Filozofická fakulta Univerzity Palackého v Olomouci
Katedra obecné lingvistiky

Lingvistická anotace sémantických jevů

magisterská diplomová práce

Autor: Bc. Jiří Kozmér
Vedoucí práce: Mgr. Dalibor Pavlas

Olomouc
2019

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Lingvistická anotace sémantických jevů“ vypracoval samostatně a uvedl jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne 24. 4. 2019 Podpis

Tato diplomová práce vznikla v rámci projektu *Počáteční fáze tvorby korpusu metafory v češtině* (číslo grantu: IGA_FF_2018_026) financovaného Ministerstvem školství, mládeže a tělovýchovy České republiky.
Kapitola 3.2 je částečně založena na publikovaném článku *Applying MIPVU Metaphor Identification Procedure on Czech* (Pavlas, Vrabeľ, & Kozmér, 2018).

Poděkování

Děkuji svému vedoucímu práce za cenné rady a vstřícnost. Také bych chtěl poděkovat svým rodičům a prarodičům za bezmeznou podporu během studia a tvorby této práce.

Abstrakt

Název práce: Lingvistická anotace sémantických jevů
Autor práce: Bc. Jiří Kozmér
Vedoucí práce: Mgr. Dalibor Pavlas
Počet stran a znaků: 87 stran, 165 456 znaků
Počet příloh: 0
Abstrakt:

Cílem této diplomové práce je poskytnout ucelený přehled o možnostech lingvistické anotace s důrazem na anotaci sémantických jevů. Práce je rozdělena do tří větších celků, přičemž první část pojednává o obecných rysech lingvistické anotace (úrovně, vývoj, standardy, anotační proces a nástroje). Pozornost je věnována též metodologii tvorby anotačních schémat, statistickým metodám v anotaci a současným trendům na poli kolaborativní anotace. Zvláštní pozornost je věnována anotaci sémantických jevů, jejím specifikům a její důležitosti pro různé obory lingvistiky (NLP, korpusová lingvistika). Ve druhé části jsou na příkladech deseti vybraných projektů představena některá současná řešení v oblasti anotace sémantických jevů. Závěrečná část je věnována konkrétnímu lingvistickému výzkumu, na kterém autor práce participoval. Projekt se zabývá adaptací procedury anotace lingvistické metafory MIPVU na český jazyk.

Klíčová slova: lingvistická anotace, sémantické jevy, komputační lingvistika, NLP, metafora, MIPVU

Abstract

Title: Linguistic Annotation of Semantic Features
Author: Bc. Jiří Kozmér
Supervisor: Mgr. Dalibor Pavlas
Number of pages and characters: 87 pages, 165 456 characters
Number of appendices: 0
Abstract:

The aim of this master's thesis is to offer a comprehensive overview of linguistic annotation with special focus on the annotation of semantic features. The work is divided into three bigger parts. The first part deals with linguistic annotation in general terms (annotation levels, development, standards, annotation process and tools). Moreover, phenomena such as annotation scheme design methodology, annotation statistics and contemporary trends in collaborative annotation are dealt with. Special attention is paid to the annotation of semantic features, its specifics and importance for various linguistic fields (NLP, corpus linguistics). The second part introduces ten selected projects dealing with semantic annotation and illustrates some of the contemporary methods and solutions in this field. The final part is dedicated to a linguistic research in which the author of this thesis participated. The research project deals with the modification of MIPVU linguistic metaphor identification procedure for Czech.

Keywords: linguistic annotation, semantic features, computational linguistics, NLP, metaphor, MIPVU

Obsah

Úvod

Lingvistická anotace byla od 70. let spojována především s korpusovou lingvistikou – manuální anotace prvních korpusů otevřela zcela nové možnosti studia jazyka a ověřování lingvistických teorií. Technologický pokrok na poli výpočetní techniky, jenž naplno propukl na přelomu 80. a 90. let, umožnil oproti předchozím dekádám zpracování nesrovnatelně větších objemů jazykových dat a odstartoval i vývoj automatických metod lingvistické anotace. Lingvistická anotace se díky tomu dostala do centra zájmu NLP *(Natural Language Processing)*, protože anotovaná jazyková data jsou základním předpokladem pro trénování a evaluaci technologií a nástrojů zpracovávajících přirozený jazyk. S množstvím projektů vznikajících v tomto období vyvstala i tendence zabývat se anotací na teoretické rovině, což zahrnovalo například sdílení osvědčených postupů, zavádění standardů a snahy o interoperabilitu. Na počátku 21. století lze již lingvistickou anotaci bez nadsázky považovat za plně etablovanou vědeckou disciplínu (též nazývanou *annotation science*), která za padesát let své existence (v dnešním, komputačně-lingvistickém slova smyslu) urazila velký kus cesty a která bezpochyby zasluhuje více pozornosti než jen kapitolu v knihách o korpusové lingvistice. Svědčí o tom mimo jiné i celosvětová vědecká komunita, pravidelně pořádané konference, ISO standardizace a množství publikací věnovaných anotační teorii.

Cílem této práce je představení lingvistické anotace v kontextu komputační lingvistiky s důrazem na lingvistickou anotaci sémantických jevů. Práce je strukturována od obšírněji pojatého teoretického základu ke konkrétnímu příkladu použití lingvistické anotace v praxi a je rozdělena do třech větších celků.

První kapitola poskytuje ucelený přehled o lingvistické anotaci jako vědecké disciplíně od jejího vývoje, přes standardizaci, metodologii tvorby anotačních schémat, statistické metody používané v anotaci až po její technické aspekty a anotační nástroje. Pozornost je věnována také crowdsourcingu v lingvistické anotaci. Druhá kapitola na příkladu deseti projektů zaměřených na anotaci vybraných sémantických jevů představuje některé současné postupy a řešení v oblasti anotace sémantiky. Projekty jsou představeny v kontextu teorie obsažené v první kapitole. Třetí, a zároveň závěrečná kapitola, detailně pojednává o lingvistickém výzkumu, na němž se autor práce podílel – o modifikaci anotačního protokolu MIPVU za účelem anotace lingvistické metafory v českém jazyce.

1 Lingvistická anotace

1.1 Typy anotace

Anotací se z pohledu komputační lingvistiky rozumí proces, při němž se ke korpusovým datům přiřazují další rozšiřující či interpretační údaje, jinými slovy metadata. Takto anotovaná jazyková data najdou v současné době využití především v korpusové lingvistice a v NLP, kde jsou cenným materiálem při vývoji algoritmů strojového učení.

V korpusové lingvistice se obecně rozlišují tři základní typy anotace. Prvním typem je administrativní anotace, jež je také nazývána anotací vnější, případně metatextovou. Při tvorbě a vnitřní organizaci korpusu slouží k evidenci údajů o vkládaném textu (v případě psaného materiálu např. autorství, žánr, rok vydání, zdroj atp.).

Druhým typem je strukturní anotace zachycující formátování textu. Slouží také k hierarchické segmentaci textů na dílčí celky. V případě textových korpusů se povětšinou jedná o párové tagy vymezující jednotlivé texty, kapitoly, odstavce a věty. Korpusy opatřené administrativní a strukturní anotací umožňují uživatelům prostřednictvím korpusových vyhledávacích nástrojů efektivní a rychlou práci s korpusem.

Třetím typem anotace je anotace lingvistická (vnitřní), jíž se tato práce primárně zabývá. Při ní se k analyzovaným segmentům v korpusu (většinou slovním tvarům) přiřazují lingvistické interpretační údaje, tj. morfologické, syntaktické, sémantické, popř. další informace. Jak zdůrazňují McEnery & Hardie (2014), lingvistická anotace do korpusu nepřidává nová data – jejím primárním účelem je explikovat ve výchozím textu informace, které mohou být na první pohled skryté a text tak obohatit (31).

Co se provedení lingvistické anotace týče, existují tři přístupy: plně manuální anotace, automatizovaná (před)anotace s následnou manuální korekcí a plně automatická anotace. Žádný z uvedených přístupů negarantuje naprosto bezchybný výstup. Obecně nejlepších výsledků dosahuje plně manuální anotace (i když ani lidský úsudek není samozřejmě bezchybný). Jak bude ovšem podrobněji vysvětleno, jedná se o činnost finančně a časově náročnou, a tak je používána především jako základ k tvorbě a trénování nástrojů sloužících k automatické anotaci, které sice zpravidla nedosahují 100 % přesnosti a kvality lidských anotátorů, ale zato jsou nesrovnatelně rychlejší, výkonnější, ekonomičtější a konzistentnější. Pro automatickou anotaci prováděnou softwarovými nástroji se také používá označení *tagování* či *značkování*.

Potenciální míra chybovosti jakékoliv anotace se navíc přímo odvíjí od charakteru anotovaného fenoménu – například slovnědruhová anotace (PoS tagging) se striktně definovanými kritérii poskytuje relativně malý prostor pro různé individuální interpretace, kdežto v případě sémantické anotace je mnohdy možno akceptovat více různých interpretací, což představuje poměrně velkou překážku pro automatické tagování sémantických jevů.

Co se typu anotovaných dat týče, nemusí se jednat výhradně o korpusy textové, ale například i o korpusy mluvené, videokorpusy nebo obecně korpusy multimodální. Je možné se setkat i s lingvistickou anotací prozodie, gestiky, mimiky, fonetických a fonologických jevů atp.

Tokenizace

Předpokladem k provedení lingvistické anotace jazykového materiálu v korpusu je identifikace a vymezení jednotek, jež mají být anotovány a následně zpracovávány či podrobeny analýze. Volba minimálních jednotek závisí na zaměření a konkrétních potřebách projektu, v případě textových korpusů se může jednat o rozsáhlé textové celky, ale např. i o morfémy (Ide et al., 2016). Velmi často jsou minimální jednotkou tokeny, tedy jednotlivá grafická slova textu (včetně interpunkčních znaků a čísel). Proces, pomocí nějž jsou tokeny vymezeny, se nazývá tokenizací, a je v dnešní době prováděn automaticky pomocí specializovaného software, tzv. tokenizéru.

Jak uvádí Schmid (2008), u jazyků s alfabetickým zápisem je tokenizace podstatně snazší záležitostí než u jazyků ideografických, v jejichž zápise nejsou jednotlivá slova jasně ohraničena (tokenizace jazyků tohoto typu, mezi něž patří například čínština, je z důvodu zcela odlišného přístupu a metodologie samostatnou disciplínou). Nejprimitivnější způsob automatické tokenizace alfabetického jazyka využívá skutečnosti, že slova jsou oddělena mezerami (tento způsob tokenizace je také znám pod názvem *white space tokenization*). Následně postačí odstranit interpunkční znaménka vyskytující se na počátcích a koncích slov. Výraznější komplikaci mající vliv na celkovou přesnost segmentace tokenizéru představují například tečky ve zkratkách a řadových číslovkách. Některé zkratky vyskytující se na konci věty mohou

mít stejný tvar jako plnohodnotné slovo[1]. Desambiguace slov vyskytujících se s tečkou proto vyžaduje informace o kontextu. Další problematickou záležitostí vyžadující desambiguaci jsou například víceslovná spojení, enklitika nebo spřežky. Samostatnou disciplínu v oblasti tokenizace představuje rozpoznávání pojmenovaných entit (Named Entity Recognition – NER), jinými slovy identifikace proprií.

Lemmatizace

Jednou ze základních forem lingvistické anotace je lemmatizace, při níž je ke každému slovnímu tvaru v korpusu přiřazen jeho základní slovníkový tvar neboli lemma. Jedná se o množinu všech forem lišících se tvaroslovnými afixy, případně pravopisnou variantou (Cvrček, 2014). Lemmata sloves jsou tedy v českém prostředí jejich infinitními tvary, u substantiv se jedná o tvar nominativu singuláru atp.

Lemmatizace může být provedena několika různými metodami. Gries & Berez (2017) uvádějí jako základní způsob použití existující databáze lemmat, která jsou následně přiřazena jednotlivým tokenům. Takovýto způsob lemmatizace je také nazýván lemmatizací slovníkovou. Pokročilejší metodou je tzv. stemming (v českém prostředí se používá i pojmu stematizace). Jedná se o automatický přístup, při jehož aplikaci jsou ze slova odstraněny morfologické afixy s cílem odhalit kmen slova. Ve stematizaci se uplatňuje celá řada algoritmů: brute force algoritmy, suffix stripping nebo stochastické algoritmy. V současné době je však nejrozšířenější hybridní přístup kombinující různé algoritmy se slovníky pokrývajícími výjimky a specifika daného jazyka (Chmelař et al., 2011).

PoS tagging (slovnědruhové tagování)

PoS tagging, nebo také morfosyntaktická anotace, je proces, při němž je slovním tokenům přiřazena informace o jejich příslušnosti ke slovnímu druhu doplněná ve většině případů i o další gramatické kategorie. PoS tagging je vůbec nejčastější a nejvyužívanější formou lingvistické anotace, a to z toho důvodu že, na něm přímo

[1] Například česká zkratka *bud.* (budoucí čas) může být na konci věty zaměněna s plurálem genitivu substantiva *bouda.*

závisí další vrstvy anotace (parsing, různé formy sémantické anotace, ale například i lemmatizace, u níž je určení slovnědruhové příslušnosti klíčové při desambiguaci).

Parsing

Syntaktická analýza neboli parsing je po PoS taggingu v pořadí dalším nejobvyklejším způsobem anotace. Jedná se o syntaktickou analýzu větných struktur v textu, jež jsou opatřeny syntaktickými tagy. Automatický parsing je prováděn pomocí softwarových nástrojů – parserů (syntaktických analyzátorů). Parsery mohou využívat statistických metod (stochastické parsery) nebo se řídí souborem předem definovaných pravidel (Cvrček, 2016).

Pro parsované korpusy se používá označení treebanky. Ty se pak typologicky liší v závislosti na tom, na základě jaké teorie jsou vystavěny (to může být ovlivněno například lingvistickou tradicí následovanou tvůrci nebo čistě pragmatickými důvody jako je vhodnost konkrétní teorie pro zpracování daného jazyka). Dvěma převládajícími typy treebank tedy jsou:

- **treebanky anotované na frázovou strukturu (bezprostředněsložkové)** Myšlenka analýzy na bezprostřední složky byla poprvé zmíněna americkým lingvistou Leonardem Bloomfieldem (1933) a proslavila se především díky dílu Noama Chomského. Proto je také běžnější v angloamerickém prostředí. Příklady takovýchto treebank jsou Penn Treebank nebo ICE-GB (International Corpus of English).
- **treebanky anotované na dependenční strukturu (závislostní)** – Tento způsob syntaktické analýzy je založen na dependenční gramatice francouzského lingvisty Luciena Tesnièra a má tak blíže evropskému strukturalismu. Mezi takto anotované treebanky patří například česká Prague Dependency Treebank (Hajič et al., 2018).

Další anotované jevy

Za zmínku stojí i některé specializované projekty zaměřené na anotaci dalších lingvistických jevů. Jedná se například o anotaci chyb v textech jedinců osvojujících si cizí jazyk. Jedním z takových projektů je ICLEv2[2] (Granger, 2009), jehož autoři vytvořili speciální tagger, který chyby nejen identifikuje, ale i kategorizuje podle oblasti, do níž spadají (gramatika, slovní zásoba, interpunkce, funkční styly, slovosled atp.). Takto anotované korpusy jsou velmi cenným zdrojem informací pro tvůrce výukových materiálů, a to především při tvorbě lokalizovaných učebnic cizích jazyků kladoucích důraz na specifické obtíže způsobené vlivem mateřského jazyka studentů.

Anotovány mohou být i korpusy obsahující data odborného charakteru, kupříkladu lékařské zprávy. Tato data mohou být anotována za účelem vývoje pokročilých NLP aplikací umožňujících z těchto textů vyextrahovat např. informace o diagnóze pacienta. Prvním krokem musí v tomto případě ovšem být anotace výchozího textu člověkem z oboru medicíny, u nějž nelze automaticky předpokládat, že bude mít zároveň lingvistické vzdělání. Pro tyto případy byly vyvinuty speciální anotační modely nazvané *light annotation tasks* (Stubbs, 2013), umožňující efektivně zaznamenat anotaci odborníků na danou problematiku tak, aby následně mohla být dodatečně zpracována lingvisty na míru konkrétního NLP projektu.

1.2 Standardy v lingvistické anotaci

Standardizace v oblasti jazykových zdrojů se týká především dvou zásadních oblastí. První z nich je formát (syntax), v němž jsou data uchovávána, druhou pak datové kategorie sloužící k identifikaci anotovaných lingvistických jevů. Obě zmíněné oblasti mají svá specifika a potenciální komplikace, lze ovšem říci, že standardizace lingvistických datových kategorií představuje podstatně větší problém spočívající mimo jiné v mnohdy subjektivních rozdílech v definicích či v příslušnosti k různým teoretickým směrům (Ide et al., 2017). V následujících odstavcích této sekce bude nastíněno, jakými kroky se celosvětová anotátorská a lingvistická komunita snaží standardizace v těchto dvou oblastech docílit.

[2] (International Corpus of Learner English) – korpus tvořený texty žáků EFL šestnácti různých mateřských jazyků obsahující 3,7 milionu slov

První snahy o standardizaci přístupů k elektronickému zpracování jazykových dat, tedy i k lingvistické anotaci, lze zaznamenat již v polovině 80. let. Nástup výpočetní techniky na poli humanitních věd přispěl k rychlému rozvoji komputační a korpusové lingvistiky. Jedním z úskalí tohoto rychlého rozvoje byla ovšem skutečnost, že nezávisle na sobě vznikalo nepřeberné množství menších komputačně-lingvistických projektů, z nichž každý používal vlastní software vyvinutý specificky pro vlastní potřeby. To nutně vedlo k tomu, že jednotlivé projekty spolu byly nekompatibilní a že data z nich bylo velmi obtížné aplikovat pro další výzkum. Případný převod dat mezi různými formáty představoval extrémně časově a finančně náročnou záležitost a výjimkou nebyl ani vývoj zcela nového software pro zpracování již jednou použitých dat pro jiné účely. Jak uvádí Ide et al. (2017), k řešení problémů s kompatibilitou nepřispěla ani tehdejší situace na trhu se softwarem. Komerční vývojáři a prodejci výpočetní techniky mezi sebou soupeřili o své místo na trhu a v honbě za ziskem se jejich obchodní strategie často vyznačovala tím, že nabízené produkty byly omezeny výhradně na jejich vlastní platformu (s. 115). Hlavními prioritami na přelomu 80. a 90. let proto byla interoperabilita v oblasti softwaru a dat a rozvoj metod, jak již jednou použitá jazyková data využít k dalším účelům.

1.2.1 TEI (Text Encoding Initiative)

Za přelomovou událost ve standardizaci kódování textu v elektronické podobě se považuje založení organizace s názvem Text Encoding Initiative (TEI) (Ide et al., 2017). Jejím hlavním posláním je poskytnout směrnice popisující způsoby kódování strojově čitelných textů v humanitních a společenských vědách a v lingvistice. První ustavující setkání TEI se odehrálo roku 1987 v Poughkeepsie ve státě New York za účasti delegátů z desítek různých zemí světa.

V rámci setkání byl formulován dokument s názvem *Poughkeepsie Principles*, jenž se stal základním kamenem pro pozdější *TEI Guidelines*, tedy soubor směrnic definujících strojově čitelný textový formát, který bude:

- vhodný k výměně dat a datové analýze
- nezávislý na software a hardware
- pečlivě definovat textové objekty
- snadno použitelný v praxi
- kompatibilní s již existujícími standardy

Autoři *Poughkeepsie Principles* předpokládali, že základem k plánovaným směrnicím by měl být značkovací metajazyk SGML (The Standard Generalized Markup Language), jenž byl roku 1986 oficiálně prohlášen za standard ISO ("Design Principles"). V polovině devadesátých let se však TEI Guidelines přeorientovaly na v té době nově nastupující značkovací jazyk XML (eXtensive Markup Language), jehož podoba je do značné míry ovlivněna právě TEI a jenž je v současné době nejrozšířenějším jazykem svého druhu. Jedná se o zjednodušenou verzi jazyka SGML a od jeho vzniku jej vyvíjí a spravuje organizace W3C.

První TEI Guidelines byly vydány roku 1994 a byly zatím třikrát aktualizovány – naposledy roku 2007. TEI působí dodnes jako nezisková organizace sdružující desítky akademických a výzkumných institucí po celém světě (patří mezi ně například i Ústav pro jazyk český) za cílem vyvíjet a udržovat standard pro reprezentaci textů v elektronické formě. TEI od roku 2001 každoročně pořádá konference, na nichž jsou prezentovány probíhající projekty, výsledky výzkumů a voleno vedení organizace. Konference jsou také místem, kde se zástupci komerční sféry a potenciální sponzoři mohou setkat s předními výzkumníky z oblasti lingvistiky a Digital Humanities (TEI).

