přehláskových pravidel ze samohlásky původní.

Při generování jsou některé formy vytvářeny ve více alternativách ať už vzhledem k dubletám daným morfologickými či fonologickými pravidly nebo k častým případům pravopisného, nářečního či diachronního kolísání forem. K výsledným formám jsou přiřazeny i informace zdůvodňující vygenerování alternativní formy, ve výsledku lze tedy např. odlišit formy různých nářečí či písařských tradic.

Každá vygenerovaná forma je také obohacena o informace o všech příslušných morfologických kategoriích, o vzorovém paradigmatu a základní morfonologické struktuře, na jejímž základě byla složena.

Zájmena a některé nepravidelné tvary sloves by bylo natolik neefektivní generovat (jejich počet je malý a formální souvislost se základním tvarem příliš malá), že jsou pro účely nástroje flektovány ručně a takto vloženy do slovníku forem.

Analyzátor v užším slova smyslu je již vlastně jen aplikace, která srovnává doklady staroanglického textu vložené uživatelem s formami vyprodukovanými generátorem a zobrazující informace o těchto formách, které k nim generátor přiřadil. Zásadní komplikaci při přiřazování dokladů však způsobuje již zmíněné kolísání forem. To je zčásti řešeno výše popsaným zavedením alternativních forem při generování, není tak ale vyřešeno zcela, jelikož řada případů kolísání je ve fázi generování obtížně předvídatelná a především, pokud by se mělo se všemi možnostmi kolísání počítat již při generování, počet vygenerovaných forem by neúnosně narostl (viz níže).

Proto analyzátor pracuje s tzv. variačními filtry, které umožňují přiřadit i doklady, které bez nich nedopovídají žádné položce slovníku forem. Tyto filtry jsou

6 Počet slabik, jejich délka atp.

paradigma	sloves	paradigma	slov.	paradigma	slov.	paradigma	slov.	paradigma	slov.
sealfian	2342	hyngrian	23	hycgan	10	spanan	6	pǽcan	3
déman	1758	sníþan	22	hebban	9	þerscan	6	streccan	3
nerian	387	bregdan	21	rǽdan	9	witan	6	unnan	3
bídan	226	dón	20	rísan	9	béatan	5	ágan	2
bindan	200	sléan	20	sécan	9	lácan	5	bréoþan	2
fremman	176	gangan	19	wríþan	9	munan	5	cunnan	2
drencan	123	cuman	18	libban	8	tellan	5	dreccan	2
faran	115	cweþan	18	scéadan	8	þeccan	5	dugan	2
béodan	112	twéogan	18	wegan	8	wesan	5	durran	2
helpan	108	hléapan	15	willan	8	gilpan	4	dwellan	2
metan	95	wyrcan	15	brengan	7	hnígan	4	fricgan	2
weorpan	81	gierwan	14	reccan	7	rǽcan	4	frignan	2
beran	73	hátan	14	séoþan	7	slǽpan	4	gellan	2
weorþan	57	lætan	14	stellan	7	smúgan	4	mígan	2
fealdan	55	giefan	13	stígan	7	tǽcan	4	mótan	2
brúcan	53	gieldan	13	stregdan	7	þyncan	4	nugan	2
téon	52	þencan	13	þryccan	7	wǽcan	4	sceþþan	2
blótan	46	búgan	12	béon	6	weccan	4	scippan	2
bláwan	40	niman	12	bycgan	6	blandan	3	stæppan	2
settan	39	swerian	12	féolan	6	cweccan	3	þurfan	2
gán	30	berstan	11	findan	6	cwellan	3	wleccan	2
céosan	28	gitan	11	ícan	6	drǽdan	3	birnan	1
séon	27	habban	11	leccan	6	hlehhan	3	magan	1
lǽcan	26	licgan	11	sellan	6	míþan	3	nágan	1
biddan	23	secgan	11	sícan	6	murnan	3	nyllan	1
fón	23	bannan	10	sígan	6	nytan	3	sculan	1
								spornan	1

TABULKA 3. Počty sloves přiřazených jednotlivým paradigmatickým vzorům

založeny na více zdrojích (předně Baker, 2012, a Wright — Wright, 1914) a jde v nich jak o jednoduché nahrazování znaků (např. záměnnost *u* a *v*), tak o komplikované nahrazování pomocí regulárních výrazů (např. kolísání zadních vokálů před nasálami). Výhodou práce s filtry až v této fáze je i jejich volitelnost z pozice uživatele — ten se např. může rozhodnout, že vzhledem ke stáří a nářečí textu chce využít jen některé, nebo žádné.