1.2.2 CES (Corpus Encoding Standard)

Jednou z praktických aplikací TEI Guidelines je CES – Corpus Encoding Standard z roku 1994. Jedná se o optimalizaci TEI Guidelines pro lingvisticky anotované korpusy vytvořenou expertní skupinou EAGLES[3]. CES jasně definuje, jaká kritéria (kódování, architektura, specifikace pro lingvistickou anotaci) korpus musí splňovat, aby jej bylo možno považovat za standardizovaný. Nástupcem CES je XCES (Corpus Encoding Standard for XML), který vznikl roku 2000 současně s přechodem TEI Guidelines na XML. Jednou z klíčových vlastností CES a dalších specifikací zveřejněných projektem EAGLES je skutečnost, že při jejich vývoji bylo použito tzv. bottom-up přístupu – vznikaly na základě zkušeností nabytých při realizaci projektů MULTEXT[4] a MULTEXT-EAST (117). Tento přístup je diametrálně odlišný od

[3] Projekt EAGLES (Expert Advisory Group for Language Engineering Standards) vznikl roku 1993, byl financován z rozpočtu EU a jeho úkolem bylo vydávat standardy, směrnice a doporučení pro rozsáhlejší projekty pracující s jazykovými zdroji (korpusy, lexikony, komputačně-lingvistické projekty atp.)

[4] MULTEXT a MULTEXT-EAST je série projektů financovaných EU, při nichž byly v druhé polovině 90. let vyvíjeny jazykové nástroje a korpusy pro celkem 18 převážně evropských jazyků (TEI).

způsobu, jímž byly tvořeny TEI Guidelines, jejichž autoři specifikace vyvíjeli a priori, aniž by vycházeli z nějakého právě probíhajícího projektu (top-down přístup). Dalším specifikem CES je to, že vůbec poprvé doporučuje použití tzv. stand-off anotace (více v sekci 2.5.3), jež je, jak uvádí Ide (2017), v dnešní době považována za nejefektivnější způsob anotace.

1.2.3 ISO standardizace

Zásadním milníkem ve standardizaci lingvistické anotace je rok 2001, kdy v rámci Mezinárodní organizace pro normalizaci (ISO) vznikla specializovaná technická podkomise SC 4 – Language Resource Management. Podkomise funguje pod hlavičkou technické komise ISO/TC 37 – Language and Terminology a k roku 2018 vydala celkem dvacet dva ISO norem.

Vůbec prvním počinem SC 4 je norma ISO 24612:2012 známá též pod názvem *Linguistic Annotation Framework* (LAF). Cílem LAF je poskytnout abstraktní datový model sloužící jako základ všech ostatních standardů pro morfosyntaktickou, syntaktickou a sémantickou anotaci vzniklých v rámci SC 4. Oproti předchozím počinům (CES, XCES atp.) je u LAF kladen ještě větší důraz na univerzálnost, srozumitelnost, obecnost a interoperabilitu[5]. Samozřejmostí je již stand-off anotace, uchovávání jednotlivých úrovní anotace v samostatných dokumentech a reprezentace hierarchických informačních schémat (např. syntaktických stromů) výhradně prostřednictvím XML. Datový model LAF navíc kromě anotace textových materiálů počítá i s anotací ostatních typů médií jako například audia, videa nebo obrazového materiálu. Hlavními rysy jsou především jasně oddělená anotační struktura, tedy fyzický formát anotace a samotný obsah anotace (kategorie a popisky užívané k popisu lingvistických jevů).

[5] S interoperabilitou u LAF souvisí i snaha umožnit opakované zpracování již existujících a použitých dat prostřednictvím vícevrstvé anotace, slučování různých projektů atp.

Architekturu LAF tvoří dvě části:

1. Datová struktura umožňující zachytit vztahy mezi anotacemi a mechanismus přiřazující lingvistické kategorie příslušným částem této datové struktury.
2. Prostředky sloužící k definici lingvistických kategorií, které nejsou vázané na žádnou konkrétní lingvistickou teorii nebo konvenci (Ide & Sundermann, 2014, s.397).

K tomuto účelu (bod 2) byl vytvořen online repozitář lingvistické terminologie (DCR – Data Category Registry) používané v projektech pracujících s jazykovými zdroji nazvaný ISOcat. Jeho cílem je sjednotit označení konkrétních tagů napříč světovou anotátorskou komunitou a dosáhnout tak sémantické interoperability (Ide et al. 136). V rámci LAF mají autoři možnost pojmenovat tagy libovolným způsobem za předpokladu, že v dokumentaci projektu uvedou, jakým tagům v repozitáři jimi použité tagy odpovídají (Pustejovsky & Stubbs 2013).

Jakožto komunitní projekt se však ISOcat postupem času začal potýkat s hromaděním duplicitních nebo téměř totožných položek a kategorií, což snižovalo jeho použitelnost. Projekt byl proto rozdělen na dvě části – první, CLARIN Data Concept Registry[6], funguje pod taktovkou projektu EU CLARIN a obsahuje zjednodušený repozitář termínů, jehož editace je výhradně v dikci koordinátorů projektu. Druhá část je nadále spravována technickou komisí ISO/TC 37 a je de facto pokračováním ISOcat v původní podobě.

Datový model LAF jako takový byl do praxe uveden roku 2007 pod názvem GrAF (Graph Annotation Format). Jedná se o implementaci modelu prostřednictvím formátu XML tak, aby mohl sloužit jako páteřní (pivotní) mechanismus pomocí nějž by různá anotační schémata mohla být porovnávána, slučována nebo případně nahrazována (Ide & Sundermann, 2014)

Neméně důležitým rysem je požadavek, aby byly vždy explicitně uvedeny všechny informace týkající se anotace – v minulosti používaná schémata často počítala s některými více či méně akceptovanými úzy: Ide & Sundermann (2014) jako příklad uvádějí práci se závorkami, jejichž správná interpretace může činit značné obtíže, jelikož někteří autoři jich používají k označení alternativních prvků, ale jiní například k označení hierarchicky řazeného seznamu. Nesrovnalosti tohoto rázu pak představují velkou překážkou pro interoperabilitu.

[6] https://concepts.clarin.eu/

1.2.4 Další standardy související s anotací

Výše zmíněné standardy se týkají bezprostředně anotace a zpracování přirozeného jazyka. Existuje však mnoho dalších standardů a vnějších faktorů majících vliv na anotační projekty: V první řadě se jedná o způsob kódování dat, s nimiž se v projektu pracuje (jinými slovy způsob, jakým jsou znaky a symboly ze znakové sady reprezentovány číselně, aby mohly být zpracovávány počítačem). Před zavedením standardu Unicode bylo používáno mnoho různých kódovacích sad, jež se překrývaly (např. jeden číselný kód mohl reprezentovat dva různé znaky atp.), mnohdy byly navzájem nekompatibilní a nepokrývaly znaky všech světových jazyků. Standard Unicode[7] uvedený roku 1991 každému znaku přiřazuje unikátní číselnou hodnotu a pokrývá znakový repertoár naprosté většiny psaných světových jazyků. Navíc je nezávislý na platformě, operačním systému a používaném zařízení, což například v oblasti NLP a komputační lingvistiky usnadňuje celosvětovou spolupráci a interoperabilitu. Unicode je vyvíjen souběžně se standardem ISO/IEC 10646 definujícím univerzální znakovou sadu a podporuje tři způsoby kódování – UTF-8, UTF-16 a UTF-32 (liší se ve způsobu, jakým se data přenášejí, 8/16/32 bitů na kódovací jednotku). Pustejovsky & Stubbs (2013) za nejvhodnější variantu pro účely NLP považují variantu kódování Unicode UTF-8 *(Unicode Transformation Format)*, jelikož umožňuje zápis a čtení znaků většiny současných světových jazyků a je podporováno téměř všemi počítačovými platformami (znaky kódované tímto způsobem mají navíc stejné číselné hodnoty jako v původní ASCII kódovací tabulce).

Kromě kódování textu je záhodno věnovat pozornost i zvyklostem týkajícím se např. formátu zápisu dat (MM-DD-YYYY versus DD-MM-YYYY atp.), který se často liší v závislosti na regionu a jehož desambiguace vyžaduje potřebný kontext. Další potenciální překážku (v tomto případě např. při anotaci pojmenovaných entit – NE) mohou představovat zvyklosti týkající se pořadí zápisu jména a příjmení – v některých zemích, mezi něž patří kupříkladu Maďarsko, je běžné uvádět na prvním místě příjmení a až poté křestní jméno.

[7] http://www.unicode.org

1.3 Tvorba anotačních schémat

1.3.1 Anotační proces

Metodologii anotačního procesu včetně jeho dílčích částí se detailně věnují Pustejovsky & Stubbs (2013) a navrhují tzv. **MATTER** cyklus. Metodologie je primárně zaměřena na lingvistickou anotaci pro strojové učení.

Cyklus se skládá celkem z šesti fází:

Model – Hlavním cílem této fáze je vytvořit a definovat anotační schéma: jev, jenž má být anotován, je strukturně popsán, dostatečně podložen lingvistickou teorií a ze studia dostupných dat jsou vyvozeny jeho vlastnosti. Následně je specifikován fyzický formát anotace a je formulován způsob aplikace anotačního schématu na korpusová data. Fáze zahrnuje i samotné obstarání korpusu, v němž má být fenomén anotován (včetně zajištění jeho vyváženosti, reprezentativnosti, licenčních náležitostí atp.).

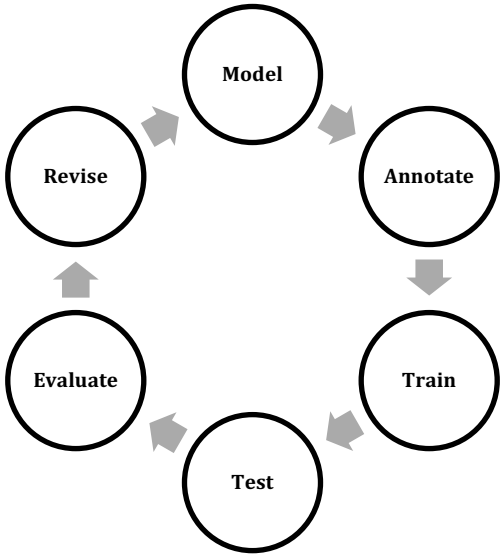
Anotace – V této fázi je korpus na základě výše zmíněných specifikací opatřen anotačními daty. Tomu však předchází celá řada důležitých kroků: data v korpusu jsou v první řadě normalizována a jsou odstraněny typografické chyby, je definován standardizovaný formát anotace spolu s metadaty a atributy vztahujícími se k anotovaným datům (viz 1.5.2). Následuje zaškolení anotátorů podle anotačního schématu a instrukcí (kritéria výběru anotátorů by měla vedle časové a finanční stránky věci přesně definovat požadované jazykové, popř. další odborné znalosti a dovednosti). Dále je podle zaměření a potřeb projektu zvolen vhodný anotační nástroj a anotátoři jsou instruováni v jeho používání. Anotátorům jsou poté dodána data k anotaci a podle předem nastaveného harmonogramu je zahájena samotná anotace, přičemž cílem je dosáhnout co nejvyšší hodnoty mezianotátorské shody (IAA), jež je v dílčích fázích projektu průběžně měřena a zaznamenávána. Anotátoři se pravidelně scházejí nad výsledky své práce, diskutují případné nesrovnalosti a často se vyskytující chyby a provádějí korekce. Cílem této fáze je dosáhnout takové kvality anotovaných dat, aby ji bylo možné prohlásit za tzv. gold standard (viz 1.4.2).

Trénování + **Testování** – Obě tyto fáze jsou v MATTER cyklu zaměřeny na vývoj algoritmů strojového učení. Přestože je tento způsob využití lingvistické anotace v současnosti velmi častý a významný, jedná se zároveň o využití značně specifické bereme-li v úvahu celou škálu dalších možností. Finlayson & Erjavec (2017) proto navrhují nahradit tyto dvě fáze obecnějším termínem ***Leverage*** – tedy ***Využití*** anotací získaných dat za jakýmkoliv blíže nespecifikovaným účelem (ať už se jedná např. o

měření výskytu lingvistických jevů, ověřování teorie nebo zmíněné strojové učení atp.).

Evaluace – fáze zahrnuje celkové zhodnocení a analýzu výsledné anotace, a to jak z kvalitativního, tak z kvantitativního hlediska – může se jednat např. o statistickou analýzu chybovosti.

Revize – tato fáze slouží v případě neuspokojivých výsledků evaluace projektu jako prostředek k návratu k některé z fází předcházejících. Pustejovsky & Stubbs (2013) navíc v rámci celého MATTER cyklu za podobným účelem počítají s opakujícím se sub-cyklem ***MAMA (Model-Annotate-Model-Annotate)***, který najde své využití především v počátečních fázích projektu: za účelem zvýšení celkové kvality anotačního schématu se po anotaci menších úseků dat provádějí úpravy samotného modelu. Po dostatečném odladění modelu je následně možno začít se naplno věnovat anotaci celého korpusu.



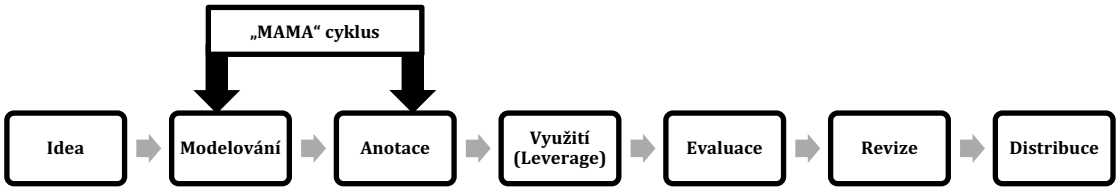
Obrázek 1: Schematické vyjádření MATTER cyklu podle Pustejovsky & Stubbs (2013)

Finlayson & Erjavec (2017) navíc MATTER cyklus rozšiřují o následující tři fáze:

Idea – tato fáze je autory řazena na úplný začátek a zahrnuje prvotní kroky předcházející tvorbě samotného modelu. Jedná se o formulaci cíle a zaměření projektu průzkum relevantní literatury, dále ověření, zdali se danou problematikou v minulosti již někdo jiný nezabýval a též průzkum již existujících korpusů a možností/výhodnosti jejich anotace na zvolený jev.

Zajištění (Procure) – tato fáze je vsazena mezi Modelování a Anotaci. Ve zkratce se jedná o nalezení vhodných anotačních nástrojů pro potřeby projektu (a v případě nutnosti i o modifikaci existujících nástrojů, případně vývoji zcela nových).

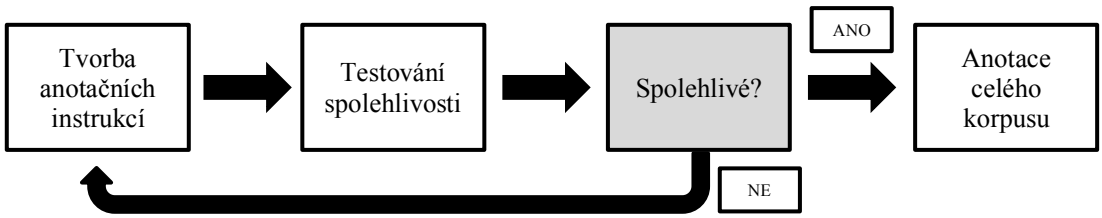
Distribuce – tento bod je řazen až na úplný konec MATTER cyklu a zahrnuje zveřejnění a distribuci výsledků projektu pro další využití komunitou (včetně exportu dat do formátu vhodného pro sdílení, licenčních podmínek, archivace dat a dokumentace v online repozitářích atp.).



Obrázek 2: Schematické vyjádření MATTER cyklu s rozšířením od Finlayson & Erjavec (2017)

1.4 Mezianotátorská shoda (IAA)

Anotační schéma lze označit za spolehlivé tehdy, když jej více různých anotátorů aplikuje na stejný úsek textu v korpusu a dosáhnou pokud možno co nejvyšší mezianotátorské shody (*Inter-Annotator Agreement* – IAA). Míra shody mezi anotátory poté určuje, jak je anotační schéma konzistentní a reprodukovatelné.[8] Jak ilustruje Artstein (2017) v následujícím schématu, standardním postupem při tvorbě anotačního schématu je nejprve tvorba anotačních instrukcí, jejichž spolehlivost je opakovaně testována více anotátory na vybraném vzorku. V rámci testování schématu je po každém kole anotace vždy měřena IAA, a pokud není dosaženo uspokojivých hodnot, přistupuje se k revizi anotačního schématu a instrukcí. Tento proces se opakuje do té doby, než je dosaženo cílové hodnoty IAA.



Obrázek 3: Postup tvorby anotačního schématu (Artstein, 2017)

[8] Dobrá reprodukovatelnost ovšem nijak nevypovídá o kvalitě výstupu anotačního schématu, pouze udává, že se anotátoři v anotaci shodují.

Zmíněná testovací fáze předpokládá zcela nezávislou práci všech anotátorů – výsledná anotace musí být produktem aplikace schémata na daný text, diskuse nad spornými případy je v této fázi kontraproduktivní. Kromě dodržování těchto požadavků anotátory je při testování také důležité zvolit reprezentativní vzorek (či vzorky) textu, v němž je sledovaný fenomén v dostatečné míře zastoupen.

1.4.1 Metody měření IAA

Podle zaměření a podoby projektu je třeba zvolit vhodnou metodu výpočtu IAA. Výpočet shody pouze na základě procentuálního vyjádření toho, kolikrát se anotátoři mezi sebou shodli na umístění tagu, nemá vypovídací hodnotu, protože neuvažuje možnost, že shoda může být v některých případech pouze náhodná (např. u projektů zaměřených na anotaci přítomnosti/nepřítomnosti daného fenoménu anotovaných dvojicí anotátorů je pak šance na náhodnou shodu poloviční). Z tohoto důvodu jsou v těchto případech aplikovány statistické metody výpočtů a koeficienty beroucí v potaz možnost náhodné shody – mezi nejrozšířenější v oblasti měření IAA v komputační a korpusové lingvistice patří Cohenovo κ (Kappa), Fleissovo κ a Krippendorfovo α (Alfa). První dvě metody jsou v následujících sekcích popsány podrobněji.

1.4.1.1 Cohenovo κ

Cohenovo κ je základním koeficientem uvažujícím možnost náhodné shody a slouží k měření IAA mezi dvěma anotátory. Rovnice pro výpočet má následující podobu (označení proměnných se v odborné literatuře liší, v této kapitole je použito označení používané J. Cohenem (Cohen, 1960)):

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

Hodnota p_o odpovídá pozorované (observed) shodě mezi anotátory, p_c pak očekávané (expected)[9] shodě, tedy případu, kdy by oba anotátoři anotovali zcela náhodně.

[9] Cohen označuje tuto proměnnou dolním indexem *c* podle slova *chance*: „the proportion of units for which agreement is expected by chance“ (Cohen, 1960).

V následujícím příkladu je Cohenovo κ použito k výpočtu IAA u anotace korpusu o velikosti 580 tokenů na metaforicky/nemetaforicky užitá slova (pomocí tagů dvojího druhu: MRW = metaphor related word a nonMRW). Data k výpočtu pocházejí z dílčí fáze anotačního projektu, jemuž je věnována teoretická část práce.

		Anotátor B	**Anotátor B**
		MRW	**nonMRW**
Anotátor A	**MRW**	72	5
Anotátor A	**nonMRW**	5	498

Tabulka 1: Data k modelovému výpočtu Cohenova κ

p_o, tedy pozorovaná shoda mezi anotátory, se vypočítá tak, že se sečtou počty případů, kdy oba anotátoři zároveň označili tokeny jako MRW/nonMRW a součet se vydělí celkovým počtem tokenů, tzn.:

$$\boldsymbol{p_o} = (72 + 498) / 580 = 0{,}983\ (98\%)$$

p_c, tedy očekávaná shoda, se vypočítá tak, že se určí, v kolika procentech případů každý z anotátorů použil konkrétní tag, načež se procentuální hodnoty obou navzájem vynásobí, aby se určilo, jak často by oba anotátoři použili stejný tag zároveň. Nakonec se obě vynásobené hodnoty sečtou:

A MRW: (72 + 5) / 580 = 0,133
B MRW: (72 + 5) / 580 = 0,133 → A MRW x B MRW = **0,018**

A nonMRW: (498 + 5) / 580 = 0,867
B nonMRW: (498 + 5) / 580 = 0,867 → A nonMRW x B nonMRW = **0,752**

$\boldsymbol{p_c}$ = 0,018 + 0,752 = **0,77** → $\kappa = \frac{0{,}983 - 0{,}77}{1 - 0{,}77} = \mathbf{0,92}\ (\mathbf{92\%})$

Výsledná hodnota Cohenovo κ je tedy v tomto případě 0,92. Jak výsledek interpretovat, je popsáno na konci této sekce.

1.4.1.2 Fleissovo κ

Fleissovo κ je oproti Cohenovu κ koncipováno tak, aby pomocí něj bylo možné vypočítat hodnotu IAA v případě anotace třemi a více anotátory – měření IAA pouze u dvou anotátorů v praxi většinou není považováno za dostatečné. Při větším počtu anotátorů je sice možné měřit IAA pro jednotlivé páry anotátorů a za výsledek prohlásit jejich průměr, ale obecně se doporučuje použít některou z generalizovaných verzí koeficientu umožňující počítat s více než dvěma anotátory již od samého počátku. Artstein & Poesio, (2008) poukazují na to, že označení κ (kappa) může být v případě Fleissova κ matoucí, protože koeficient má historicky blíže k jinému koeficientu, Scottovu π, než k Cohenovu κ (jedná se o rozšíření Scottova π, a tak se v literatuře lze setkat i s označením *multi-π*). Skutečností však zůstává, že výchozí rovnice Fleissova κ je v podstatě identická s rovnicí Cohenova κ (jedná se opět o výpočet podílu pozorované shody a očekávané shody). Rozdíl spočívá ve způsobu výpočtu hodnot obou shod (označení proměnných dle Fleisse (Fleiss, 1971)).