ID	form_i	BT	lemma	stem	form
240508	ætswummon	1014	æt-swymman	swymman	ætswummon
240509	ætswumon	1014	æt-swymman	swymman	ætswummon
240510	ætswumme	1014	æt-swymman	swymman	ætswumme
240511	ætswume	1014	æt-swymman	swymman	ætswumme
240512	ætswumme	1014	æt-swymman	swymman	ætswumme
240513	ætswume	1014	æt-swymman	swymman	ætswumme
240514	ætswummen	1014	æt-swymman	swymman	ætswummen
240515	ætswumen	1014	æt-swymman	swymman	ætswummen
240516	ætswamm	1014	æt-swymman	swymman	ætswamm
240517	ætswam	1014	æt-swymman	swymman	ætswamm
240518	ætswamm	1014	æt-swymman	swymman	ætswamm
240519	ætswam	1014	æt-swymman	swymman	ætswamm
240520	ætswummen	1014	æt-swymman	swymman	ætswummen
240521	ætswumen	1014	æt-swymman	swymman	ætswummen
240522	ætswimman	1014	æt-swymman	swymman	ætswimman
240523	ætswiman	1014	æt-swymman	swymman	ætswimman

TABULKA 4. Ukázka generovaných forem slovesa *æt-swymman*

Výsledná analýza dokladu *andswarodon* pak vypadá např. takto:

ANDSWARODON
- ***and-swarian*** (infl. like *sealfian*)
  - **verb**, Pl. Ind. Pret. (w, 2, a, *and-sw-a-r-o-od-on*)[7]

4.3 VÝSLEDKY

Výsledky nástroje byly testovány na úryvcích deseti staroanglických textů v celkové délce přibližně 2 500 slov. Jednotlivé úryvky byly vybrány tak, aby bylo zaručeno pokrytí různých nářečí, žánrů, míst a období vzniku (viz *tabulka 5*).

Jak jsme již uvedli v části popisující teoretický rámec, účelem nástroje je pouze samotná analýza, nikoliv disambiguace výsledků analýzy. Měřítkem úspěšnosti nástroje je pro nás tedy především pokrytí (*recall*) spíše než přesnost (*precision*). Pokrytí ukazuje, jaké části slov/dokladů z analyzovaného textu byly analýzou přiřazeny správné morfologické informace. Neukazuje však, kolik alternativních avšak chybných řešení k jednotlivým dokladům analýza dále přiřadila. Ač přesnost bude zásadní pro případnou disambiguaci, a ne přímo pro náš nástroj, čím méně alternativních analýz náš nástroj nabídne při zachování správné analýzy, o to bude následná disambiguace snazší.

7 Jde tedy o indikativ plurálu préterita slabého slovesa 2. třídy podtypu a, v základním slovníkovém tvaru *and-swarian*, časovaného dle vzoru slovesa *sealfian*.

formParts	var	prob.	function	wright	pEx	pID	wclass	c1	c2	c3
0-æt-sw-u-mm-0-on	0		painpl	Null	bindan	10	verb	s	3	a
0-æt-sw-u-mm-0-on	0	1	painpl	null	bindan	10	verb	s	3	a
æt-sw-u-mm-0-e	0		PaInSg2	null	bindan	10	verb	s	3	a
æt-sw-u-mm-0-e	0	1	PaInSg2	null	bindan	10	verb	s	3	a
æt-sw-u-mm-0-e	0		PaSuSg	null	bindan	10	verb	s	3	a
æt-sw-u-mm-0-e	0	1	PaSuSg	null	bindan	10	verb	s	3	a
æt-sw-u-mm-0-en	0		PaSuPl	null	bindan	10	verb	s	3	a
æt-sw-u-mm-0-en	0	1	PaSuPl	null	bindan	10	verb	s	3	a
0-æt-sw-a-mm-0-0	0		painsg1	null	bindan	10	verb	s	3	a
0-æt-sw-a-mm-0-0	0	1	painsg1	null	bindan	10	verb	s	3	a
æt-sw-a-mm-0-0	0		PaInSg3	null	bindan	10	verb	s	3	a
æt-sw-a-mm-0-0	0	1	PaInSg3	null	bindan	10	verb	s	3	a
0-æt-sw-u-mm-0-en	0		papt	null	bindan	10	verb	s	3	a
0-æt-sw-u-mm-0-en	0	1	papt	null	bindan	10	verb	s	3	a
0-æt-sw-i-mm-0-an	0		if	null	bindan	10	verb	s	3	a
0-æt-sw-i-mm-0-an	0	1	if	null	bindan	10	verb	s	3	a

Pokrytí v testovaných úryvcích bylo v průměru 95 %, po zapnutí variačních filtrů stoupla na 97,3 %, tedy jen u 2,7 % nebyly analýzou přiřazeny správné morfologické informace. Jak vyplývá z *tabulky 5*, nejlepší pokrytí nástroj dosahuje v prozaických textech západosaské provenience přibližně z roku 1000. Takový výsledek odpovídá našim očekáváním, protože právě tyto texy jsou nejblíže varietě, na které je založena mluvnice a do značné míry i BT (byť části slovníku jsou orientovány spíše na mladší a špatně doložené západosaské období). Problémy naopak nástroji způsobují texty neprozaické a především nářeční. Proto také na nářečí nejlépe reagují variační filtry, jak je patrné z výsledků pro *Rushworth Gospels*.

5. POUŽITÉ TECHNOLOGIE

Technologicky jsou první dvě fáze zpracovány pomocí skriptů v jazyce Perl, který pracuje s daty v textových souborech v kódování Unicode. Předpokládá se, že tyto skripty spouští pouze administrátor nástroje v případě, že dojde ke změně ve zdrojových datech (slovníku či pravidlech). Naopak třetí fáze je zpracována do podoby webové aplikace programované v jazyce PHP a s daty v relační databázi MySQL, jelikož tvoří samotné prostředí pro koncové uživatele.

Uživatelské prostředí v tuto chvíli ještě není veřejně přístupné, hostováno je však společně se slovníkem BT online na serveru Filozofické fakulty Univerzity Karlovy. Zájemci o přístup se mohou obracet na autora.

název	popis	žánr	datace	nářečí	pozic	pokrytí	pokrytí s filtry
House on the Rock (or the Parable of the Wise and the Foolish Builders)	West Saxon Gospels, Matthew 7:24–27	náboženská próza	990	Západosaské	100	100%	100%
Wise and Foolish Virgins (or the Parable of the Ten Virgins)	West Saxon Gospels, Matthew 25:1–13	náboženská próza	990	Západosaské	194	100%	100%
The Voyages of Ohthere and Wulfstan	from Historiarum Adversum Paganos Libri VII	cestopis v próze	890	Západosaské	347	99,4%	99,4%
Cynewulf and Cyneheard	entry for 754 AD in the Anglo-Saxon Chronicle (Parker MS)	kronika v próze	890	Západosaské	313	99%	99%
Bede's „The Sun and the Moon"	from Ælfric's translation of Bede's De Temporibus	odborná próza	1000	Západosaské	146	98,6%	98,6%
Wulfstan's Sermo Lupi ad Anglos	from Wulfstan's sermon	kázání v próze	1015	Západosaské (sepsáno v Northumbrii)	320	97,1%	97,8%
Beowulf	lines 702–757	hrdinská poezie	ca. 1000	směs	313	96,8%	98,4%
Rushworth Gospels	interlinear gloss to Latin text of Matthew 6:1–6	náboženská próza	late 10c	Mercijské	167	94,6%	97%
Lindisfarne Gospels	interlinear gloss to Latin text of Matthew 6:1–6	náboženská próza	10c	Northumbrijské	183	89,6%	90,7%
Oswulf's Charter	charter	právní	806	Kenstké	471	84,7%	94,2%
Celkem					2554	95%	97,3%

TABULKA 5. Výsledky analýzy na úryvcích staroanglických textů

6. DISKUSE A ZÁVĚR

Jak je patrné z tabulky 6, celkový počet vygenerovaných forem je vzhledem k počtu lemmat vysoký: ca. 71 tis. lemmat odpovídá téměř 12 mil. flektovaných forem. Tato skutečnost je dána dvěma faktory.