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Čitatel ve vzorci vyjadřuje stupeň skutečně dosažené shody nad úroveň shody náhodné, jmenovatel pak maximální možný stupeň shody dosažitelný nad úroveň náhody.

Nechť n = počet anotátorů, N = počet tokenů a k = počet kategorií tagů. První kategorie je opět MRW, druhá nMRW.

Nejprve se vypočítá p_j (poměr počtu tagů v j-té kategorii k celkovému počtu tagů):

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$$

Následně se vypočítá P_i, tedy míra shody mezi jednotlivými anotátory pro každý token (jinými slovy kolik dvojic anotátorů je ve shodě vztaženo k počtu všech možných dvojic anotátorů).

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}^2 \, (n_{ij} - 1)$$

Nakonec se vypočítají veličiny $\bar{P}$ a $\bar{P}_e$:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i \, \bar{P}_e = \sum_{j=1}^{k} p_j^2$$

a podle následujícího vzorce se vypočítá Fleissovo κ:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

V následujícím příkladu je použit stejný způsob anotace a korpus o 580 tokenech jako u předchozího příkladu na výpočet Cohenovo κ. Nyní je však v datech navíc přítomna ještě anotace třetího anotátora. Pro výpočet byl z důvodu velikosti datového souboru použit tabulkový procesor MS Excel, níže uvedená tabulka naznačuje postup výpočtu. Hodnoty proměnných jsou: $n = 3$, $N = 580$ a $k = 2$.

n_{ij}	**MRW**	**nonMRW**	P_i
1	0	3	1
2	0	3	1
3	3	0	1
4	0	3	1
12	2	1	0,333
41	1	2	0,333
579	0	3	1
580	0	3	1
celkem	**230**	**1510**	**564,659**
P_j	0,132	0,868	
P_j^2	0,017	0,753	
P_e	**0,771**		**P = 0,974**

Tabulka 2: Data k modelovému výpočtu Fleissova κ

$$\kappa = \frac{0{,}974 - 0{,}771}{1 - 0{,}771} = \mathbf{0,884}\ (\mathbf{88,4}\%)$$

1.4.1.3 Krippendorfovo α

Krippendorfovo α je svým charakterem podobné Fleissovu κ, s tím rozdílem, že zatímco κ přikládá všem neshodám mezi anotátory stejnou váhu, α obsahuje speciální metriku umožňující brát v potaz specifickou míru neshody mezi konkrétními páry tagů (Artstein, 2017). Využití tak najde v projektech, v nichž anotátoři udávají míru výskytu daného jevu na číselné škále (anotace sémantických jevů jako je humor, sentiment atp.) nebo při evaluaci výsledků strojového překladu atp.

1.4.1.4 Interpretace κ koeficientů

Problematika interpretace výsledných kappa koeficientů je od jejich uvedení předmětem diskusí a je nutno podotknout, že dosud neexistuje žádný univerzální klíč. Výsledná hodnota je silně ovlivněna celkovou charakteristikou projektu – autoři projektů zaměřených na anotaci teoreticky velmi precizně podložených a definovaných jevů (např. PoS nebo parsing) budou mít tendenci akceptovat pouze κ atakující hranici 1. Na druhou stranu, u projektů zaměřených na sémantickou anotaci bude vytyčená hranice o něco nižší, protože anotované jevy mohou mít více správných interpretací.

Velmi obecné vodítko interpretaci κ koeficientů, jež je dodnes poměrně často využíváno, vytvořili Landis & Koch (1977). Sami autoři ovšem zdůrazňují, že se jedná o zcela arbitrární rozdělení spektra výsledných hodnot:

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Tabulka 3: Interpretace κ koeficientů podle Landis & Koch (1977)

Artstein & Poesio (2008) na základě vlastního dlouholetého výzkumu v oblasti lingvistické anotace dospěli k názoru, že pouze hodnoty κ nad 0,8 zaručují anotaci obstojné kvality. Zároveň ovšem podotýkají, že tato hranice není univerzálně platná a že velmi záleží na konkrétním projektu (jako příklad uvádějí, že při tvorbě korpusů

zaměřených na diskurz jsou běžně akceptovatelné i hodnoty kolem 0,7). Rozhodnutí, zdali je korpus dostatečně kvalitní k publikaci, by mělo podle autorů být založeno na více faktorech než jen na hodnotě IAA – například na nezávislé evaluaci výsledků založených na korpusových datech (591).

Pokud bychom se vrátili zpět k výsledkům obou výše zmíněných příkladů výpočtů Cohenova κ a Fleissova κ, lze konstatovat, že obě hodnoty, tzn. 0,92 a 0,884, spadají do kategorie téměř dokonalé shody a značí velmi dobrou reprodukovatelnost daného anotačního schématu.

1.4.2 Gold standard

Pro účely trénování algoritmů strojového učení, vývoje automatických taggerů nebo dalších NLP aplikací je třeba zajistit, aby trénovací korpus anotovaný lidskými anotátory dosahoval dostatečných kvalit (nesrovnalosti v manuální anotaci mají zpravidla negativní vliv na výsledky výstupu strojového učení). Takovýto korpus je označován jako **Gold Standard Corpus (GSC)**.

Postup tvorby GSC poté v obecné rovině vypadá následovně: Po dosažení uspokojivých výsledků IAA anotátoři přistoupí k anotaci celého korpusu. Poté následuje tzv. adjudikace, tedy proces, při němž jsou anotační data všech anotátorů sloučena a pomocí adjudikačního software[10] jsou opět manuálně překontrolována. Tuto kontrolu provádí nezávislý pracovník, adjudikátor, jenž je důkladně obeznámen s anotačním schématem – Pustejovsky & Stubbs (2013) v ideálním případě doporučují využít služeb člena týmu, jež se přímo podílel na tvorbě anotačního schématu (školení nového pracovníka je minimálně stejně časově náročné jako školení anotátorů a vstup dalšího člověka by v této fázi mohl do projektu vnést zmatek a mít tak vliv na kvalitu výstupu). V případě, že adjudikaci provádí více adjudikátorů současně, je opět vhodné průběžně měřit jejich IAA (v tomto případě Inter-adjudicator Agreement), aby se předešlo kvalitativním rozdílům v jejich práci (134). Časová náročnost adjudikace je srovnatelná (ne-li větší) než v případě anotace a vyžaduje taktéž velkou míru soustředění – kromě striktního následování instrukcí musí mít adjudikátor stále na paměti, že shoda mezi anotátory může být pouhým dílem náhody.

[10] Příkladem jednoduchého adjudikačního software je **MAI** (Multi-document Adjudication Interface) (Stubbs, 2018). Umožňuje vizualizaci jednotlivých anotací, filtrování tagů a rychlé opravy chyb a neshod mezi anotátory.

Po dokončení procesu adjudikace je korpus prohlášen za GSC a může se přistoupit k jeho dalšímu využití, například zmíněnému trénování algoritmů strojového učení (viz fáze **TT** v rámci MA**TT**ER cyklu).

1.5 Fyzický formát anotace

Fyzický formát anotace zprostředkovává propojení lingvistických informací s regiony (segmenty) dat, jež jsou anotována. Skládá se ze značek (tagů) definujících, co se v daném regionu z lingvistického hlediska nachází (token, morfém, pojmenovaná entita atp.), případně i jaké má daný segment vlastnosti (lemma, gramatické kategorie, sémantika atp.).

Fyzický formát anotace zachycuje také vztahy mezi jednotlivými anotovanými elementy – může se jednat například o strukturní vztahy (např. v treebankách), funkční vztahy (např. řídící vztahy mezi větnými členy, koreference), ale i zcela arbitrární vztahy odvíjející se od zaměření projektu (Ide et al., 2016).

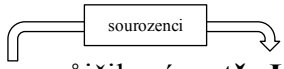
1.5.1 Druhy tagů

Z hlediska anotační terminologie jsou analogicky k výše uvedeným funkcím fyzického formátu anotace podle Pustejovsky & Stubbs (2013) rozlišovány následující druhy tagů:

- **extent tagy (segment tagy)** – tagy označující určitý úsek textu
  příklad anotace sémantických rolí větných členů:
  [Tomáš]$_{\textbf{AGENS}}$ vymaloval [zeď]$_{\textbf{PATIENS}}$ [štětcem]$_{\textbf{NÁSTROJ}}$.
  příklad anotace entit:
  [Hypotéku]$_{\textbf{PRODUKT}}$ [Novákovým]$_{\textbf{KLIENT}}$ poskytla [ČSOB]$_{\textbf{INSTITUCE}}$.

  V případě používání extent tagů k označení např. víceslovných názvů organizací a institucí je v zájmu konzistentnosti anotace (a následné hodnoty mezianotátorské shody) důležité již při tvorbě anotačního schématu jasně definovat rozsah těchto tagů, jelikož úsudek jednotlivých anotátorů se může podstatně lišit.

- **link tagy** – tagy vyznačující určitý vztah mezi dvěma různými extent tagy:



Honza půjčil své **sestře Lence** kolo.

Dále jsou rozlišovány dva druhy tagů z hlediska jejich vztahu k anotovaným datům:

- **consuming tagy** – váží se k určitému segmentu dat anotovaného textu, doslova jej obepínají
- **non-consuming tagy** – k žádnému segmentu dat se neváží, ale i tak jsou v anotačních datech přítomny (například tehdy, když mezi segmenty anotovaného textu existuje relace, která ale není explicitně vyjádřena – v praxi se může jednat kupříkladu o případy elipsy typické pro mluvený projev).

1.5.2 DTD (Document Type Definition)

DTD je soubor pravidel definujících strukturu XML dokumentu a umožňující následně konkrétní dokument validovat a zajistit tak správnost formátování. Z pohledu anotace je DTD prostředkem, jak zadefinovat množinu tagů (tagset) a jejich atributů, které budou v daném anotačním projektu figurovat. DTD navíc specifikuje, jaké druhy vztahů mezi sebou jednotlivé elementy mohou mít (Pustejovsky & Stubbs, 2013). DTD poté v praxi slouží jako šablona, kterou lze importovat do anotačního software a přizpůsobit jej tak na míru potřebám konkrétního projektu (viz např. MAE – Multi-Purpose Annotation Environment). Pokročilejší nástroje umožňují přímou editaci DTD i v průběhu anotace.

1.5.3 Implementace anotačních dat

Pokud jsou anotovány jednotlivé dokumenty jako celky (například zákaznické recenze produktu), je možné anotaci realizovat např. formou tabulkového seznamu nebo databáze s názvy recenzí v jednom sloupci a odpovídajícím tagem v sloupci druhém. V případech, kdy je však zapotřebí anotovat konkrétní pasáže textu, je nutné zvolit jemnější přístup. V současnosti nejpoužívanějšími metodami jsou:

1.5.3.1 Inline anotace

Při inline anotaci jsou tagy fyzicky přítomny přímo v anotovaném textu (obklopují anotované segmenty). Tento způsob anotace je stále poměrně běžný i přes značné množství nevýhod – přímo zasahuje do anotovaného textu a zásadním způsobem ovlivňuje jeho čitelnost a pozdější orientaci v něm. Další nespornou nevýhodou je fakt, že případné sloučení takto anotovaných dat s další vrstvou anotace (např. NE současně s PoS tagy) nevyhnutelně zhorší již tak chaotickou podobu dokumentu – tagy se budou překrývat a orientace v původním textu již bude téměř nemožná. Ztížená je i anotace více slovních tokenů najednou (např. při tagování NE s víceslovným názvem)

Mezi výhody inline anotace naopak patří fakt, že na rozdíl od stand-off anotace nevyžaduje, aby byla jakkoli evidována pozice anotovaných segmentů výchozího textu.

1.5.3.2 Stand-off anotace

Základní charakteristikou stand-off anotace je skutečnost, že anotační data nejsou přítomna v anotovaném textu – jsou uchovávána v externím souboru. Tento způsob anotace je, jak již bylo zmíněno, doporučován předními anotačními standardy v čele s LAF.

Stand-off anotace tokenů (hybridní stand-off)

Prvním typem stand-off anotace je anotace tokenů – text tedy musí být nejprve zpracován tokenizérem (viz tokenizace), přičemž každému tokenu je přiřazeno číselné ID značící jeho pořadí. Toto ID následně v dokumentu s anotačními daty funguje jako odkaz na daný token ve výchozím textu (podle potřeby je samozřejmě možné očíslovat i větší textové jednotky, např. věty či odstavce). Výhodou je, že úplně odpadá problém nepřehledných, překrývajících se tagů. Navíc nedochází k tak zásadnímu narušení výchozího textu, i když je nutné podotknout, že i tokenizace původní formát textu samozřejmě naruší.[11]

Anotace tokenů má však i svá úskalí: I v tomto případě je problematická anotace více tokenů najednou (je ji možno provést, ale je nutné vždy uvádět ID tokenů, jimiž je

[11] Tomu je možné zabránit tím, že se tokenizace provede pokaždé, když je třeba spárovat anotaci a výchozí text (za předpokladu použití identického tokenizéru) (Pustejovsky & Stubbs, 2013).

daný víceslovný segment vymezen). Další nevýhodou této metody je nemožnost jednoduše anotovat části slov – není tedy např. vhodná k anotaci morfémů nebo fonémů (jedinou možností je u každého tokenu zvlášť vyznačovat, jakých jeho znaků se tag týká, což se už ovšem blíží dalšímu typu stand-off anotace, a to anotace s vymezením pozice znaku). Ide et al. (2016) v souvislosti s touto metodou navíc upozorňuje na problematiku zatíženosti teorií, tedy na skutečnost, že její použití nevyhnutelně vyžaduje, aby se autor přiklonil k některé z definic konceptů jako např. token nebo věta.

Stand-off anotace s vymezením pozice znaku

Tento způsob stand-off anotace se od předchozího liší tím, že k vymezení anotovaného obsahu jsou v něm použity číselně vyjádřené pozice znaků (*start* a *end* atributy). Při pohledu na anotační data je na základě těchto údajů samozřejmě prakticky nemožné vyčíst, k čemu ve výchozím textu odkazují, ale pro účely počítačového zpracování se jedná o velmi efektivní a přesnou metodu, jež je v současné době vhodná i pro projekty zaměřené na strojové učení (Pustejovsky & Stubbs, 2013). Výchozí text není v tomto případě vůbec narušen a je možné jej opatřit libovolným počtem různorodých vrstev anotace, aniž by se jakkoliv překrývaly. Důležitým předpokladem při používání této metody je zvolit jednotné kódování znaků (např. UTF-8) a striktně jej dodržovat, aby pořadí znaků v anotovaném korpusu bylo stejné ve všech zařízeních, na nichž se bude v průběhu daného projektu pracovat.

1.6 Anotační nástroje

Hlavní motivací pro používání anotačních nástrojů je potřeba anotace mnohdy rozsáhlých trénovacích korpusů lidskými anotátory. Ruční anotace je ve většině případů činností velmi náročnou na soustředění, a tak je nezbytné, aby se anotátor mohl plně věnovat své úloze a nemusel se zatěžovat dalšími formálními úkony – anotační nástroje proto nabízejí možnost zprostředkovat kódování anotačních dat prostřednictvím uživatelsky přívětivého prostředí do strojově čitelné podoby (ve formátu XML, JSON, CoNLL/IOB[12] atp.), aby tato mohla být dále zpracována například pro účely strojového učení. Jak bude nastíněno dále v této kapitole, některé

[12] Formát užívaný k vymezení entit (datových bloků): B = begin (začátek entity), I = inside (pokračování entity), O = outside (není součástí entity)

sofistikovanější nástroje nabízejí pro anotační projekty velmi komplexní systém podpůrných prostředků zahrnující mimo jiné rozhraní pro celkovou organizaci a monitorování workflow projektu, preprocessing dat atp.

Vlastnostem, jimiž by anotační nástroje v ideálním případě měly disponovat, se ve svém výzkumu[13] podrobně věnuje Dipper et al. (2004): Jedním z hlavních měřítek kvality je zde poměr mezi jednoduchostí nástroje a kvalitou jeho datového výstupu. Dalšími důležitými požadavky jsou funkcionalita a použitelnost. Mírou funkcionality nástroje se rozumí jeho vhodnost pro zpracovávání konkrétních úkolů a interoperabilita s dalšími systémy. Hledisko použitelnosti naopak hodnotí uživatelský komfort nástroje: tedy jak obtížné je naučit se s ním pracovat, jak kvalitní dokumentací disponuje, jak uživatelsky atraktivní je jeho AUI[14] a jestli splňuje relevantní standardy a konvence. Důležitá je také jeho nezávislost napříč IT platformami, uživatelská podpora, vybavenost lingvistickými utilitami, a finanční dostupnost (v ideálním případě zdarma pro účely nekomerčního výzkumu). Velmi užitečnou funkcí je také možnost výpočtu IAA přímo v rozhraní aplikace či adjudikační nástroje.

Mezi další kritéria formulovaná autory patří:

- **Diverzita dat** – schopnost nástroje zpracovávat jazyková data různého charakteru (psaný/mluvený jazyk, anotace segmentů různého rozsahu – slovní tokeny/větné celky atp., podpora široké škály znakových sad).
- **Vícevrstvá anotace** – podpora anotace dat ve více na sobě nezávislých vrstvách reprezentujících různé informace.
- **Diverzita anotace** – podpora dostatečně široké škály tagů a funkcí, pomocí nichž jsou vstupní data anotována (relace mezi prvky, atributy, stromy, grafy atp.). Užitečnou funkcí je mnohdy i možnost tzv. cross-level anotace, tedy anotace napříč jednotlivými vrstvami.
- **Možnosti přizpůsobení** – jednoduchá a intuitivní optimalizace tagsetu podle potřeb projektu (i v jeho průběhu)[15].

[13] Cílem výzkumu bylo formulovat kritéria pro evaluaci anotačních nástrojů založených na XML. Vybrané nástroje byly následně testovány v rámci rozsáhlého projektu na Postupimské univerzitě, jenž byl zaměřen na vícevrstvou anotaci velkého množství jazykových dat.
[14] AUI = annotation user interface
[15] Praxe ukazuje, že v počátečních fázích tvorby anotačních schémat jsou úpravy tagsetu a anotačních instrukcí velmi častým jevem, proto je nezbytné, aby je bylo možno provést co nejjednodušeji (Dipper et al., 2004).

- **Konvertibilita** – možnosti převodu, importu a exportu dat za účelem jejich dalšího zpracování nebo opětovného využití pro další projekty.

1.6.1 Vybrané anotační nástroje

Anotačních nástrojů v současné době existuje nepřeberné množství a všeobecně je lze rozdělit podle toho, zdali byly vytvořeny pouze na míru konkrétního projektu, nebo jestli jsou koncipovány jako univerzální. V této sekci bude představeno několik vybraných zástupců z řad univerzálně zaměřených open source nástrojů, jež v současnosti patří mezi nejhojněji využívané. Přestože za anotační nástroj je považován v první řadě jakýkoliv software vyvinutý za účelem zprostředkování anotačního procesu, je třeba zmínit, že v anotačních projektech často najdou využití i běžné, nespecializované nástroje, jakými jsou například tabulkové a textové editory (MS Excel, Word atp.).

1.6.1.1 Standalone nástroje

Hlavním rysem standalone nástrojů je skutečnost, že jsou instalovány lokálně na zařízení každého anotátora a anotace poté probíhá off-line (data se ukládají nezávisle na sobě a úkony jako adjudikace mohou být prováděny až po jejich manuálním sloučení).

MAE a MAI

Reprezentantem tohoto typu nástrojů je dvojice aplikací vyvinutá Stubbs (2011), *Multi-purpose Annotation Environment* a *Multi-document Adjudication Environment*. Jejich hlavní předností dobrý poměr mezi jednoduchostí a kvalitou výstupu – MAE poskytuje přehledné grafické AUI s podporou link tagů a extent tagů (včetně non-consuming extent tagů) a umožňuje rychlou a jednoduchou tvorbu a konfiguraci tagsetu pomocí DTD. Výstupem je poté stand-off anotace ve formátu XML plně v souladu s LAF. Druhá z aplikací, MAI, umožňuje následnou adjudikaci anotačních dat vytvořených v MAE a tvorbu gold standardu. Oba nástroje najdou pro svou jednoduchost uplatnění v projektech, v nichž je žádoucí, aby proces anotace mohl začít co nejdříve (uvedení nástroje do provozu a naučení se práce s ním není nikterak časově náročné). Naopak nehodí se pro projekty vyznačující se větší mírou komplexity, např.

k hierarchické anotaci a v porovnání s většinou web-based nástrojů neposkytuje žádné dodatečné lingvistické utility ani prostředky k managementu a monitorování práce na projektu.

Mezi další standalone nástroje patří například Callisto (Day et al., 2004).