Z praktického hlediska nebylo příliš třeba řešit obávaný problém autorů podobných nástrojů, tedy tzv. přegenerování. Velkou výhodou materiálu určeného pro náš nástroj totiž je, že pro starou angličtinu existuje vyčerpávající staroanglický korpus (Healey et al., 2000) a nelze očekávat, že by se v budoucnosti objevily další rozsahem významné doklady tohoto období. Je proto možné srovnat všechny vygenerované formy se všemi doklady korpusu a tím se snadno zbavit všech nedoložených („přegenerovaných") forem. Proto jsme si mohly dovolit řešit část výjimek či víceznačností staroanglické morfologie prostým vygenerováním dalších teoreticky možných forem, aniž bychom hleděli na jejich případnou doložitelnost.

Druhým důvodem je samotný stav staroanglické morfologie. Jak je opět patrné z tabulky 6, největší počet forem na jedno lemma vykazují adjektiva a adjektivně skloňovaná participia, což je dáno tím, že každé adjektivum lze ve staré angličtině skloňovat min. dle dvou vzorů (slabého a silného) a to ve všech třech stupních. Tím dramaticky roste počet možných forem, byť je to právě adjektivní morfologie, která je jinak formálně nejslabší. Tedy z průměrných 300 vygenerovaných forem na jedno adjektivum připadá v průměru pouze 47 skutečně formálně rozlišených (unikátních) forem. V těchto počtech se tedy dobře ukazuje formální „slabost" staroanglické adjektivní morfologie, což odpovídá i jejímu velmi brzkému zániku.

Problém přegenerování je tedy palčivý především s ohledem na počty doložených, ale homonymních forem, které bude potřeba následně disambiguovat. V průměru jsou ke každému analyzovanému dokladu přiřazeny ca 2 lemmata a několik desítek možných morfologických interpretací (konkrétní čísla se liší dle slovních druhů, viz výše).

Přes takto vysoké počty vygenerovaných forem zůstává malé procento (viz Výsledky) dokladů neurčeno, resp. určeno nesprávně. Existují tedy i doložené formy, které nástroj negeneruje. Většinu z nich pokrývají variační filtry, jelikož generovat všechny možné nářeční podoby a podoby odpovídající všem dobovým způsobům zápisu by bylo neefektivní (přegenerování by pak bylo o několik řádů vyšší a slovník forem by byl i z praktického hlediska špatně použitelný). Existují ale jak nedoložené potenciální formy (tedy gramatické formy v nepravděpodobných kombinacích pravopisných, diachronních a nářečních variet), tak formy doložené, které bychom jako potenciální z hlediska našeho nástroje neoznačily, neb jsou z hlediska slovníku, pravidel a popsaných variet nepředvídatelné (např. chyby). Takové formy budou muset být ve fázi disambiguace označkovány ručně.

Disambiguace samotná může probíhat zčásti strojově (pomocí syntaktických pravidel či za pomoci tréninkových dat a strojového učení), nápomocná v tomto procesu mohou být také data přiřazená k alternativním formám během generování, která mohou naznačit pravděpodobnost, že se daná forma vyskytne mezi doklady. Čím více kombinací předgenerovaných variací a variačních filtrů je využito, tím je forma pokládána za méně pravděpodobnou. Podobně některé morfologické tvary pokládáme již na základě příruček v určitém tvaru za méně pravděpodobné. Jistou roli ale bude

slovní druh	lemmat	forem	forem / slov	unikátních forem	u.f. / slov
adjektivum	6 235	1 882 110	301,9	294 349	47,2
adverbium	1 863	16 640	8,9	15 652	8,4
spojka	54	61	1,1	55	1,0
citoslovce	25	26	1,0	26	1,0
substantivum	19 835	259 496	13,1	140 837	7,1
maskulinum	8 668	93 874	10,8	59 251	6,8
femininum	7 306	122 508	16,8	57 133	7,8
neutrum	4 067	43 114	10,6	27 477	6,8
číslovka	110	3 344	30,4	773	7,0
předložka	148	160	1,1	153	1,0
zájmeno	43	4 996	116,2	1 010	23,5
sloveso	6 959	561 736	80,7	395 506	56,8
slabé	5 094	461 239	90,5	325 768	64,0
silné	1 786	103 313	57,8	74 128	41,5
préteritoprésentní	32	1 245	38,9	813	25,4
nepravidelné	59	2 481	42,1	1 569	26,6
participium	35 642	8 715 027	244,5	1 247 483	35,0
minulé	12 959	2 930 381	226,1	433 783	33,5
přítomné	22 683	5 784 646	255,0	814 928	35,9
Celkem (prům. poměr)	**70 914**	**11 443 596**	**161,4**	**2 095 844**	**29,6**

TABULKA 6. Počty vygenerovaných forem po slovních druzích v porovnání k počtům lemmat

muset při disambiguaci sehrát i ruční práce zkušených filologů, i kdyby jen kvůli morfologicky neoznačkovaným, nebo špatně označkovaným dokladům.

Vzhledem k nezbytnému a předpokládanému kroku poloautomatické disambiguace pokládáme výsledky našeho nástroje za úspěšné, byť předpokládáme, že právě z této následné další fáze nutné pro kvalitně označkovaný text vyplyne řada poznatků, které zpětně pomohou náš nástroj dále zlepšovat.

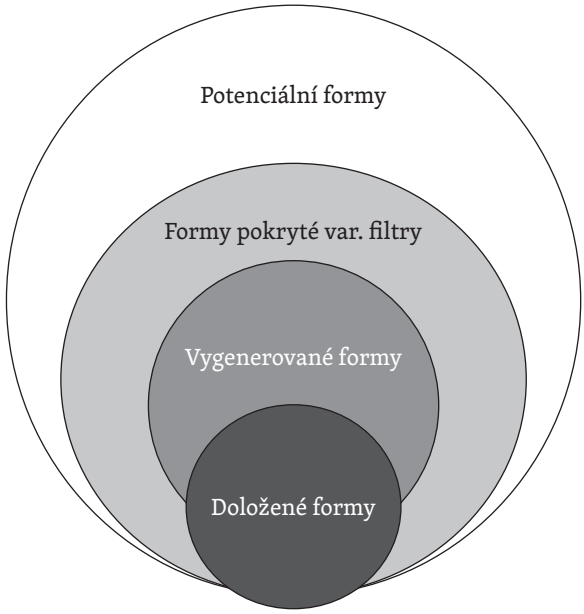
LITERATURA

Adams, J. (2007): Retrieved from *A Morphological Analyzer for Old English Verbs*: http://www.cs.cmu.edu/~jmadams/NLPLabProject.pdf

Baker, P. S. (2012): *The Electronic Introduction to Old English*: http://www.wmich.edu/medieval/resources/IOE/

Bosworth, J. (1921): *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of the Late Joseph Bosworth.* (T. N. Toller, Ed.). Oxford: Clarendon Press.

Calle Martín, J. — Triviño Rodríguez, J. L. (1998): Algoritmos de derivación de



OBRÁZEK 1. Překryv množiny potenciálních, vygenerovaných a doložených forem

palabras con ortografía irregular en el análisis morfológico automático del Inglés Antiguo. *Interlingüística*, s. 67–70.

Campbell, A. (1983): *Old English Grammar*. Oxford: Clarendon Press.

Čermák, J. — Znojemská, H. (2001): *Čítanka staroanglických, středoanglických a raně novoanglických textů*. Praha: Nakladatelství Karolinum.