1.6.1.2 Web-based nástroje

Web-based nástroje nevyžadují lokální instalaci žádného software – data se ukládají a zpracovávají na vzdáleném serveru a uživatelské rozhraní je přístupné prostřednictvím standardního webového prohlížeče (který je tak jediným požadavkem pro fungování nástroje). Rozhraní je uživatelsky přívětivé a je jej ve většině případů možné přizpůsobovat na míru konkrétního projektu. Mezi výhody takovéhoto řešení patří především fakt, že se data neustále ukládají a zálohují na serveru a odpadá starost s jejich ukládáním na více samostatných PC. Celý projekt je díky této centralizaci také možné daleko efektivněji spravovat a koordinovat. Většina takto navržených nástrojů umožňuje využívat předpřipravený vzdálený server (v některých případech zdarma, případně za poplatek, jako tomu je kupříkladu v případě služby GATE Cloud) nebo nabízí možnost provozovat instalaci na serveru vlastním.

BRAT (Rapid Annotation Tool)

BRAT (Stenetorp et al., 2012), disponuje propracovaným grafickým rozhraním a intuitivní vizualizací anotace založenou na vektorové grafice (formát SVG). Umisťování tagů je možné provádět jedním kliknutím myši. Samozřejmostí je podpora extent tagů, předností BRAT je ale především velmi přehledná vizualizace link tagů, a to i (např. při anotaci koreference, Named Entity Recognition nebo jakýchkoliv jiných relací v textu). Anotační data jsou na server ukládána ve stand-off formě a je možno je samostatně vyexportovat. Mezi další výhody patří mód umožňující anotaci stejného korpusu více anotátory v reálném čase (ovšem z podstaty věci bez možnosti výpočtu IAA), adjudikační prostředí, propracovaný vyhledávací nástroj, konkordancer, tokenizér a integrace dalších externích zdrojů (databáze, ontologie atp.). Užitečnou funkcí je pak implementace vzdálených webových nástrojů poskytujících automatickou předanotaci založenou na strojovém učení (v současné době je tímto způsobem možné

využít např. systém automatické desambiguace sémantických tříd v angličtině). Autoři uvádějí, že tato funkce může ušetřit až 15 % času při anotaci.

GATE Teamware

GATE Teamware: a web-based, collaborative annotation framework (Bontcheva et al., 2013) stojí na základech NLP platformy GATE[16] a patří mezi nejpropracovanější nástroje svého druhu, protože poskytuje komplexní podporu celého anotačního projektu a široké možnosti přizpůsobení. Členům týmu mohou v rámci projektu být přiděleny různé role a práva (anotátor, administrátor, projektový manažer atp.) s tím, že každá z rolí má své vlastní uživatelské rozhraní. Uživatelské rozhraní pro samotnou anotaci je koncipováno tak, aby umožňovalo co nejefektivnější práci bez nutnosti dalších úkonů. Nástroj disponuje velmi propracovaným systémem monitoringu, kontroly kvality a adjudikace (včetně výpočtu Fleissova Kappa a dalších koeficientů). K dispozici je celá řada utilit fungujících pod křídly GATE a umožňujících preprocessing korpusových dat a automatickou předanotaci (PoS tagging, NER, analýza sentimentu atp.). Kromě angličtiny je podporováno i několik dalších jazyků včetně češtiny (Universal Dependencies PoS tagger). Samozřejmostí je podpora relevantních ISO standardů (ISO/TC 37/SC4) a výstup ve formě stand-off XML.

WebAnno

Velmi podobnou alternativou GATE Teamware je nástroj WebAnno (Yimam et al., 2013) využívající grafické rozhraní výše zmíněného BRATu. V porovnání s GATE Teamware nestojí WebAnno na základech tak zralé a časem ověřené platformy, což mu ale neubírá na konkurenceschopnosti – jeho devizou je například integrovaná podpora crowdsourcingu (zatím pouze pro anotaci pojmenovaných entit, ovšem autoři počítají s rozšířením i na další oblasti). Rozhraní využívá API crowdsourcingové platformy CrowdFlower (viz následující kapitola). Anotační data pocházející od crowdsourcerů v projektu figurují jako samostatný virtuální anotátor.

[16] https://gate.ac.uk/

1.7 Crowdsourcing v anotaci

Tradiční metody anotace (v podobě, v níž jsou popsány v sekci 2.3.1. v odstavci věnovaném fázi *Anotace* v rámci MATTER cyklu) jsou jak časově, tak finančně náročným procesem. Přesto je skutečností, že většina anotovaných korpusů vznikajících v současnosti těchto metod využívá. Pro některé projekty však tradiční metody nejsou vhodné například z důvodu příliš velkého rozsahu anotovaných dat a tím pádem neúnosných finančních nároků, nebo naopak z důvodu malého rozsahu korpusu. V těchto případech se jako alternativa nabízí možnost outsourcingu procesu anotace vzdáleným spolupracovníkům prostřednictvím internetu[17]. Tato metoda se v posledních deseti letech v komputační lingvistice etablovala natolik, že se v podstatě stala standardem pro anotační projekty menšího rozsahu a uplatnění našla i v již zmíněných rozsáhlých projektech (Poesio et. al., 2017). Každou z metod zmíněných v této sekci je možno implementovat do MATTER cyklu. Jak bude nastíněno, největší rozdíl oproti standardnímu postupu v tomto případě tkví ve způsobu tvorby anotačních instrukcí.

1.7.1 Typy crowdsourcingu

V závislosti na zdroji motivace spolupracovníků k tomu, aby se do projektu zapojili, lze rozlišit tři typy crowdsourcingu:

1. **CS jako forma výdělku** – hlavní motivací řešitele je finanční ohodnocení jeho práce.

2. **CS jako zábava** – spolupráce na projektu je pro řešitele zábavnou činností.

3. **CS jako společný cíl** – chuť podílet se na komunitním projektu (bez nároku na odměnu), podpora vědy či počinu považovaného řešitelem za prospěšný a záslužný.

[17] Estellés-Arolas & González-Ladrón-De-Guevara (2012) definují crowdsourcing jako participativní online aktivitu, při níž zadavatel zadá heterogenní skupině spolupracovníků dobrovolný úkol za podmínek, jež jsou zpravidla výhodné pro obě strany.

1.7.1.1 Crowdsourcing jako forma výdělku

Amazon Mechanical Turk

Příkladem zavedené crowdsourcingové platformy je Amazon Mechanical Turk (zkráceně MTurk). Jedná se o platformu sdružující velký počet online spolupracovníků (pro něž se také užívá označení *turkers*), kteří si mohou vydělat peníze plněním úkolů menšího rozsahu (microtasks). Tyto úkoly se nazývají HITs – human intelligence tasks, zadavatelé je zveřejňují prostřednictvím uživatelského rozhraní MTurk a turkeři se mohou volně ucházet o jejich plnění, které je ve většině případů záležitostí minut. Finanční ohodnocení za zpracování jednoho HITu se v závislosti na jeho obtížnosti většinou pohybuje v řádu jednotek až desítek centů USD, což se v kontrastu s tradičními metodami anotace, kde je zvykem anotátorům platit standardní hodinovou sazbu, může jevit jako výhodnější.

Tento systém má ovšem i svá úskalí – některé druhy anotace již ze své podstaty nejsou vhodné k tomu, aby byly zadávány formou microtasks v rámci HIT systému, protože je není možné jednoduše rozdělit na sérii dílčích úkonů (kupříkladu manuální parsing atp.). Je také zpravidla náročné formulovat anotační instrukce pouze několika větami a zároveň způsobem pochopitelným i pro nelingvisty. Jak poukazuje Fort et al. (2011), pro zadavatele je navíc prakticky nemožné zajistit dodržení i tak základního požadavku, jakým je například status řešitele jako rodilého mluvčího požadovaného jazyka (v případě angličtiny se tak lze setkat s velmi kolísavou kvalitou výsledků, bereme-li v úvahu fakt, že v roce 2010 pocházelo 50 % turkerů z Indie, kde je pro většinu obyvatelstva angličtina až druhým jazykem). MTurk přesto disponuje poměrně propracovaným systémem validace a kontroly kvality: Stejný úkol může být přiřazen více řešitelům, aby se odfiltrovaly nekvalitní výsledky. Zadavatel může navíc specifikovat kritéria, jež řešitel musí splňovat, aby se mohl o HIT ucházet (například dostatečně velký počet již akceptovaných úkolů). Zadavatel může také zcela odmítnout práci řešitelů, s jejichž výkonem není spokojen. Dostatečnou úroveň řešitelů může též zajistit vstupní test ověřující jejich schopnost splnit sérii úkolů, jejichž řešení je součástí zadavatelem poskytnutého zlatého standardu.

Otázkou kvality lingvistické anotace z MTurk se ve svém výzkumu zabýval Snow et al. (2008) a na několika anotačních projektech porovnával práci profesionálních

anotátorů s prací turkerů. Při analýze projektu, jehož cílem byla anotace afektu v nadpisech novinových článků došel k následujícím závěrům: IAA mezi profesionály je vyšší než IAA mezi turkery a profesionály (ale rozdíly nejsou nikterak zásadní). Zároveň se potvrdilo, že jednotliví profesionálové dosáhli lepších výsledků než jednotliví turkeři. Dále se ukázalo, že anotace turkerů v tomto konkrétním případě byla natolik kvalitní, že pokud se zakomponovala do celkového zlatého standardu, došlo ke zvýšení jeho celkové hodnoty IAA. Autor závěrem uvádí, že k dosažení ekvivalentu hodnoty IAA u profesionála bylo v tomto případě třeba v průměru čtyř anotací turkerů (což hovoří silně ve prospěch používání MTurku pro projekty tohoto typu – 3500 anotací od turkerů přišlo na 1 USD, což odpovídá minimálně 875 profesionálních anotací, náklady na něž by však byly nesrovnatelně vyšší).

Samostatnou kapitolou týkající se fungování MTurk je ovšem jeho sociální a etická stránka. Faktem je, že MTurk je neregulovaným online pracovním trhem s absencí ochrany a práv pracovníků a férového ohodnocení práce. Fort et al. (2011) v této souvislosti hovoří o dvou znepokojivých skutečnostech: Průměrný hodinový výdělek na MTurk je pod 2USD, přičemž 20 % turkerů v průzkumu autorů uvádí MTurk jako svůj primární zdroj příjmů, 50 % pak jako sekundární. Pouze 30 % považuje MTurk za volnočasovou aktivitu. Z pohledu vyspělých zemí se tak dá hovořit o formě zneužívání levné pracovní síly z chudších zemí (a je nutno podotknout, že Amazon Mechanical Turk není zdaleka jedinou crowdsourcingovou platformou s nepříliš férovou pracovní politikou). Tento nedostatek byl v posledních deseti letech reflektován a stal se impulzem pro vznik celé řady především startupových společností kladoucích důraz na férové ohodnocení svých pracovníků. Příkladem jsou společnosti Figure Eight (dříve CrowdFlower), Clickworker, OneSpace nebo SamaSource. Poslední zmíněná společnost se přímo zaměřuje na outsourcing zpracování dat do zemí třetího světa, především pak do Ugandy a Keni.

1.7.1.2 Crowdsourcing jako zábava

Games-with-a-purpose (GWAP)

Alternativou placeného crowdsourcingu jsou tzv. games-with-a-purpose, tedy počítačové hry vyvinuté tak, aby během jejich hraní jako vedlejší efekt docházelo k cílenému zpracovávání dat, řešení komputačních problémů či trénování algoritmů

umělé inteligence. Tvůrce tohoto konceptu, Luis von Ahm, již ve své pilotní studii (Von Ahm & Dabbish, 2008) jako jeden z hlavních argumentů pro používání GWAP uvádí fakt, že průměrný jedenadvacetiletý Američan má za svůj život již odehráno 10 000 hodin počítačových her, tedy přibližný ekvivalent pěti let zaměstnání na plný úvazek.

Autoři takovýchto aplikací musí klást důraz na rovnováhu mezi zábavností hry a její schopností produkovat kvalitní datový výstup. Z pohledu MATTER cyklu musí být anotační instrukce vhodně zakomponovány do pravidel a uživatelského prostředí hry. Design aplikace by zároveň měl zaručovat to, že datový výstup bude vykazovat co nejméně chyb (problém může ovšem představovat například snaha některých uživatelů dosáhnout co nejlepších výsledků hledáním mezer v designu hry a zadáváním chybných odpovědí – tomu je však možné předcházet např. zavedením systému odměn pro hráče zadávající mimořádně kvalitní data). Nevýhodou GWAP jsou mnohdy vysoké počáteční náklady spojené s vývojem aplikace, a to jak časové, tak finanční (vývoj jedné aplikace se může v některých případech rovnat až několika letům práce). To je ale částečně vyváženo skutečností, že jakmile je aplikace uvedena do provozu, může generovat data při minimálních nákladech na údržbu. U GWAP navíc na rozdíl od placeného crowdsourcingu zcela odpadá otázka potenciálního zneužívání pracovní síly – lidé se takovéto aktivity účastní už z její zábavné podstaty dobrovolně.

V následujícím výčtu jsou uvedeny některé z nejúspěšnějších GWAP:

Phrase Detectives

Cílem této hry je sbírat data týkající se koreferenčních vztahů v textu. Hráčům jsou předkládány krátké textové úryvky s barevně vyznačenými slovy nebo frázemi. Jejich úkolem je určit, zdali k barevně vyznačenému úseku již bylo v textu referováno.

ESP Game

Cílem je sbírat data pro trénování systému na rozpoznávání obrázků. Hra spočívá v tom, že jsou náhodně spárováni dva hráči, jimž je vždy přidělen stejný obrázek. Oba musí v časovém limitu navrhnout stejné slovo, jímž by každý z obrázků (celkem 15) popsali. Aby se předešlo příliš obecným či běžným popisům, je ke každému obrázku připojen seznam zakázaných slov. Komerční verze hry funguje od roku 2006 pod taktovkou společnosti Google jako Google Image Labeler a má za cíl zlepšit kvalitu

vyhledávání obrázku na Google Images (jejím cílem je zajišťovat, aby k indexovaným obrázkům byla přiřazena správná klíčová slova).

Sentiment Quiz

Cílem hry je na základě odpovědí hráčů sbírat data o emočním náboji zadávaných frází. Hráčům jsou předkládány krátké fráze (případně i slova či věty) a jejich úkolem je prostřednictvím stupnice určit, jestli vyjadřují negativní, neutrální nebo pozitivní emoci. Odpovědi hráče jsou následně porovnány s odpověďmi ostatních a jeho skóre se odvíjí od toho, do jaké míry se s nimi shodoval.

České GWAP

V České republice se vývojem GWAP zabývali například Hladká, Mírovský & Schlesinger (2009). Na půdě Ústavu formální a aplikované lingvistiky MFF UK tak vznikl portál LGame[18] na němž jsou dodnes provozovány tři flashové hry: **PlayCoref** (hledání koreferenčních vztahů v textu), **Place the Space** (hledání mezer mezi slovy) a **The Shannon Game** (doplňování záměrně skrytých chybějících slov ve větách).

Grafické GWAP

Jednou z potenciálních slabin her, jež jsou alternativou k tradiční lingvistické anotaci, je skutečnost, že jsou i přes svůj zábavný charakter stále založené primárně na práci s textem, a mohou tedy více či méně připomínat klasický anotační proces. Jurgens & Navigli (2014) proto navrhli vlastní anotační paradigma, v rámci nějž je možno produkovat lingvistická anotační data prostřednictvím graficky atraktivnějších videoher. Autoři výhody svého konceptu demonstrují na následující dvou hrách, které vyvinuli:

Puzzle Racer

Účelem hry je přiřazovat obrazovým materiálům významy z lexikální databáze WordNet. Jedná se o kombinaci klasické lineární 2D závodní hry (s překážkami a sbíráním různých bonusů a bodů) a řešení rébusu - na začátku každého kola je zobrazena trojice obrázků (nápověda), u nichž má hráč za úkol najít společné téma.

[18] https://lgame.ms.mff.cuni.cz//lgame/sb/

V průběhu závodu poté hráč musí procházet náhodně se objevujícími speciálními branami, při jejichž překonávání se zobrazí několik obrázků, z nichž je třeba zvolit ten, který tematicky odpovídá počáteční trojici. Nesprávná volba znamená ztrátu jednoho bodu života. Existuje ještě druhý typ brány, jež hráči dává na výběr obrázky, které s nápovědou souvisí pouze potenciálně. V tomto případě hráč nemůže přijít o životy. Hra končí tím, že hráč projede celou trasu a zadá slovo, které podle něj odpovídá tématu rébusu. V případě správné odpovědi dojde ke zdvojnásobení jeho skóre.

Experiment provedený Jurgens & Navigli (2014), při němž byl výstup této GWAP porovnán s ekvivalentním objemem dat anotovaných prostřednictvím komerční crowdsourcingové platformy CrowdFlower[19], ukázal, že Puzzle Racer je schopný produkovat kvalitativně velmi podobná anotační data při podstatně nižších nákladech. Komerční crowdsourcing je v tomto srovnání ale jasným vítězem na poli rychlosti anotace.

Ka-boom!

Druhá videohra je zaměřena na desambiguaci lexikálních významů.[20] V počáteční fázi hry je zobrazena věta se zvýrazněným slovem (slovo je tedy užito v kontextu, což je pro tento účel nezbytné). Hráčům jsou následně zobrazovány obrázky a jejich úkolem je kliknutím či dotykem ničit ty, které podle nich významově nesouvisí se zvýrazněným slovem z úvodu. Nezničení irelevantního obrázku znamená penalizaci, přičemž tempo zobrazování nových obrázků se postupně zvyšuje. Na konci hry se hráčům zobrazí řetězec obrázků, které ponechali. I tato hra se na základě experimentu autorů ukázala být velmi obstojnou alternativou v oblasti WSD.

Autoři her Puzzle Racer a Ka-boom! vyzdvihují fakt, že obě mohou být díky svému open-source charakteru modifikovány a využívány za účelem jakéhokoliv jiného anotačního projektu. Co se funkčnosti uvedených her týče, lze říci, že většina z nich (kromě ESP Game a Sentiment Quiz) je stále v provozu. Některé z nich fungují, jak již bylo zmíněno, pod hlavičkou Google Crowdsource.[21] Obecně vzato, potenciál GWAP

[19] Nyní fungující pod názvem Figure Eight (www.figure-eight.com)
[20] Užívá se též anglické zkratky WSD (Word Sense Desambiguation)
[21] **Google Crowdsource** je portál umožňující dobrovolníkům přispívat ke zlepšení kvality online služeb poskytovaných společností Google (evaluace strojového překladu, manuální NER, evaluace emočního náboje, rozpoznávání psaného textu, tagování obrázků atp.).

se v současné době zdá být nejsilnější v oblasti sociálních sítí, prostřednictvím nichž je možno oslovit ohromné množství uživatelů.

1.7.1.3 Crowdsourcing jako společný cíl

Hlavní výhodou tohoto typu crowdsourcingu je, že se mu přispěvovatelé věnují dobrovolně a z přesvědčení, že projekt, jemuž se věnují, má smysl a je společensky prospěšný. Oproti GWAP tedy prakticky odpadá riziko, že by docházelo k cílenému zadávání špatných dat s cílem obejít systém a dosáhnout lepších výsledků než ostatní uživatelé. Zároveň ovšem nelze zcela zaručit ani to, že všechna data přispěvovatelů budou bezchybná (ke snížení tohoto rizika je nutné ze strany autorů projektu věnovat dostatečnou pozornost designu anotačního schématu, uživatelským instrukcím a zároveň celkové prezentaci projektu, jež měla motivovat dobrovolníky k tomu, aby mu zdarma věnovali svůj volný čas).

Projektů fungujících díky společné práci velkého počtu dobrovolníků existuje na internetu v současnosti celá řada – jedním z nejstarších a zároveň nejznámějších je například internetová encyklopedie Wikipedia.org (jejíž úspěch dokládá i fakt, že patří mezi pětici nejnavštěvovanějších webů na světě). Kromě svého přínosu v podobě nejobsáhlejší otevřené encyklopedie na světě jsou data z ní využívána i pro účely výzkumných projektů[22] v oblasti komputační lingvistiky (množství volně přístupného a sdíleného textového materiálu dostupného pod volnými licencemi typu Creative Commons).

Dalším z projektů využívajících práce dobrovolníků je například Zooniverse[23], jenž sdružuje bezmála sto crowdsourcingových projektů napříč vědními obory (především přírodní, humanitní a společenské vědy). Dobrovolníci se tak mohou podílet například na transkripcích historických textů, klasifikaci galaxií, či třídění záznamů z fotopastí. I přesto, že většina úkolů od dobrovolníků nevyžaduje žádné odborné znalosti, jsou výsledky projektů díky vhodně zvolenému designu úkolů hojně využívány v seriózním vědeckém výzkumu. Dalším podobným počinem je projekt Foldit[24], v němž se dobrovolníci mohou podílet na skládání proteinů a přispět tak k výzkumu v oblasti

[22] Příkladem takového projektu je DBpedia, která má za cíl těžit data z webů Wikipedie a organizovat je ve strukturované a strojově čitelné podobě – OKG (Open Knowledge Graph). Využívání takto strukturovaných vědomostních databází přináší zefektivnění práce s informacemi na internetu (rozšířené vyhledávání, sběr dat atp.) ("DBPedia", 2018).
[23] https://www.zooniverse.org/
[24] https://fold.it/portal/

genetiky a bioinformatiky (Foldit díky svým motivačním prvkům, jakými jsou například žebříčky nejproduktivnějších přispěvovatelů, stojí na pomezí dobrovolnického crowdsourcingu a GWAP).

Na stejném typu crowdsourcingu je založen i projekt Open Mind Common Sense (OMCS) spuštěný roku 1999 na MIT. Přispěvovatelé v rámci něj budují databázi konceptů, které považují za *common sense* (českým ekvivalentem užívaným v komputační lingvistice je slovní spojení *zdravý rozum*), tedy „rozsáhlý soubor znalostí a zkušeností týkajících se našeho světa“ (Nevěřilová, 2017). Tyto znalosti jsou v mezilidské komunikaci součástí presupozice komunikantů a nejsou tak explicitně sdělovány – v komunikaci s počítačem, který *common sense* nedisponuje, je však nezbytné, aby existovaly dostatečné podklady k tomu, aby byl obsah sdělení systémem správně interpretován. OMCS tedy shromažďuje jednoduché věty a fráze typu „V hospodě se pije pivo.“ nebo „Kočky chytají myši.“, aby počítače byly schopny lépe porozumět lidské komunikaci. Data z OMCS v současné době slouží pro potřeby většího projektu ConceptNet[25] sdružujícího data z různých zdrojů (GWAP, dobrovolnický crowdsourcing, expertní zdroje typu WordNet atp.) v rozsáhlé vícejazyčné znalostní databázi obsahující fráze, koncepty a vztahy mezi nimi vyskytující se v přirozeném jazyce.

[25] http://conceptnet.io/

2 Vybrané metody anotace sémantických jevů

2.1 Význam a specifika

Důležitost sémantické anotace spočívá v tom, že ačkoliv člověku ve většině situací nečiní problém rozpoznat v kontextu komunikace kýžený smysl výpovědi, pro automatické počítačové systémy v oblasti NLU a NLI (*Natural Language Understanding / Interpretation*) představuje tento úkol v současné době jeden z vůbec nejzásadnějších problémů. Co se specifik anotace sémantiky týče, tak se na rozdíl od anotace morfologické či syntaktické vyznačuje menší mírou automatizace, a tedy i nutností zapojit do procesu anotace (nebo alespoň do jeho počátečních fází) lidské anotátory. Pro takto vzniklé korpusy je typické to, že nejsou hlavně z finančních a časových důvodů tak rozsáhlé, ale zato jsou často velmi hustě a zevrubně anotované (Dipper et al., 2014). Nemalou komplikaci ovšem představuje fakt, že ani lidští anotátoři při anotaci sémantiky nedosahují tak vysoké míry shody, jak by se dalo očekávat (Gries & Berez, 2017).

Sémantickou anotaci lze rozdělit na dvě podskupiny. Prvním typem je anotace lexikálních významů slov v korpusu (WSD – *Word Sense Desambiguation*), při níž jsou slovům podle kontextu, v němž figurují, přiřazovány významy z předem definovaného inventáře významů. Tuto formu sémantické anotace lze v současnosti v některých případech relativně úspěšně provádět i za pomoci automatických taggerů. Druhým typem je anotace konkrétního sémantického jevu, například metafory, metonymie, sarkasmu atp. (tímto typem se bude podrobněji zabývat tato sekce). Sémantické jevy se všeobecně vzato mohou velmi lišit svou komplexitou, mírou teoretického zpracování a vymezení a v neposlední řadě také tím, jak moc je jejich pochopení závislé na individuální interpretaci adresáta. Současná řešení a postupy při anotaci konkrétních sémantických jevů budou představena na deseti vybraných projektech. Na příkladech bude ilustrována nejen rozmanitost a komplexita tohoto odvětví komputační lingvistiky, ale i to, že rigorózní zpracování a anotace některých jevů vyžaduje propracovanou a jasně definovanou metodologii.

Následuje stručná charakteristika dotyčných sémantických jevů.

Humor – v kontextu jazyka je humor komunikátem, jenž v adresátovi vzbuzuje emocionální reakci provázenou pobavením a smíchem. Jedná se o jev poměrně těžko vymezitelný a operující na různých jazykových rovinách (od hry se slovy, přes

dvojsmysly až po komplexní vtipy s příběhovou strukturou). Specifikem humoru je také jeho spjatost s konkrétní kulturou, konvencemi a individuálními preferencemi (jak bude uvedeno v kapitole 3.2, lidští anotátoři se např. většinou shodnou na tom, že se má jednat o humor, ale velmi se rozchází v tom, do jaké míry je text vtipný).

Sarkasmus – cílená změna polarity výpovědi – jinými slovy, intence mluvčího při formulaci výpovědi je opakem jejího konvenčního významu (Nekula, 2017). Z pohledu komputační lingvistiky je detekce sarkasmu zásadní v tom, že nesprávná interpretace těchto konstrukcí může dramaticky snížit efektivitu NLU systémů (hlavně těch zaměřených na sentiment). Problematické je také to, že pochopení sarkasmu je mnohdy závislé na znalosti širšího kontextu a činí potíže i lidem.

Přestože literární věda používá jemnější dělení a jasně odlišuje např. sarkasmus od ironie a satiry, v kontextu NLP je z důvodu poměrně tenké hranice mezi těmito jevy (a jejich v jádru velmi podobné charakteristiky) termíny ironie/sarkasmus povětšinou označován tento fenomén v širším slova smyslu (Reyes, Rosso, & Buscaldi, 2012, s. 3).

Metafora – přenesení významu slova nebo jeho kombinace z jednoho denotátu na druhý na základě vnější podobnosti. Největší a dodnes trvající zájem o studium metafory odstartovali (Lakoff & Johnson, 1980) uvedením teorie konceptuální metafory (CMT), díky níž se metafora stala prominentním předmětem zájmu kognitivní lingvistiky (CMT vnímá metaforu jako kognitivní proces). Ať již na úrovni myšlení, nebo na úrovni jazyka, metafora je velmi komplexní fenomén sahající od kognitivní lingvistiky, přes lexikologii až k literární teorii: v jazyce se lze setkat jak s metaforami novými, kreativními, až po konvencializované metafory, které se stávají součástí polysémie slova, případně ustálených kolokací (frazémů, idiomů).

Metonymie – přenesení významu slova nebo jeho kombinace z jednoho denotátu na druhý na základě soumeznosti (vnitřní podobnosti) (Čermák, 220). Metonymie funguje nejčastěji na bázi kauzálního, prostorového, nebo časového přenosu významu. Stejně jako metafora, i metonymie se pohybuje na škále od nových případů až po metonymie lexikalizované hraničící s polysémií.

Přirovnání – přímé srovnání dvou entit nebo abstraktních konceptů, které může být uvozeno signálním slovem (v češtině např. částicí *jako*). Jedná se o druh metafory, jenž někteří autoři nazývají přímou metaforou. (Více k teorii přirovnání v kapitolách 3.6 a 3.10.2).

Entailment – druh logického vyplývání, pro jehož pochopení je nebytná znalost lexikálního významu daného výrazu. Z pohledu NLP se jedná spolu s problematikou presupozice o jevy, jejichž zpracování by umožnilo rekonstrukci nevyjádřených, ale současně vyrozumívaných částí výpovědi (Karlík, 2017). Počítačové zpracování entailmentu vyžaduje manuální anotaci, která ovšem do celého procesu nevyhnutelně vnáší prvek toho, jakými znalostmi o světě disponují jednotliví anotátoři.

Polarita – detekce polarity je jednou ze základních úloh postojové analýzy (sentiment analysis) a zabývá se určováním toho, zdali je daný úsek textu (nejčastěji věty) pozitivní, negativní nebo neutrální orientace. Třemi základními prvky hodnotících konstrukcí jsou zdroj hodnocení, hodnotící výraz a cíl hodnocení, přičemž jejich paralelu lze často nalézt sémantických rolích větných členů (agens-predikát-patiens) (Veselovská, 2017).

2.2 A Crowd-Annotated Spanish Corpus For Humor Analysis

Anotovaný jev	**humor**
Jazyk	španělština
Autoři	Castro et al. (2017)
Korpus	27 282 tweetů
Odkaz na korpus	https://pln-fing-udelar.github.io/humor/
Počet anotátorů	1271 (crowdsourcing)
Typ anotace	stand-off (CSV)
Anotační nástroj	vlastní webové rozhraní

Tento projekt spadá do podoblasti komputační lingvistiky označované jako Computational Humor, mezi jejíž hlavní sféry zájmu patří detekce a generování humoru a jeho hodnocení.

Hlavním cílem projektu bylo vytvořit španělský korpus, v němž bude anotován humor. Autoři obstarali korpus čítající celkem 27 282 tweetů ze španělských účtů na sociální síti Twitter (tweety pocházejí jak z účtů zaměřených na humor, tak z účtů bez primárně humorného zaměření), přičemž jejich cílem bylo korpus prostřednictvím crowdsourcingu opatřit anotací dvojího druhu – zaprvé, zdali je daný tweet humorný nebo ne a zadruhé, pokud je humorný, tak do jaké míry: Pro účely crowdsourcingové anotace autoři vytvořili vlastní webové rozhraní, v němž anotátoři vždy po zobrazení tweetu měli za úkol rozhodnout, zdali se jedná o humor nebo ne. Pokud označili tweet jako humor, zobrazila se jim škála od jedné do pěti (ve formě emotikonů), pomocí níž měli vyjádřit, jak moc vtipný je.

První tři zobrazované tweety byly vždy stejné, aby se anotátoři seznámili s prostředím nástroje a aby se otestovala jejich kvalita (jednalo se o autory vybrané příklady, u nichž byla zřejmá odpověď – jeden humorný a dva nikoliv). V případě, že anotátor udělal v testovacích tweetech chybu, byly všechny jeho následující odpovědi při finální kompilaci korpusu posuzovány jako méně důvěryhodné. Další tweety byly již vybírány náhodně (s tím, že systém automaticky zaznamenával ID již splněných tweetů, aby se předešlo duplikátům).

Během anotace autoři narazili na problém, že anotátoři měli tendenci označovat nepoměrně větší část tweetů jako nehumorné (tzn. vůbec se nejedná o humor). Prvním opatřením bylo, že tweetům, jež byly již alespoň třikrát označeny jako nehumorné, byla

snížena priorita, aby se anotátoři mohli soustředit na dosud neanotované tweety. Do korpusu bylo navíc doplněno 4500 tweetů náhodně vybraných z twitterových účtů zaměřených na humor.

Každý tweet byl v průměru anotován 3,8x při směrodatné odchylce 1,16 (s výjimkou testovacích tweetů). Za přispění všech 1271 anotátorů bylo celkem zaznamenáno 107 634 anotací. K měření IAA použili autoři Krippendorfovo α (z toho důvodu, že podporuje neomezený počet anotátorů, a navíc bere v úvahu hodnocení na číselné škále, což je právě případ hodnocení, do jaké míry je tweet vtipný). IAA v případě rozhodování humorný / nehumorný dosáhla hodnoty 0,5710, což je hodnota spadající do kategorie průměrné až dostačující shody. IAA v případě hodnocení míry vtipnosti tweetu však byla pouhých 0,1625 (což je hodnota hraničící s téměř náhodnou distribucí anotace). Autoři provedli měření IAA znovu, a to pouze s výsledky dvanácti nejaktivnějších anotátorů (každý z nich vytvořil tisíc a více anotací, v celkovém součtu 50 939) a došli k hodnotám α 0,6345 a 0,2635. V případě rozpoznání toho, zdali se jedná o humor nebo ne, tedy anotátoři dosáhli poměrně uspokojivé shody, ale v případě hodnocení vtipnosti dosažené hodnoty nejsou zdaleka dostačující pro účely trénování ML algoritmů. Mezi důvody nízké IAA autoři uvádějí velkou míru subjektivity v oblasti vnímání humoru.

Autoři výsledný dataset prezentují jako první krok v budování klasifikátoru humoru ve španělštině a jako inspiraci pro budování podobných korpusů i pro další jazyky. Jako cíl pro budoucí výzkum navrhují opatřit jednotlivé tweety ještě více anotacemi než dosud a při analýze výsledků se zaměřit na sociální stratifikaci anotátorů s ohledem na jejich hodnocení tweetu, což by potenciálně mohlo zodpovědět mnoho otázek týkajících se subjektivního vnímání humoru.

2.3 A Large Self-Annotated Corpus for Sarcasm (SARC)

Anotovaný jev	**sarkasmus**
Jazyk	angličtina
Autoři	Khodak et al. (2017)
Korpus	1,3mil. sarkastických + 533mil. nesarkastických Reddit statusů
Odkaz na korpus	http://files.pushshift.io/reddit/
Počet anotátorů	N/A
Typ anotace	stand-off (CSV)
Anotační nástroj	Reddit (self-annotation)

Cílem projektu bylo vytvořit rozsáhlý korpus sloužící jako výchozí bod k výzkumu sarkasmu a k trénování a evaluaci systémů jeho detekce. Autoři navazují na předchozí podobně zaměřené projekty, které ve většině případů jako zdroj dat využívaly Twitter. Autoři SARC ovšem dali přednost sociální sít Reddit (hlavně z toho důvodu častého výskytu zkratek ve tweetech, což značně snižuje kvalitu korpusu). V souvislosti s anotací sarkasmu se autoři rozhodli využít jednoho ze specifik Redditu – jeho uživatelé poměrně stabilně označují sarkasmus ve svých příspěvcích tagem „/s“ umístěným na konci sarkastické pasáže (jedná se tak v podstatě o druh sémantické anotace, pro níž autoři používají označení *self-annotation*). Důvodem pro označování sarkasmu je fakt, že příspěvky jsou z velké části anonymní, a tak se nelze spoléhat na sdílený kontext. Všeobecně vzato s sebou takovýto neorganizovaný způsob anotace (podobně jako využívání hashtagů na Twitteru) přináší množství komplikací, především z toho důvodu, že ne všichni uživatelé jsou v používání označení konzistentní. Autoři podnikli hned několik kroků, aby odfiltrovali pokud možno co nejvíce chybných dat.

První vyskytujícím se problémem je tzv. chyba typu I (false positive), tedy situace, kdy byl příspěvek nesprávně označen jako sarkasmus (to může nastat zaprvé, když autor neví, že „/s“ značí sarkasmus, nebo zadruhé, když pouze hovoří o užívání tagu ke značení sarkasmu nebo zatřetí, když tagem myslí něco zcela jiného – např. hovoří-li o HTML). Aby se předešlo první možnosti, brali autoři v potaz pouze příspěvky uživatelů, kteří „/s“ tag použili již v minulosti, ve druhém případě vyfiltrovali pouze výskyty, kdy byl tag umístěn výhradně na konci příspěvku. Třetí možnost by podle autorů vyžadovala automatickou desambiguaci smyslu příspěvku, tedy záležitost

značně komplikovanou. Částečným řešením by bylo vyfiltrování pouze těch příspěvků, u nichž je dostatečně dobře znám kontext, a tedy malá pravděpodobnost použití tagu v jiném smyslu.

V případě chyby typu II (false negative), tedy absence tagu u sarkastického příspěvku se řešení ukázalo být složitějším především z důvodu nesrovnatelně většího počtu neotagovaných příspěvků. Tento typ chyby nastává, když uživatel o konvenci neví, nebo když se domnívá, že jeho sarkasmus je evidentní. V mnoha případech se podle autorů navíc stává, že na sebe sarkastický příspěvek naváže množství komentářů, které ale již jako sarkasmus označené nejsou (toto autoři prozatím vyřešili vymazáním takovýchto komentářů z korpusu).

Následná manuální evaluace datasetu byla provedena tak, že byl vybrán náhodný vzorek pěti set komentářů označených jako sarkasmus spolu se stejně velkým vzorkem komentářů nesarkastických. Na základě množství výskytů chyb typu I a II ve vzorcích (1 % a 2 %) byly odhadnuty hodnoty pro celý dataset.

Na základě korpusových dat autoři také vytvořili algoritmus k detekci sarkasmu, jehož efektivitu porovnali s manuální lidskou anotací (pět evaluátorů). Úkolem jak pro evaluátory, tak pro algoritmus bylo na vzorku čítajícím 100 příspěvků vždy rozhodnout, který ze dvou obsahuje sarkasmus a který ne. Výsledky manuální evaluace byly ve srovnání s algoritmem podle očekávání znatelně lepší, ale ani evaluátoři nebyli v zdaleka perfektní shodě – hodnota Fleissova κ na vzorku z celého korpusu byla 0,5 a na vzorku příspěvků z oblasti politiky (o němž měli evaluátoři obecně větší povědomí, takže se jim snáze rozpoznávalo, zdali jde o sarkasmus nebo ne) dosáhlo κ hodnoty 0,67.

Celý dataset (v současné době největší svého druhu) i s evaluovanými vzorky dávají autoři volně k dispozici online pro účely dalšího výzkumu (detekce sarkasmu atp.).

2.4 Sarcasm Detection on Czech and English Twitter

Anotovaný jev	**sarkasmus**
Jazyk	čeština a angličtina
Autoři	Ptáček et al. (2014)
Korpus	7000 tweetů
Odkaz na korpus	http://liks.fav.zcu.cz/sarcasm/
Počet anotátorů	2+1
Typ anotace	stand-off (ID dokumentu – tweetu)
Anotační nástroj	Microsoft Excel

Následující projekt je výjimečný tím, že se jedná o vůbec první výzkum v oblasti automatické detekce sarkasmu na úrovni dokumentu (tweetu) v češtině a všeobecně vzato i ve slovanských jazycích. Většina předchozích počinů v této oblasti byla zaměřena především na angličtinu. Jedním z hlavních důvodů je fakt, že slovanské jazyky, typologicky spadající do jazyků flektivních, disponují velmi bohatou, nepravidelnou morfologií a syntaxí a pro NLP představují daleko větší výzvu než např. zmíněná angličtina. Souběžně s češtinou autoři prováděli testování i na anglickém datasetu, následující odstavce však budou věnovány primárně české části projektu.

Autoři nejprve vytvořili gold standard korpus, na jehož základě následně evaluovali několik vlastních nástrojů automatické detekce sarkasmu založených na částečně řízeném učení ML algoritmů.

Co se týče teoretického vymezení sarkasmu, vycházeli autoři z předpokladu, že člověk je přirozeně schopen spolehlivě rozpoznat, zdali se jedná o sarkasmus či ne. Data pro trénování ML algoritmů tedy pocházejí z manuální anotace lidskými anotátory (kteří se nemuseli řídit žádnými specifickými instrukcemi a následovali vlastní jazykový cit – výjimku tvořily pouze tweety obsahující vulgarity, vtipy a sarkastické situace, které byly zpravidla považovány za nesarkastické).

Dataset byl obstarán za pomoci nástrojů Twitter Search API a Java Language Detector. Komplikaci představovala skutečnost, že Twitter v době výzkumu plně nepodporoval český jazyk, takže autoři použili při filtrování parametr *geocode* a vyfiltrovali tweety zveřejněné poblíž Prahy. Získat se takto podařilo 140 000 českých tweetů. V případě českých tweetů nebylo možno efektivně využít fenoménu, jež je běžný v anglické části Twitteru, a sice, že uživatelé často, podobně jako tomu je

v případě Redditu (viz SARC), označují sarkastické příspěvky hashtagy #irony a #sarcasm. Jak bylo zjištěno, čeští uživatelé hashtagy #sarkasmus a #ironie využívají naprosto minimálně (10 + 100 výskytů z celkových 140 000). Jediným východiskem byla tedy manuální anotace náhodně vybraného subsetu čítajícího 7000 tweetů (po anotaci 325 sarkastických a 6675 nesarkastických).

Anotaci datasetu prováděli dva anotátoři, jejichž IAA byla měřena pomocí Cohenova κ a dosáhla hodnoty 0,54 (k neshodě došlo ve 403 případech, k nimž byl povolán třetí anotátor, přičemž hodnoty κ byly následující: – 0,4 (anotátor 1 + 3) a 0,6 (anotátor 2 + 3). Finální hodnotou však zůstává κ 0,54, tedy výsledek bez třetího anotátora. Při následné evaluaci výsledků se ukázalo, že dataset obsahuje 48 chyb prvního typu (false positive) a 52 chyb druhého typu (false negative). K porozumění sarkasmu bylo v některých komplikovanějších případech nutné znát širší kontext, aktuální společenské a politické dění atp. Některé sarkastické tweety bylo též těžké rozluštit pro jejich rafinovanou hru se slovy nebo jinak komplikovaný způsob vyjádření.

Jak již bylo zmíněno, autoři dataset následně použili k testování ML algoritmů na automatickou detekce sarkasmu v češtině. Celý dataset anotovaný jako gold standard korpus (ve formátu .xlsx) je k dispozici online.

2.5 Towards a Corpus Annotated for Metonymies: the Case of Location Names

Anotovaný jev	**metonymie** (místní souvislost – e.g. *place-for-people*)
Jazyk	angličtina
Autoři	Markert & Nissim (2002)
Korpus	subkorpus BNC – 2000 anotovaných místních názvů
Počet anotátorů	2 (autorky schématu)
Typ anotace	stand-off (XML)
Anotační nástroj	MATE

Cílem projektu bylo vytvořit schéma pro anotaci metonymie, jež by sloužilo jako vzor pro budoucí projekty zaměřené na zpracování tohoto sémantického jevu ve větším měřítku. Protože metonymie pokrývá široké spektrum sémantických domén, rozhodly se autorky v první fázi zaměřit na metonymii operující na bázi místní souvislosti.