Hajič, J. (2004): *Disambiguation of Rich Inflection*. Prague: The Karolinum Press.

Hall, J. R. (1894): *A Concise Anglo-Saxon Dictionary*. London: Swan Sonnenschein & Co.

Healey, A. d. (Ed.). (2014): *Dictionary of Old English A to G Online*. http://www.doe.utoronto.ca/

Healey, A. d. — Holland, J. — McDougall, I. — Mielke, P. (2000): *The Dictionary of Old English Corpus in Electronic Form*. Toronto: University of Toronto.

Hogg, R. M. — Fulk, R. D. (2011): A Grammar of Old English: Morphology. Wiley-Blackwell.

Kay, C. — Edmonds, F. — Roberts, J. — Wotherspoon, I. (2005): *A Thesaurus of Old English*. Amsterdam: Rodopi.

Kiernan, K. (2011): *Electronic Beowulf*. London: British Library.

Kleinman, S. (2012): *A demo version of the OE lemmatiser*. Retrieved from http://www.csun.edu/english/lemmatise/main.php

Lindberg, T. (2014, February 7): *Verbix*: http://www.verbix.com/

Martín Arista, J. (2014, February 7): *Nerthus Project*: http://www.nerthusproject.com/

McGillivray, M. (2014): Cynewulf and Cyneheard. *Old English Texts*: http://www.ucalgary.ca/UofC/eduweb/engl401/texts/ohthfram.htm

McGillivray, M. — Chevallier, G. (2014): The Voyage of Ohthere. *Old English Texts*: http://www.ucalgary.ca/UofC/eduweb/engl401/texts/ohthfram.htm

Miranda García, A. — Calle Martín, J. — Moreno Olalla, D. — Muñoz González, G. (2006): The Old English Apollonius of Tyre in the light of the Old English Concordancer. In: A. Renouf — A. Kehoe, *The Changing Face of Corpus Linguistics.* Amsterdam: Rodopi, s. 91–98.

Miranda García, A. — Triviño Rodríguez, J. L. — Calle Martín, J. (2000): MAOET: Morphological Analyser of Old English Texts. *Proceedings of the 10th International Conference of SELIM*. Zaragoza: Institución Fernando el Católico, s. 127–145.

Mitchell, B. — Robinson, C. F. (2001): *A Guide to Old English* (6th ed.). Oxford: Blackwell Publishing.

Osolsobě, K. (1996): *Algoritmický popis české formální morfologie a strojový slovník češtiny* (nepublikovaná disertační práce). Brno: Masarykova Univerzita.

Oxford University Press (2014): *OED Online*. http://www.oed.com/

Quirk, R. — Wrenn, C. L. (1957): *An Old English Grammar*. London: Routledge.

Rissanen, M. — Kytö, M. — Kahlas-Tarkka, L. — Kilpiö, M. — Nevanlinna, S. — Taavitsainen, I. — Raumolin-Brunberg, H. (1991): *The Helsinki Corpus of English Texts*. Helsinki: University of Helsinki.

Sedláček, R. (1999): *Morfologický analyzátor češtiny* (nepublikovaná diplomová práce). Brno: Masarykova Univerzita.

Sedláček, R. — Smrž, P. (2001, June): Automatic Processing of Czech Inflectional and Derivative Morphology. *FI MU Report Series*.

Sweet, H. (1887): *A Second Anglo-Saxon Reader: Archaic and Dialectal*. Oxford: Calendon Press.

Taylor, A. — Warner, A. — Pintzuk, S. — Beths, F. (2003): *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*.

Taylor, A. — Warner, A. — Pintzuk, S. — Plug, L. (2001): *The York-Helsinki Parsed Corpus of Old English Poetry*.

Tichý, O. (2007): *Digitization of Old and Middle English Dictionaries* (nepublikovaná disertační práce). Praha: Univerzita Karlova, Filozofická fakulta.

Wright, J. — Wright, E. M. (1914): *Old English Grammar*. London: H. Milford, Oxford University Press.

Ondřej Tichý | Ústav anglického jazyka a didaktiky, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
ondrej.tichy@ff.cuni.cz