Autorky v první řadě poukazují na nedostatečné pokrytí metonymických významů ve slovnících (které navíc většinou nezahrnují propria, jež jsou často užívána metonymicky) a lexikálních databázích a na jejich nesystematičnost. Další významnou komplikací je fakt, že odborná literatura má při popisu metonymie tendenci uvádět příliš zřejmé příklady tohoto fenoménu, přičemž v praxi je hranice mezi doslovným a metonymickým významem mnohdy tenká. Aplikace teoretických studií o metonymii z literatury se navíc ukázala být velmi problematickou, a to hlavně z důvodu neexistence žádné všeobecně uznávané kategorizace metonymických významů. Přesto se autorky rozhodly ověřit, zdali je anotace metonymie na základě odborné literatury možná. Jazykové vzorky k anotaci byly vybrány z British National Corpus (BNC) hlavně z toho důvodu, že se jedná o referenční korpus pokrývající vyváženou škálu žánrů a stylů.

V pilotním experimentu tedy autorky sestavily klasifikaci metonymie z dostupné literatury a slovníků. Na základě této klasifikace nejprve anotovaly 100 příkladů metonymie, načež o výsledcích diskutovaly. Po diskusi následovala anotace dalších 200 příkladů. Hodnota κ byla v prvním případě 0,39 a ve druhém (po diskusi) 0,55. V prvním případě šlo o výsledek velmi nespolehlivý, ve druhém nepatrně lepší, ale stále neuspokojivý (viz kap. 1.4.1.4). Po analýze výsledků tak autorky došly k závěru, že je pro další fázi projektu nutné vytvořit komplexní anotační schéma – kategorie z odborné

literatury byly mnohdy příliš specifické a nebylo je možné aplikovat na všechny případy z korpusu, v němž se navíc vyskytovaly o poznání komplexnější syntaktické konstrukce než v příkladech. Potvrdilo se také, že se při anotaci metonymie nelze spoléhat na individuální intuici.

Autorkami vyvinuté anotační schéma je po úpravách aplikovatelné na jakoukoliv sémantickou třídu metonymie a poskytuje tak výchozí bod pro další podobné projekty. V případě anotace metonymie s místní souvislostí je schéma organizováno hierarchicky a bere v potaz i odfiltrování případů, které je záhodno z korpusových dat vyloučit (takovéto případy spadají do kategorií *unsure*, *noapp*, *homonym*). Jak lze vidět v následující tabulce převzaté z Markert & Nissim (2002), kategorizace metonymických vyjádření se dělí na tzv. supertypy (obj-for-rep, place-for-people atp.) a subtypy (CapGov, Off atp.).

Understanding	App	Base-type	Reading	Pattern	Subtype
unsure					
yes	no				
	yes	homonym			
		location	literal		
			metonymic	obj-for-rep	
				obj-for-name	
				place-for-event	
				place-for-people	CapGov
					Off
					Org
					Pop
				place-for-product	
				othermet	
			mixed		

Tabulka 4: Grafické znázornění anotačního schématu podle Markert & Nissim (2002)

Reprodukovatelnost schématu byla testována na subsetu čítajícím celkem 1000 místních pojmenování (názvy zemí). IAA bez zahrnutí subtypů dosáhla velmi obstojné hodnoty κ 0,88, se subtypy pak 0,81. IAA byla následně měřena ještě jednou, přičemž byly vynechány případy *unsure*, *noapp* a *homonym* (vyřazené z anotace), protože jejich identifikace byla snazší než anotace metonymie / doslovné. Bez těchto případů bylo dosaženo hodnot κ 0,87 (supertypy) a 0,78 (subtypy). Dalším krokem byl výpočet toho, jak velké potíže anotátorům činilo identifikovat jednotlivé kategorie (byla použita Krippendorfova statistika). Adjudikace při následné tvorbě gold standard korpusu probíhala za účasti obou anotátorů.

Výstup projektu je tedy dvojí – propracované a zevrubně otestované anotační schéma a gold standard korpus obsahující 2000 místních názvů použitých v kontextu a anotovaných na metonymické, popř. doslovné užití. Dalším cílem je rozšířit anotační schéma i na další sémantické typy metonymie

2.6 Annotating Similes in Literary Texts

Anotovaný jev	**přirovnání (similes)**
Jazyk	angličtina, francouzština
Autoři	Mpouli (2017)
Korpus	1456 fragmentů prozaické poezie (viz níže)
Počet anotátorů	nezveřejněno (crowdsourcing)
Typ anotace	stand-off (XML)
Anotační nástroj	(Dis)Similitudes (založeno na scribeAPI)

Autorka tohoto projektu předkládá návrh anotačního schématu určeného k anotaci stylistického prostředku přirovnání a představuje crowdsourcingovou platformu, jejímž prostřednictvím schéma v praxi testuje. Dlouhodobým cílem projektu je vytvořit manuálně anotovaný korpus, který bude sloužit k evaluaci systémů automatické detekce přirovnání v prozaických literárních textech. Kromě toho autorka vyzdvihuje význam anotace přirovnání (jakožto i dalších stylistických prostředků a básnických figur) pro účely vzdělávání a corpus-based výzkumu.

Autorka vychází z literárně-teoretické definice přirovnání a pracuje se třemi základními termíny: **tenor** (porovnávaná entita), **báze**/ground (na základě čeho je porovnání uskutečněno) a **vehikulum**/vehicle (na co je přenášen význam). Na následující příkladové větě uvedené v Mpouli (2017) jsou ilustrovány jednotlivé konstituenty přirovnání:

She	is soft, crinkled	like	a fading rose
tenor	*ground*	*marker*[26]	*vehicle*

Tabulka 5: Konstituenty přirovnání podle Mpouli (2017)

Přirovnání se v obecné rovině vyskytuje ve dvou podobách, jako doslovné a idiomatické. Idiomatické přirovnání je strukturně identické s přirovnáním doslovným, rozdíl spočívá v sémantice – idiomatické přirovnání spoléhá při přenosu významu z jedné entity na druhou, sémanticky vzdálenější, na adresátovu *znalost světa*.

[26] V souvislosti s markery přirovnání zmiňuje autorka jeden z tagů používaný ve VUMC (Vrije Universiteit Amsterdam Metaphor Corpus) (Steen et al. (2010)), a sice mFlag, který ve zmíněném korpusu do určité míry označuje případy přirovnání.

Doslovné přirovnání pouze udává, zdali si jsou dvě entity rovny, či ne. Jak sémantická, tak syntaktická rovina přirovnáni je autorkou detailněji rozebrána.

Samotná anotace je realizována (projekt stále probíhá) formou dobrovolnického crowdsourcingu. Anotační rozhraní je založené na platformě scribeAPI[27] a funguje pod názvem (Dis)Similitudes[28]. Předmětem anotace je 1456 fragmentů (ve formě obrázků) francouzské, britské a americké prozaické poezie z rozmezí 18. – 21. století (texty tohoto druhu byly zvoleny z toho důvodu ideálního rozsahu – v rámci jednoho HITu je možno je přehledně zobrazit jako celek). Každý fragment sestává z věty obsahující alespoň jeden marker přirovnání a vět, jež ji obklopují (kontext). Anotátoři si posléze mohou vybrat, zdali chtějí zodpovědět na otázky týkající struktury fragmentu (určení, o jaký druh přirovnání se jedná, jakou má funkci a pragmatickou hodnotu a identifikace a popis jejích komponentů), nebo zdali se chtějí věnovat transkripci již anotovaných fragmentů a doplňování jejich sémantických kategorií (seznam kategorií je definován v anotačním schématu). Co se druhů přirovnání týče, rozlišuje schéma celkem pět druhů přirovnání: idiomatické, perceptuální, proverbiální, aktualizované idiomatické a originální (kreativní).

Po dokončení bude anotovaný korpus volně dostupný online. Kromě již uvedených příkladů využití budoucího korpusu autorka navrhuje zkombinovat data s anotací dalších sémantických jevů (metafora, ironie atp.) a zkoumat jejich vzájemné vztahy a provázanost. S pomocí vyvíjeného automatického taggeru bude v budoucnu též možné vytěžit další, daleko rozsáhlejší korpusy literárních textů a studovat vývoj idiomatických přirovnání (např. jejich přechod od přirovnání originálních směrem k ustáleným, případně až ke klišé).

[27] scribeAPI je open-source platforma sloužící k crowdsourcingu v oblasti transkripce digitalizovaných textových materiálů, u nichž je (např. z důvodu špatné čitelnosti) nutná manuální transkripce. Dostupné na adrese: http://scribeproject.github.io/

[28] http://dissimilitudes.lip6.fr:8180/#/

2.7 A Large Annotated Corpus for Learning Natural Language Inference

Anotovaný jev	**entailment**
Jazyk	angličtina
Autoři	Bowman et al. (2015)
Korpus	SNLI Corpus (Stanford Natural Language Inference) – **570 152** párů vět anotovaných na entailment, rozpor a sémant. nezávislost + **56 941 gold standard korpus** (subset)
Odkaz na korpus	https://nlp.stanford.edu/projects/snli/
Počet anotátorů	2500 (crowdsourcing – MTurk)
Typ anotace	stand-off (JSON, TSV[29])
Anotační nástroj	Amazon Mechanical Turk

Tento projekt, vypracovaný Bowman et al. (2015) a vyvinutý na Stanfordově univerzitě, přináší vůbec největší dataset párů vět anotovaných na NLI (*Natural Language Inference*). Autoři k jeho tvorbě přistoupili z důvodu absence dostatečně rozsáhlého zdroje dat nezbytného pro vývoj modelů strojového učení na automatické rozpoznávání logických vztahů (entailment a rozpor) v textech. Díky takto velkému trénovacímu korpusu se autorům v automatické klasifikaci entailmentu podařilo dosáhnout lepších výsledků, než dosahovaly do té doby nejvyspělejší modely založené především na sofistikované symbolické logice. Všechny ostatní dostupné datasety (autoři zmiňují například korpusy série RTE – *Recognizing Textual Entailment*), byly i přes svou kvalitní manuální anotaci pro účely strojového učení nevhodné pro svůj malý rozsah (méně než tisíc vět).

Celý anotační proces probíhal prostřednictvím crowdsourcingové platformy Amazon Mechanical Turk a zúčastnilo se jej celkem přibližně 2500 turkerů. V každém úkolu (HITu) byl turkerovi představen popisek[30] k fotografii (bez fotografie samotné), který sloužil jako výchozí premisa. Úkolem poté bylo *čistě na základě popisku a znalostí o světě* vytvořit tři alternativní popisky: 1) rozhodně pravdivý, 2) možná pravdivý a 3) rozhodně nepravdivý. Tyto tři alternativní popisky odpovídají

[29] TSV = Tab Separated Text (alternativa k CSV, jako oddělovač zde slouží tabulátor)

[30] Popisky pocházely ze samostatného korpusu Flickr30k, který obsahuje 160tis. popisků k přibližně 30tis. fotografiím. Popisky nebyly vytvořeny autory fotografií, ale vznikly v rámci jiného, dřívějšího crowdsourcingového projektu. Vyznačují se tím, že se převážně jedná o velmi doslovné popisy obsahu fotografií bez osobních komentářů, což je pro účely anotace NLI ideální.

anotovaným kategoriím logického vyplývání, tedy entailmentu, sémantické nezávislosti a rozporu.

K tvorbě gold standard korpusu byl použit subset obsahující zhruba 57tis. párů vět (asi 10 % z celkového datasetu). Tento subset byl evaluován třiceti pečlivě vybranými turkery, a to následujícím způsobem: anotátorům byly vždy představeny páry vět ve skupinkách po pěti a jejich úkolem bylo přiřadit každému z párů tag (entailment, rozpor, sémantická nezávislost). Každý pár byl distribuován mezi čtyři anotátory, přičemž pátý pár již obsahoval původní anotaci z prvního kola anotace – každý z párů subsetu byl tedy ve finále anotován celkem pětkrát. Pokud se na tagu shodli alespoň tři anotátoři z pěti, byl prohlášen za gold standard, pokud se tak nestalo (pouze 2 % případů), byl pár vyřazen z následného trénování ML klasifikátoru. Při měření IAA u jednotlivých tagů (u párů vět anotovaných pětkrát) bylo dosaženo následujících hodnot Fleissova κ: rozpor 0,77, entailment 0,72 a sémantická nezávislost 0,60. Celková hodnota κ pro všechny kategorie byla 0,70. Vzhledem k povaze úkolu a zvolené metodě anotace, tedy crowdsourcingu, se jedná o velmi dobrý výsledek. Jak uvádějí autoři, proporce neshod plně odpovídá situaci, kdy je sémantický jev (jehož interpretace je do určité míry závislá na individuální intuici) anotován velkou, heterogenní skupinou v podstatě anonymních anotátorů. Uveřejněný gold standard korpus je podle autorů dostatečně kvalitní na to, aby jeho využití v úlohách strojového učení bylo reálné (byť potenciálně náročnější). Dataset je volně dostupný za účelem dalšího výzkumu v NLP.

2.8 A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

Anotovaný jev	**entailment**
Jazyk	angličtina
Autoři	Williams, Nangia, & Bowman (2018)
Korpus	MultiNLI (Multi-Genre NLI Corpus) – **433 000** párů vět
Odkaz na korpus	https://www.nyu.edu/projects/bowman/multinli/
Počet anotátorů	387 (crowdsourcing – Hybrid)
Typ anotace	stand-off (JSON, TSV)
Anotační nástroj	Hybrid[31] (komerční crowdsourcingová platforma)

Korpus MultiNLI představený Williams et al. (2018) přímo navazuje na předchozí projekt, SNLI korpus. Je anotován podle totožného anotačního schématu a skládá se tedy z párů vět anotovaných na entailment, sémantickou nezávislost, či rozpor. Na rozdíl od SNLI, který je založen výhradně na popiscích fotografií, však MultiNLI obsahuj vyvážená data z deseti různých žánrů psané i mluvené americké angličtiny – umožňuje tedy při evaluaci automatických klasifikátorů postihnout jazyk daleko šířeji než jeho předchůdce. Autoři totiž poukazují na skutečnost, že úspěšné NLU systémy často staví na modelech, jež jsou trénovány na velmi podobných datech, na základě nichž jsou evaluovány – MultiNLI tedy pro klasifikátory přináší výzvu v tom smyslu, že jsou aplikovány na široké spektrum často zcela neznámých domén. Tuto myšlenku potvrzuje i to, že když autoři na MultiNLI otestovali klasifikátory vyvinuté pro SNLI, tak zjistili, že větší rozmanitost korpusu měla za následek poměrně rapidní snížení jejich efektivity, a to i přesto, že oba korpusy mají podobné hodnoty IAA.

Výchozí premisou v zadání pro anotátory byla vždy věta z non-fiction textu (literatura faktu) popisující určitou situaci nebo událost. Hlavním zdrojem textů premis byl korpus OANC[32], z nějž autoři vyextrahovali celkem devět žánrů (viz tabulka). Desátý žánr, nazvaný FICTION, byl zkompilován z volně dostupných literárních děl současné fikce napsaných mezi lety 1912–2010.

[31] http://www.gethybrid.io/

[32] Open American National Corpus, rozsáhlý a volně přístupný korpus rozličných psaných žánrů a transkriptů mluvené americké angličtiny od roku 1990 až do současnosti. Dostupný na: http://www.anc.org/

Žánr	**popis**
FICTION	viz výše
GOVERNMENT	Ofic. reporty, proslovy a korespondence z vládních webů
SLATE	Články z archivu časopisu Slate (1966-2000)
TELEPHONE	Přepisy telefonních konverzací (1990-91)
TRAVEL	Bedekry od Berlitz Publishing (počátek milénia)
9/11	Reportáž o útocích na WTC (2004)
FACE-TO-FACE	Transkripty rozhovorů (počátek milénia)
LETTERS	Korespondence z Indian Center for Intercultural Communication of Philantropic Fundraising Discourse (1990-2000)
OUP	5 odborných publikací o textilním průmyslu a vývoji dítěte od Oxford University Press
VERBATIM	Články o lingvistice pro laiky z časopisu Verbatim (1990-1996)
(CAPTIONS)	Původní korpus SNLI

Tabulka 6: Žánrové složení MultiNLI korpusu

Anotační proces i tentokrát probíhal prostřednictvím komerční crowdsourcingové platformy – v tomto případě se jednalo o platformu Hybrid. Jako motivační prvek se autoři rozhodli sami manuálně oanotovat 1 % datasetu a nabídli crowdsourcerům bonus $1 pokaždé, když použili stejný tag jako oni.

Evaluace probíhala stejným způsobem jako u SNLI a jak je vidět z následující tabulky, spolehlivost (reprodukovatelnost) obou schémat se ukázala být téměř totožnou:

Statistic	**SNLI**	**MultiNLI**
Pairs w/ unanimous gold label	58.3%	58.2%
Individual label = gold label	89.0%	88.7%
Individual label = author's label	85.8%	85.2%
Gold label = author's label	91.2%	92.6%
Gold label ≠ author's label	6.8%	5.6%
No gold label (no 3 labels match)	2.0%	1.8%

Tabulka 7: Srovnání reprodukovatelnosti SNLI a MultiNLI (Williams, Nangia, & Bowman, 2018)

Korpus ve své aktuální verzi je pod svobodnou licencí CC 3.0 k dispozici online a může být za účelem výzkumu modifikován a redistribuován. Autoři věří, že korpus v budoucnu díky své reprezentativnosti a diverzitě poslouží k vylepšování ML klasifikátorů v oblasti NLI.

2.9 Creating Annotated Resources for Polarity Classification in Czech

Anotovaný jev	**polarita**
Jazyk	čeština
Autoři	Veselovská, Hajič & Šindlerová (2012)
Korpus	Aktualne.cz (410 vět), ČSFD (405 vět), Mall.cz (10 177 uživatelských recenzí – 158 955 slov)
Počet anotátorů	2
Typ anotace	N/A
Anotační nástroj	N/A

Veselovská et al. (2012) jako výstup tohoto projektu představují prvotní fázi metody anotace evaluativních konstrukcí v češtině a na třech různých typech anotovaných textových dat otestovali vlastní ML klasifikátor.

Polaritu je možno anotovat na třech úrovních – na úrovni slovního spojení (výrazu), na úrovni věty (segmentu) a na úrovni dokumentu. Autoři se v této fázi rozhodli pro anotaci na úrovni věty s dlouhodobým cílem vytvořit hustě anotovanou treebanku. Anotace na úrovni věty umožňuje, na rozdíl od anotace na úrovni dokumentu, podrobnou analýzu lingvistických jevů podílejících se na vyjádření polarity (slovní druhy, struktura věty atp.). Při klasifikaci polarity na úrovni věty v češtině autoři rozlišují tři funkční komponenty: **zdroj** (osoba nebo entita vyjadřující svůj osobní postoj), **evaluaci** vyjádřenou elementy polarity (např. slova či fráze mající pozitivní/negativní hodnotu) a **cíl evaluace**.

Z důvodu absence anotovaných zdrojů pro analýzu sentimentu v češtině byla první fáze projektu zaměřena na tvorbu korpusů, na nichž by se analýza dala testovat. Celkem byly sestaveny tři datasety zahrnující články ze serveru Aktuálně.cz, recenze filmů z webu CSFD.cz a uživatelské recenze domácí elektroniky z e-shopu Mall.cz.

Prvním anotovaným datasetem byly články z Aktuálně.cz. Instrukce pro anotátory v této fázi zněly následovně: *Měl by čtenář na základě toho, jak daný segment o entitě referuje, k entitě chovat pozitivní, či negativní sympatie?* Volba tohoto přístupu, tj. anotace z pozice čtenáře, byla motivovaná snahou vyvinout systém, který by dokázal chování čtenáře simulovat a vybudovat si sympatie (přičemž trénovací data by tento proces měla zachycovat. Po anotaci 410 segmentů (6868 slov) z dvanácti náhodně vybraných článků dosáhli anotátoři IAA 0,63 (Cohenovo κ). Ukázalo se, že anotátorům

činí potíže oprostit se od osobních sympatií a antipatií vůči danému cíli (obzvláště v případě textů s politickou tematikou). Dalším problémem bylo lingvistické zaměření obou anotátorů – měli mnohdy tendenci označovat v textu položky, jež jim z hlediska polarity připadaly lingvisticky zajímavé, přičemž samotná polarita se nacházela v jiné části segmentu nebo se vůbec nejednalo vyjádření sentimentu. K vylepšení schématu autoři upustili od anotace z perspektivy čtenáře a dospěli k názoru, že je třeba anotovat nejen cíl evaluace, ale i cíl evaluace a její vyjádření. Problém neutrality by mohl být např. crowdsourcingem (více anotátorů tím pádem i reprezentativnější čtenářský vzorek).

Ve druhém kole bylo anotováno celkem 405 segmentů z recenzí filmů na CSFD.cz s o něco lepší hodnotou IAA 0,66. Následná analýza anotace odhalila, že se anotátoři často neshodli ve vymezení cíle a vyjádření evaluace, a to hlavně u vět se sponovými slovesy. Neshody často vznikaly i v případech metaforického vyjádření polarity.

Jako vylepšení schématu autoři v této fázi zavedli dva nové tagy: NONPOS (not that good) a NONNEG (not that bad) vhodné pro situace, kdy evaluace zdroje jde proti předpokládané evaluaci čtenářem. Jedná se tak o zjemnění klasifikace v případech, kdy se nejedná o neutrální, pozitivní ani negativní sentiment. Dalším vylepšením bylo uvedení nového tagu TOPIC sloužícího k označování případů, kdy je evaluace zacílena na dokument jako celek a nelze zakotvit v daném segmentu. Další dva tagy nazvané Good News a Bad News slouží k označení vět, jež vyjadřují věci běžně považované za příjemné/nepříjemné (např. vítězství, bohatství / smrt, nemoc).

Dataset z uživatelských recenzí z Mall.cz (10 177 recenzí celkem o 158 955 slovech) byl specifický v tom, že jednotlivé recenze již byly roztříděny na pozitivní a negativní (6365 pozitivních a 3812 negativních). Práce s ním byla navíc o to snazší, že ze své podstaty obsahoval mnoho prototypických příkladů evaluativního jazyka a zároveň nebyl komplikovaný z hlediska syntaxe a sémantiky. Mezi jeho nevýhody patří občasné gramatické chyby, překlepy, v některých případech nesprávná kategorizace (narozdíl od problémů s datasetem z Aktuálně.cz bylo však řešení těchto záležitostí o poznání snazší). Dataset byl pro svůj prototypický charakter použit k ověření fungování klasifikátoru při jeho testování.

Výstupem z testování představeného schématu je především poznatek, že hodnota IAA při anotaci sentimentu je úzce spjatá s typem anotovaného textu. To, že dosažená IAA přesáhla hodnotu κ 0,6 se podle autorů dá považovat za úspěch (vzhledem ke komplikovanosti úkolu a jeho subjektivitě). Navíc vyšlo najevo, že jednoduchý naivní

Bayesův klasifikátor polarity je schopen dosahovat solidních výsledků i na základě velmi malého anotovaného datasetu.

2.10 Lingvistická anotace metafory

V této sekci je věnována zvláštní pozornost lingvistické anotaci metafory. Bude představen v současné době nejrozšířenější protokol k anotaci metafory, MIP, a jeho aktuálnější, vylepšená verze MIPVU. Následně bude popsáno, jak byl protokol v praxi aplikován k anotaci korpusu a na několika příkladech budou ilustrovány snahy o vytvoření jeho mutací pro použití v dalších jazycích (původní verze byla vytvořena pro anotaci angličtiny). Závěrečná část detailně pojednává o projektu, na němž se autor této práce podílel, a sice o aplikaci protokolu MIPVU na český jazyk.

2.10.1 MIP: A Method for Identifying Metaphorically Used Words in Discourse

Anotační protokol MIP (*Metaphor Identification Procedure*) je výstupem společné práce skupiny vědeckých pracovníků a expertů na metaforu zvané Pragglejaz Group (název je zkratkou prvních písmen křestních jmen jejích členů) (Pragglejaz Group, 2007). Jak uvádí autoři, vznik MIP byl v první řadě motivován tendencí komunity výzkumníků v oblasti metafory zkoumat užívání metafor v reálném diskurzu. Příklady metaforického jazyka vytvořené introspektivně jsou sice cenným zdrojem poznatků o tomto jevu, ale vyvozování reálných závěrů o tom, jak je metaforický jazyk užíván ve skutečnosti, je možné především prostřednictvím corpus-based výzkumu, jemuž stála v cestě absence spolehlivého anotačního protokolu, pomocí nějž by bylo možno metaforu identifikovat v jakémkoliv textu a který by vyřešil problém mnohdy velmi rozdílných intuicí výzkumníků o tom, co je a co není metafora.

Pragglejaz Group tedy MIP prezentuje jako „explicitní, spolehlivou a flexibilní metodu sloužící k identifikaci metaforicky užitých slov jak v mluveném, tak v psaném jazyce“ (Pragglejaz Group, 2007). Protokol operuje na úrovni lexikální jednotky (není zaměřen na identifikaci konceptuální metafory) a pro každou z nich má za cíl určit, zdali je v daném kontextu užita metaforicky a vyžaduje, aby o (ne)metaforicitě jednotky bylo učiněno jasné rozhodnutí. Autoři ale uznávají, že stupeň metaforičnosti slov a jazyka obecně se v reálu může značně lišit. Je nutno brát v úvahu i fakt, že není-li jednotka označena jako metafora, neznamená to, že je užita doslovně (může být stále

předmětem jiného jevu jako například metonymie, hyperbola atp.). Je též zdůrazňováno, že metaforičnost jednotky nemusí nutně korespondovat se záměrem mluvčího nebo pisatele vyjádřit se metaforicky.

Postup aplikace MIP na text je definován následovně:

1. Přečíst celý text a získat tak obecnou představu o jeho celkovém vyznění
2. Určit v textu lexikální jednotky
3. **(a)** Pro každou lexikální jednotku v textu určit její význam v kontextu, tj. jak je aplikována na entitu, vztah či atribut v situaci, jež je evokována textem *(contextual meaning)*. V úvahu je třeba brát co jednotce předchází a co následuje.

   (b) Pro každou lexikální jednotku určit, zdali v současném jazyce může mít v jiných kontextech základnější význam než ten, který má v tomto kontextu *(basic meaning)*. Základní významy jsou typicky:

   ⇒ Konkrétnější – to, co evokují je snazší si představit, vnímat zrakem, sluchem, hmatem, čichem a chutí.

   ⇒ Spojené s tělesnou činností

   ⇒ Preciznější (ve smyslu opaku vágnosti)

   ⇒ Historicky starší

   Základní významy nemusejí být nutně těmi nejfrekventovanějšími významy dané lexikální jednotky.

   (c) Pokud u jednotky v současném jazyce existuje základnější význam vyskytující se v jiných kontextech než v tomto konkrétním kontextu, rozhodnout, zdali je kontextuální význam v kontrastu s významem základním, ale je zároveň na základě porovnání s ním srozumitelný.
4. Pokud ano, označit lexikální jednotku jako metaforickou.

Tabulka 8: Postup aplikace MIP (Pragglejaz Group, 2007)

Součástí popisu procedury je i doporučení autorů týkající se způsobu reportování výsledků výzkumu. To by mělo být co nejpodrobnější a mělo by obsahovat mimo jiné informace o anotovaném textu, o cílové skupině, pro níž je primárně určen, způsobu určování lexikálních jednotek, způsobu kódování, dále pak informace o anotátorech a o způsobu statistického ověření spolehlivosti anotace (IAA).

Při anotaci autoři doporučují nespoléhat se pouze na vlastní intuici (hlavně je-li problematické stanovit základní význam), ale vyhledávat problematické případy v externích zdrojích (slovníky, korpusy). Autoři přímo doporučují používat pokud možno co nejrozsáhlejší synchronní slovníky založené na korpusech (Pragglejaz Group, 13). Volba hlavních referenčních slovníků by měla být pečlivě zvážena a měla by předcházet zahájení samotného anotačního procesu, protože má nezanedbatelný vliv na konzistentnost anotace a tím pádem i na IAA (autoři MIP se navíc při vymezování lexikálních jednotek řídili tím, že každé heslo odpovídá až na výjimky uvedené níže lexikální jednotce). Při testování protokolu byly použity:

- *Macmillan English Dictionary for Advanced Learners* (Macmillan Education, 2002) - hlavní referenční slovník
- *Shorter Oxford English Dictionary on Historical Principles* (Little et al., 1973) (doplňkový slovník sloužící jako reference v otázkách etymologie)

Následující výčet zahrnuje některá z lingvistických rozhodnutí, jež autoři museli učinit, aby proces anotace při pilotním testování protokolu byl konzistentní. Jednalo se především o tyto oblasti:

- **Víceslovná spojení:** když lze víceslovné spojení sémanticky rozložit na dílčí konstituenty, každý z komponentů je posuzován zvlášť jako samostatná lexikální jednotka
- **Polywords:** anglické výrazy typu *of course*, *all right* atp. jsou posuzovány jako samostatná lexikální jednotka
- **Frázová slovesa:** frázová slovesa jsou posuzována jako samostatné lexikální jednotky, protože by jejich rozložením ve většině případů došlo ke ztrátě významu
- **Klasické idiomy:** každý z komponentů idiomu je posuzován jako samostatná lexikální jednotka
- **Ustálené kolokace:** každý z komponentů kolokace je posuzován jako samostatná lexikální jednotka
- **Slovní druhy:** metaforická užití slova se často od základního významu odlišují tím, že spadají pod jiný slovní druh, přičemž vyvstává otázka, zdali by se dvě

homonyma patřící různým slovním druhům měla posuzovat jako dva samostatné lexémy. Autoři se v tomto případě přiklánějí k tomu nebrat otázku příslušnosti ke slovnímu druhu v potaz, což jim např. umožnilo zaznamenat metaforičnost mezi substantivem *squirrel* a slovesem *to squirrel*.

- **Určení základního významu:** autoři uvádějí, že v angličtině je všeobecně snazší určit základní význam u autosémantik než u synsémantik. Výjimku mezi autosémantiky tvoří například některé sémanticky vyprázdněné případy sloves (např. sloveso *make* ve frázi *to make a promise*). V těchto situacích autoři za základní význam vždy považují fyzický význam slovesa (sami však poukazují na to, že při rozhodování tohoto charakteru může hrát velkou roli zatíženost konkrétní teorií a potřeby projektu).

  U předložek vyjadřujících prostorové relace (*in, on, into* aj.) není podle MIP problém identifikovat jejich metaforické užití. Naopak u předložek s abstraktnějším významem (*with, for, of* aj.) je někdy prakticky nemožné odlišit kontextuální a základní význam (podobně je tomu i u spojek, pomocných sloves, zájmen a determinátorů – jednotky náležící k těmto slovním druhům nebyly autory nikdy označovány jako metaforické). Výjimku tvoří osobní zájmena, která mohou použita k personifikaci nebo depersonifikaci (za jejich základní význam je pak považován jejich osobní / neosobní význam). Další výjimku tvoří demonstrativa užívaná metaforicky prostřednictvím empatické deixe – jedná se o konstrukce typu *what's this?* a *what's that?,* v nichž *that* a *this* nevyjadřují deixi, ale emocionální postoj mluvčího. Za základní tedy autoři považovali právě jejich deiktický význam.
- **Mrtvé metafory:** U slov metaforického původu, u nichž se ale v současném jazyce tento původ vytratil autoři odkazují na Lakoffa (1987), který za mrtvé považuje pouze ty metafory, u nichž došlo k úplnému zániku metaforického mapování[33]. Sami autoři se rozhodli za metaforická označovat ta slova, u nichž je kontrast mezi základním a kontextuálním významem současnými mluvčími rozšířen a stále vnímán.

[33] Lakoff (1987) tento jev popisuje na příkladu anglického slova *pedigree* (rodokmen) vzniklého na základě francouzského *pied de grue* (noha jeřába) - tehdejší rodokmeny svým vzhledem nohu jeřába údajně skutečně připomínaly. Toto metaforické mapování je ale v současnosti zcela mrtvé, proto lze slovo i *pedigree* označit za mrtvou metaforu (143-146).

- **Metafora a polysémie:** záměně metafory za nemetaforickou polysemii mají předcházet body 3(a) a 3(b) – pro polysémii není možné nalézt základnější význam, pouze další významy (které ale nelze prohlásit za základnější).
- **Metafora a metonymie:** MIP obsahuje mechanismus předcházející záměně metonymie za metaforu, konkrétně se jedná o bod 3(c) – klíčovým je zde slovo *porovnání* (…zdali je kontextuální význam v kontrastu s významem základním, ale je zároveň **na základě porovnání** s ním srozumitelný.) Metonymie není založena na bázi porovnání, ale většinou má funkci přenosu významu (typicky např. z části na celek nebo naopak). V případě nejistoty autoři doporučují provést následující test: pokud lze do vztahu „A je B" bez ztráty smyslu vložit slovo *jako*, je výraz metaforický (např. bulvární novináři jsou *jako* hyeny). V případech, kdy se výraz nachází na pomezí metafory a metonymie je doporučeno detailně analyzovat kontext.
- **Metafora a přirovnání:** MIP neidentifikuje přirovnání jako metafory, a to ani z formálního, ani z rétorického hlediska.

2.10.1.1 Případová studie autorů MIP

Pilotní testování protokolu probíhalo následovně:

- Šest anotátorů (všichni lingvisté s předchozí zkušeností se studiem metafory, pět z nich rodilí mluvčí angličtiny) bylo seznámeno s protokolem na dvou třídenních kurzech s rozestupem jednoho roku.
- Anotovány byly dva texty o délce 668 a 676 tokenů, první novinový článek (psaný jazyk) a druhý přepis televizního diskuzního pořadu (mluvený jazyk).
- Anotace probíhala ve dvou kolech s rozestupem jednoho týdne, po druhém kole anotátoři porovnali své výsledky s výsledky z kola prvního, učinili finální rozhodnutí a výsledky odevzdali koordinátorovi.
- Diskuse anotátorů nad výsledky proběhla po odevzdání, ale její výsledky nebyly při výpočtu IAA brány v potaz.
- Spolehlivost výsledků byla testována pomocí Cochranova Q a Cohenova κ. Hodnoty κ byly 0,72 u novinového článku a 0,62 u televizní diskuse. Test pomocí Cochranova Q odhalil statisticky významné rozdíly mezi jednotlivými

anotátory, což autoři přisuzují velkému množství hraničních případů metaforičnosti (Pragglejaz Group, 2007, s.22).

2.10.2 MIPVU

Steen et al. (2010) na základě vlastních zkušeností s aplikací MIP v praxi navazují[34] na Pragglejaz Group a přinášejí řadu rozšíření a systematických vylepšení s cílem zvýšit spolehlivost schématu. Dvojice písmen přidaných k původnímu názvu odkazuje na Vrije Universiteit Amsterdam, kde autoři MIPVU působí. Autoři protokol aplikovali při anotaci korpusu VU Amsterdam Metaphor Corpus (VUAMC) a detailně analyzovali specifické problémy a výzvy při jeho aplikaci na čtyři různé žánry v angličtině. Korpus bude představen v rámci této sekce.

Následující výčet vycházející ze Steen et al. (2010) zahrnuje hlavní rozdíly MIPVU oproti MIP:

- MIPVU používá pozměněné paradigma vymezení lexikálních jednotek. Za samostatné lexikální jednotky (a to i když je slovník uvádí pod jedním heslem) jsou považována propria a sekvence dvou substantiv nedisponující vzorcem slovního přízvuku typickým pro kompozita.
- Příslušnost k slovnímu druhu je posuzována jako součást lexikální jednotky, takže homografická slova příslušící různým slovním druhům jsou vnímána jako různé jednotky. Jinými slovy, lexikálními jednotkami jsou z pohledu MIPVU slovní druhy, ne lemmata (již zmíněné substantivum *squirrel* a sloveso *to squirrel* jsou optikou MIPVU dvě různé jednotky). Cílem protokolu je postihnout užití slova v kontextu, ne výsledky metaforických a slovotvorných procesů (16-17).
- Tranzitivní a intranzitivní varianty stejného slovesa nemohou být ve vztahu základního a kontextuálního významu, to samé platí pro počitatelné a nepočitatelné varianty stejného substantiva.

[34] Gerard Steen je jedním z původních členů Pragglejaz Group a zároveň předsedou organizace Metaphor Lab Amsterdam, která podporuje výzkum metafory napříč mnoha vědními obory a každoročně pořádá konferenci s názvem Metaphor Festival. (http://www.metaphorlab.org/)

- MIPVU nebere až na výjimečné případy při určování základního významu v potaz diachronní hledisko (viz poslední odrážka bodu 3(b) u MIP), identifikace metafor je prováděna z pohledu současného mluvčího (22).
- Autoři MIPVU doporučují stejně jako Pragglejaz Group (2007) používání slovníků založených na korpusu. K tomu, jak se slovníkem při určování základního významu pracovat, Steen et al. (2010) poskytují konkrétní instrukce. Oproti MIP je doporučeno pracovat kromě hlavního referenčního slovníku ještě s jedním, srovnatelným slovníkem, aby se pokryly případy, v nichž by primární slovník anotátorům nebyl schopen poskytnout např. dostatečně jemnou klasifikaci významů konkrétního hesla. Při anotaci VUAMC autoři používali následující dva slovníky (OED byl použit pouze velmi výjimečně, a to tehdy, kdy jediným východiskem bylo posouzení etymologie).
    - *Macmillan English Dictionary for Advanced Learners*
    - *The Longman Dictionary of Contemporary English*
    - *(Oxford English Dictionary (OED))*

MIPVU přináší novou klasifikaci typů **MRW** (metaphor-related words), jež je reflektována v tagsetu VUAMC. Jak lze vidět v tabulce, základními třemi kategoriemi jsou **nonMRW** (nemetaforicky užitá slova), **WIDLII** (hraniční případy) a **MRW** (metaforicky užitá slova).

<table>
<tr><td>nonMRW</td><td></td></tr>
<tr><td>WIDLII</td><td rowspan="2">Nepřímá metafora</td></tr>
<tr><td rowspan="2">MRW</td></tr>
<tr><td>Přímá metafora</td></tr>
<tr><td>Implicitní metafora</td><td rowspan="2"></td></tr>
<tr><td>Metaphor signals (mFlag)</td></tr>
<tr><td>Personifikace</td><td></td></tr>
</table>

Tabulka 9: Schematické vyjádření tagsetu MIPVU

1) **Non-MRW** – nemetaforicky užité slovo
2) **WIDLII** *(When In Doubt, Leave It In)*

Hraniční případy, v nichž ani po individuálním posouzení anotátorů, ani po následné skupinové diskusi nebylo možno jednotku označit jako jasné MRW nebo non-MRW. Do této skupiny spadají také případy, jejichž metaforičnost nelze s jistotou určit z důvodu víceznačného kontextu. V následujícím příkladu vybraném z korpusu

VUAMC (Steen et al., 2010) je sloveso ***opens*** označeno jako hraniční případ, protože v tomto kontextu může nabývat jak fyzického (ve smyslu uvolnění vstupu do budovy), tak přeneseného významu (ve smyslu otvírací doby).

*The Sugar House only **opens** er Thursday Friday Saturday*

Ve VUAMC spadají všechny WIDLII pod kategorii nepřímé metafory.

3) MRW

a. **Nepřímá metafora** *(indirect metaphor)*

Kontextuální význam je v kontrastu s významem základním a lze jej s ním porovnat. Kontextuální význam může být konvencionalizovaný a tím pádem přítomný ve slovníku, ale může se jednat i o nový, neotřelý význam, jenž ve slovníku není. Základní význam je specifičtější, konkrétnější (ve smyslu bodu 3(b) MIP) a nachází se ve slovníku.

*While the first **wave** of popular interest arose…*

V příkladové větě z VUAMC (Steen et al., 2010) je substantivum ***wave*** označeno jako nepřímá metafora, protože na základě definice ze slovníku *Macmillan English Dictionary Online* existuje kontrast mezi kontextuálním významem:

„a sudden increase in a particular type of behaviour or activity, especially one that is unpleasant or not welcome“ („wave (noun)“, n.d.)

a základním významem:

„a line of water that rises up on the surface of a sea, lake, or river“ („wave (noun)“, n.d.).

b. **Přímá metafora** *(direct metaphor)*

Jedná se o případ, který by podle MIP nebyl považován za metaforu. Kontextuální význam není v kontrastu s významem základním, srovnání je vyjádřeno doslovně, přičemž může (ale nutně nemusí) být uvozeno signálem (**mFlag**) jako např. *like, as, as if*.

Veale et al. (2016) přímou metaforu ilustrují na anglické větě *Juliet is like the sun.* Julie je zde přirovnávána ke Slunci ve smyslu nebeského tělesa (substantivum *sun* není MRW). Obě entity v příkladu jsou tedy použity ve svém základním významu, ovšem jejich srovnání už jako doslovné být vnímáno nemůže – vlastnosti Julie a Slunce jakožto zcela jiných domén nejsou v tomto kontextu vzájemně kompatibilní a přirovnání ve smyslu *Julie je jasná, rozzářená, přitažlivá (...) jako Slunce* je již ze své podstaty figurativní (68).

V příkladové větě z VUAMC (Steen et al., 2010) je podstatné jméno ***snake*** přímou metaforou a ***like*** slouží jako mFlag:

*He turned on me **like** a **snake**.*

c. **Implicitní metafora** *(implicit metaphor)*

Kategorií implicitní metafory jsou označena slova, která v textu mají vztah koreference k metaforicky užité lexikální jednotce (např. anaforická osobní zájmena v koreferenčním vztahu s MRW). Na příkladu z VUAMC (Steen et al., 2010) lze tento vztah nalézt mezi substantivem ***things*** (nepřímá metafora) a k němu referujícím zájmenem ***them*** (implicitní metafora):

*I don't think you can see **things** the way I see **them***

Personifikace

Personifikace je v rámci MIPVU vnímána jako konceptuální mapování založené na přiřazení lidských kvalit nelidské entitě (nebo jevu atp.). Ve VUAMC jsou případy MRW, kde potenciálně hraje roli personifikace, označeny komentářem *possible personification* a v online verzi je lze pomocí vyhledávacího nástroje vyfiltrovat. V následujícím příkladu z VUAMC (Steen et al., 2010) je sloveso ***share*** označeno jako nepřímá metafora vyvolaná personifikací (váže se na nelidskou entitu, substantivum ***theories***):

*These **theories** **share** certain similarities with biological explanations.*

DFMA *(Discarded For Metaphor Analysis)*

Autoři MIPVU používají označení DFMA pro případy, v nichž není možné určit kontextuální význam (nejedná se o tag, ale spíše o termín vyskytující se v instrukcích k protokolu). Příkladem DFMA je situace, kdy věta z anotovaného korpusu (nejčastěji mluveného) obsahuje nedokončenou nebo jinak nesrozumitelnou výpověď.

2.10.3 VU Amsterdam Metaphor Corpus (VUAMC)

Anotovaný jev	**metafora**
Jazyk	angličtina
Autoři	Steen et al. (2010)
Korpus	**VU Amsterdam Metaphor Corpus** **187 570** lexikálních jednotek z korpusu **BNC Baby**
Odkaz na korpus	http://www.vismet.org/metcor/search/ (online s vyhledávačem) http://purl.ox.ac.uk/ota/2541 (XML)
Počet anotátorů	6
Typ anotace	stand-off (XML)
Anotační nástroj	N/A

VUAMC (Steen et al., 2010) je v současné době největší korpus disponující manuální anotací metafory na základě MIPVU. Korpus pokrývá čtyři žánry (academic, conversation, fiction, news – každý přibližně o 50tis lexikálních jednotkách) a je založen na anglickém korpusu BNC Baby (subkorpus BNC o velikosti přibližně 4mil slov). VUAMC využívá PoS tagging z BNC. Korpus lze volně stáhnout ve formátu XML a zároveň existuje i jeho online verze disponující základními korpusovými vyhledávacími nástroji (KWIC/KWOT, tabulkové zobrazení atp.)

Za účelem otestování spolehlivosti schématu autoři provedli celkem šest měření IAA (Fleissovo κ a Cochranovo Q). V průběhu anotace korpusu byly z každého z žánrů náhodně vybrány fragmenty o srovnatelné velikosti a čtveřice anotátorů je anotovala podle mírně zjednodušené anotační procedury: Proces anotace byl Steen et al., (2010) zredukován na binární úlohu bez výše zmíněné jemnější klasifikace MRW, takže anotátoři lexikálním jednotkám v přiřazovali pouze tagy MRW (1) a nonMRW (0). Hraniční případy (WIDLII) nebyly brány v potaz, protože jsou zpravidla konzultovány v rámci společné diskuse a jejich role v samostatné anotaci tedy není rozhodující. Z podobného důvodu byla vynechána i anotace mFlagů, jelikož jejich výskyt v datech

není tak častý, a navíc se nepotvrdilo, že by byly výraznějším zdrojem neshod mezi anotátory (152-153).

Níže uvedená tabulka uvádí výsledná data z posledního, šestého kola měření IAA.

File ID in BNC	Number lexical units	Percentage unanimous			Fleiss' κ	Min MRWs	Max MRWs	Cochran's Q (df=3)
		Not MRW	MRW	Total				
FEF (acad.)	534	73.4	14.0 (n=75)	87.4	0.79	102	126	17.07***
KNR (conv.)	602	87.2	6.1 (n=37)	93.3	0.78	44	66	30.39***
J54 (fict.)	401	82.3	11.0 (n=44)	93.3	0.85	52	61	6.57
K58 (news)	384	77.9	19.5 (n=75)	97.4	0.96	77	85	16.03***
Total	1921	80.4	12.0 (n=231)	92.5	0.85	282	317	20.36***

*** $p = 0.001$

Tabulka 10: Výsledky 6. kola měření IAA u jednotlivých žánrů VUAMC (Steen et al., 2010)

Jak lze vidět, míra IAA se při aplikaci MIPVU může značně lišit v závislosti na žánru anotovaného textu. Všechny hodnoty κ uvedené v tabulce ovšem spadají do kategorie velmi obstojné shody (celková průměrná hodnota κ 0,85 překonává dokonce i hranici 0,8 nastavenou Artstein & Poesio (2008)). Zmíněná hodnota κ 0,85 spolu s průměrnou jednomyslnou shodou přes 92 % může být podle Steen et al. (2010) považována za pomyslný cíl při budoucích aplikacích MIPVU na další texty (164). V následující kapitole bude popsána aplikace MIPVU texty v českém jazyce.

3 Aplikace MIPVU na češtinu

3.1 Představení projektu

Cílem projektu je modifikace protokolu MIPVU tak, aby mohl být použit k anotaci metafory v textech psaných v českém jazyce. Jedná se o prvotní fázi tvorby českého korpusu anotovaného na metaforu, jenž by mohl být velmi užitečným zdrojem pro lingvistický výzkum v oblastech jako např. komputační, kognitivní a korpusová lingvistika.

Tato fáze projektu zahrnuje:

1) Modifikaci protokolu MIPVU tak, aby pomocí něj bylo možné spolehlivě identifikovat metaforu v českém jazyce
2) Představení alternativního tagu (paralelně s původními tagy z MIPVU), který by v případě potřeby umožnil vyfiltrovat silně lexikalizované případy metafory. Motivací pro zavedení tohoto tagu je využití výsledného korpusu k trénování systémů automatické identifikace metafory (podrobně v sekci 3.6).

3.2 Projekty aplikující MIPVU na další jazyky

Badryzlova et al. (2013) modifikovali MIPVU pro použití v ruštině a pokusili se rozšířit anotaci i na úroveň konceptuální metafory (deep annotation). V průběhu vývoje ruské mutace protokolu autoři měřili IAA a srovnávali dosažené hodnoty s hodnotami, jež Steen et al. (2010) zveřejnili pro dílčí testy během vývoje MIPVU. Ve druhém kole měření, po aplikaci nutných modifikací, již hodnota IAA ruského týmu dokonce mírně překonala hodnoty reportované autory korpusu VUAMC. Projekt byl později přerušen, ale Badryzlova & Lyashevskaya (2017) čtyři roky na to obnovily práci na ruském korpusu anotovaném na metaforu. Jejich cílem je přidat do ruského korpusu SynTagRus (dependenční treebanka) vrstvu anotace metafory na základě jimi modifikované procedury MIPVU.

Dalším z příkladů aplikace protokolu na další jazyky je projekt Justiny Urbonaitė (2015), která s použitím MIPVU zkoumala metafory vyskytující se v textech z oblasti práva v litevštině a angličtině. Přestože její výzkum neposkytuje data o IAA (autorka byla jediným anotátorem), uvedené postřehy týkající se problémů a specifik při aplikaci protokolu na litevštinu jakožto na flektivní jazyk jsou pro účely aplikace MIPVU na češtinu velmi cenné.

Pro současnou fázi tohoto projektu je použit model na bázi (Badryzlova et al., 2013) s využitím poznatků ze všech tří výše zmíněných zdrojů.

3.3 Anotace a první kolo testování spolehlivosti

Anotované texty

Předmětem prvního kola anotace byly dva úryvky textů o délce přibližně 600 tokenů. První úryvek (598 tokenů) spadá do žánru fikce a pochází z povídky s názvem *Zasraný vánoce* od Michala Viewegha. Druhý úryvek (611 tokenů) pochází z přepisu jednání Evropského parlamentu (dále Europarl) dostupného z paralelního korpusu InterCorp (Rosen et al., 2017), jenž je součástí Českého národního korpusu.

Anotátoři

Texty anotovali tři anotátoři – dva Ph.D. studenti a jeden student magisterského stupně (autor této práce). Všichni tři anotátoři jsou rodilými mluvčími českého jazyka, studenty lingvistiky a mají předchozí zkušenosti se studiem konceptuální metafory.

Použité slovníky

Při určování základních významů sloužily jako reference dva slovníky: *Slovník spisovného jazyka českého* (SSJČ) (Vácha et al., 1971) a *Slovník spisovné češtiny* (Kroupová et al., 2005).

Měření IAA

IAA byla měřena Fleissovým κ. Pro výpočet a určení případů neshod byl použit pro tento účel speciálně vyvinutý program v jazyce Python.

Anotace

Proces anotace probíhal v podobném duchu, jako když autoři VUAMC při jeho vývoji prováděli dílčí měření IAA. Anotátoři tedy pracovali s tokenizovaným textem v tabulkovém procesoru Microsoft Excel a každé jednotce přiřazovali hodnotu 1 (MRW) nebo 0 (nonMRW).

Text	Počet tokenů	Jednomyslná shoda (%)			Fleissovo κ
		NonMRW	MRW	Celkem	
Viewegh	598	87,46	4,85	92,31	0,65
Europarl	611	76,76	10,97	87,73	0,72
Fleissovo κ celkem					**0,70**

Tabulka 11: Výsledky 1. kola měření IAA

Co se dosažených hodnot κ týče, je kromě všeobecně uznávaných minimálních hranic (viz interpretace κ koeficientů v sekci 1.4.1.4) pro tento projekt obzvlášť relevantní srovnání s výsledky IAA, jichž v podobné fázi dosahovali tvůrci VUAMC a (Badryzlova et al., 2013):

Aplikace MIPVU na češtinu 3 anotátoři **Test 1**	**Russian corpus of conceptual metaphor** 3 anotátoři cca 2000 tokenů (Badryzlova et al., 2013) **Test 1**	**Russian corpus of conceptual metaphor** 3 anotátoři cca 2000 tokenů (Badryzlova et al., 2013) **Test 2**	**VUAMC** 4 anotátoři 1921 tokenů (Steen et al., 2010) **Test 6**
0,70	**0,68**	**0,90**	**0,85**

Tabulka 12: Porovnání IAA z 1. kola anotace s ostatními projekty

Z tabulky vyplývá, že dosažená hodnota κ po prvním kole anotace stále nedosahuje dostatečné úrovně, ale zároveň je velmi podobná hodnotě, kterou po prvním kole a za použití nemodifikované verze MIPVU reportují Badryzlova et al. (2013).

3.4 Rozbor chyb a navrhované modifikace

Případy neshod

Následující tabulka obsahuje data o počtu neshod mezi anotátory pro oba anotované úryvky podle slovních druhů. Nejproblematičtějším slovním druhem v tomto směru byla slovesa a hned po nich předložky (u textu od Michala Viewegha) a substantiva (u Europarl).

Slovní druh	**Viewegh**	**Europarl**	**Neshod celkem**
Substantiva	6	18	24
Slovesa	18	30	48
Adjektiva	6	6	12
Adverbia	5	4	9
Předložky	11	16	27
Spojky	0	1	1
Celkem	46	75	121

Tabulka 13: Počty neshod podle slovních druhů

Za povšimnutí stojí, že anotace úryvku Europarl sice obsahuje více případů neshod, ale současně vykazuje vyšší hodnotu IAA (viz tabulka č. 13). Důvodem je skutečnost, že úryvek obsahuje více než dvakrát tak velké množství MRW než úryvek druhý

(Viewegh – fikce). Tento poznatek plně koresponduje se zjištěními, která přináší Steen et al. (2010), a sice, že ze všech čtyř analyzovaných žánrů (academic, news, fiction a conversation) obsahuje méně MRW než fikce už jen conversation.

Převážná část neshod v anotaci sloves byla pravděpodobně způsobena individuálními biasy anotátorů spíše než systémovou chybou v protokolu. V případě úryvku Europarl jeden z anotátorů neoznačil několik metaforicky užitých lexikálních jednotek jako MRW. Jednalo se o slovesa, u nichž anotátor přehlédl vztah personifikace vůči subjektu (který byl ve všech případech značně abstraktní: např. štěstí, šance, právo, svoboda). Anotátor si toho přehlédnutí uvědomil ihned po tom, co byla anotace ukončena.

Předložky

Předložky jsou v mnoha jazycích metaforicky nejbohatším slovním druhem – Steen et al. (2010) uvádí, že tvoří 38,5-46,9 % MWR ve VUAMC. České předložky se oproti anglickým vyznačují ještě větší mírou homonymie a v prvním kole anotace byly zdrojem mnoha případů neshody mezi anotátory.

S cílem odstranit tento problém byl po vzoru Badryzlova et al. (2013) vytvořen seznam základních významů nejčastěji užívaných českých předložek. V souladu s českou lingvistickou tradicí byly pak významy předložek rozděleny podle pádů (Veselková, 1986; Štícha et al., 2013). Sestavení tohoto seznamu umožnilo snáze odfiltrovat homonymii a dojít tak k jednomu základnímu významu.

Následující čtveřice vět ilustruje rozmanitost významů předložky *za*:

1) Petr stojí **za** mnou.
2) Chytil jsem ho **za** nohu.
3) **Za** dva roky to bude hotové.
4) Vyměnil jsem kolo **za** auto.

Zatímco ve větách 3) a 4) je zřejmé, že předložka *za* je MRW, u vět 1) a 2) nastává problém, jelikož se očividně jedná o dva rozdílné významy, jež jsou optikou MIPVU stejně konkrétní a fyzické. Pokud ovšem budeme uvažovat rozdíl mezi předložkou *za* v instrumentálu (věta 1) a v akusativu (věta 2), budou tyto dva základní významy moci být rozlišeny. Zmíněná předložka *za* v akusativu navíc slouží jako základní význam pro metaforicky užitou předložku *za* ve větách 3) a 4).

Zvratná zájmena *se/si* a pomocná slovesa

Zvratná zájmena *se/si* jsou užívána buď když je objekt totožný se subjektem (věta 5) nebo jako nedílná součást zvratného slovesa (na jehož lexikálním významu se často podílí). Přítomnost zvratného zájmena může změnit význam konkrétního slovesa (bod 6):

5) Umyji **se**.
6) rozvést (myšlenku) / rozvést (manželský pár) / rozvést **se**

Původní MIPVU protokol s výskytem tohoto fenoménu nepočítá. V následující tabulce je zachycen jeho vliv úryvek z anotovaného textu:

Text	Když	se	před	třemi	lety	rozvedl (…)
Původní MIPVU	0	0	1	0	0	1
Modifikované MIPVU	0	0	1	0	0	0

Tabulka 14: Příklad věty, v níž zvratné zájmeno způsobuje změnu významu

Považujeme-li šedě zvýrazněné tokeny za samostatné lexikální jednotky, bude muset být sloveso *rozvedl* vnímáno jako tranzitivní (např. *Soudce rozvedl manželský pár.*). Pokud ale chceme zachytit jeho skutečný význam v této větě, tedy reflexivní *Rozvedl **se*** fungující zde jako konkurenční forma pasivní konstrukce *Byl rozveden*, je bezpodmínečně nutné za součást lexikální jednotky považovat i zvratné zájmeno *se*.[35] Za tohoto předpokladu je jak zájmeno *se*, tak sloveso *rozvedl* označeno jako nonMRW, protože jde o jednu lexikální jednotku. Tento přístup je v rozporu s pravidlem uváděným (Steen et al., 2010, 30), a sice, že tranzitivní a intranzitivní varianty stejného slovesa nemohou rozlišovat základní a kontextuální význam. Z pohledu anotace metafory v češtině jsme ale došli k názoru, že je z důvodu rozdílného sémantického chování určitých sloves v některých případech nutné považovat reflexivní a tranzitivní varianty stejného slovesa za různé lexikální jednotky.

Podobně jako v případě některých zvratných zájmen, i pomocná slovesa jsou v české mutaci MIPVU považována za nedílnou součást tvaru slovesa (tvoří spolu s ním jednu lexikální jednotku). Na druhou stranu, při aplikaci MIPVU na český jazyk odpadá

[35] V pilotních kolech anotace byla zvratná zájmena a pomocná slovesa označována stejnou hodnotou (MRW/nonMRW) jako sloveso k nim příslušící. Tento způsob anotace je ospravedlnitelný přihlédneme-li k tomu, že se jedná o ranou fázi projektu, jejímž cílem je modifikovat anotační schéma. Při anotaci korpusu by tento způsob již vhodný nebyl, protože by docházelo k ovlivňování statistik o výskytu metafory u různých slovních druhů.

mimo jiné celá problematika frázových sloves (jež jsou protokolem považována za jednu lexikální jednotku) – to, co angličtina vyjádří frázovým slovesem, je v češtině ve většině případů vyjádřeno prefixem, jenž je součástí slovesa (viz např. ***ze**sílit / turn **up***).

Ustálená slovní spojení

Co se ustálených spojení týče, bylo postupováno podle (Steen, 2017), tedy že každé slovo, z nějž se ustálené slovní spojení skládá, má být analyzováno jako samostatná lexikální jednotka (83). Používání slovníku při práci s ustáleními spojeními po vzoru (Badryzlova et al., 2013) se v případě českého jazyka ukázalo být značně problematickým, protože dostupné české slovníky nejsou ani založené na korpusech (jak doporučují autoři MIP i MIPVU), ani dostatečně nereflektují současný stav jazyka (velké množství archaismů).

3.5 Druhé kolo testování spolehlivosti

Druhé kolo testování spolehlivosti, tentokrát již modifikovaného protokolu, proběhlo přibližně půl roku po kole prvním. Anotace proběhla zcela identickým způsobem jako je popsáno výše (stejná trojice anotátorů, stejné referenční slovníky a metoda výpočtu IAA).

Anotovány byly celkem čtyři texty žánrově korespondující se čtveřicí textů, na níž testovali spolehlivost MIPVU Steen et al. (2010) a Bardzylova et al. (2013):

- **Academic (oborová literatura):** úryvek (640 tokenů) z populárně-naučné historické knihy *Boleslav II.* od českého historika Petra Charváta dostupné z korpusu SYN2015 (Křen et al., 2016).
- **Fiction (beletrie):** úryvek (614 tokenů) z románu *Poslední Aristokratka* od českého spisovatele Evžena Bočka dostupné z korpusu SYN2015 (Křen et al., 2016).
- **News (publicistika):** úryvek (580 tokenů) ze článku z publicistického časopisu pro mládež ABC dostupné z korpusu SYN2015 (Křen et al., 2016).
- **Spoken (mluvená čeština):** úryvek (629 tokenů) z korpusu neformální mluvené češtiny ORAL2013 (Benešová, Křen, & Waclawičová, 2013)

Text	Počet tokenů	Jednomyslná shoda (%)			Fleissovo κ
		NonMRW	MRW	Celkem	
Academic	640	81,4 (n=521)	14,38 (n=92)	95,78	0,90
Fiction	614	87,95 (n=540)	8,63 (n=53)	96,58	0,88
News	580	84,66 (n=491)	11,38 (n=66)	96,04	0,88
Spoken	629	89,83 (n=565)	6,52 (n=41)	96,35	0,84
Fleissovo κ celkem					**0,88**

Tabulka 15: Výsledky 2. kola měření IAA

Jak lze vidět z tabulky č. 15, po druhém kole již celková IAA pro všechny čtyři texty dosáhla hodnoty κ 0,88, což ilustruje fakt, že navržené modifikace protokolu měly pozitivní vliv na jeho reprodukovatelnost. Opět se navíc potvrdilo, že žánr s nejmenším podílem MRW je mluvená čeština, následovaný žánrem beletrie.

Následující tabulka poskytuje srovnání celkového výsledku měření IAA ve druhém kole s výsledky, jichž po modifikacích dosáhli (Badryzlova et al., 2013), a s výsledky finálního, šestého kola měření u (Steen et al. 2010). Jak lze vidět, dosažený výsledek je plně srovnatelný s ostatními dvěma projekty, a dokonce je mírně nad hodnotou dosaženou autory VUAMC.

Aplikace MIPVU na češtinu	**Russian corpus of conceptual metaphor**	**VUAMC**
3 anotátoři 2463 tokenů	3 anotátoři cca 2000 tokenů	4 anotátoři 1921 tokenů
	(Badryzlova et al., 2013)	(Steen et al., 2010)
Test 2	**Test 2**	**Test 6**
0,88	**0,90**	**0,85**

Tabulka 16: Porovnání výsledků 2. kola měření IAA s Badryzlova et al., (2013) a Steen et al., (2010)

3.6 Alternativní tag NLE (No Literal Equivalent)

Důvodem zavedení alternativního tagu NLE *(No Literal Equivalent)* je skutečnost, že MIPVU je protokol zaměřený v první řadě na anotaci metafory pro účely corpus-based výzkumu, ale ne pro účely komputační lingvistiky. Pro každé z těchto dvou odvětví je z hlediska anotace metafory preferován odlišný způsob vymezení hranice mezi doslovným a metaforickým užitím lexikální jednotky.

Shutova (2015) uvádí, že VUAMC, jakožto v současnosti největší korpus anotovaný na metaforu (a tedy i korpus nejčastěji používaný k evaluaci automatických systémů identifikace metafory) obsahuje velké množství silně lexikalizovaných a konvencionalizovaných metafor (597). Pro reálně využitelné NLP aplikace má podle

autorky ovšem daleko větší smysl zabývat se pouze takovými metaforami, s jejichž interpretací si neporadí současné WSD systémy. Přítomnost silně lexikalizovaných nebo zcela mrtvých metafor v trénovacím korpusu vede ke snížení efektivity celého systému automatické identifikace těch opravdu relevantních případů metafory (616). Stejný problém je reflektován i autory a koordinátory projektu VUA Metaphor Shared Task (2018) (Shutova, Beigman Klebanov, & Leong, 2018), v rámci kterého byly porovnávány systémy automatické identifikace metafory ve VUAMC. Z anotovaného datasetu VUAMC, na základě něhož byly systémy evaluovány, byla vyfiltrována pouze autosémantika a všechny druhy metafor byly zařazeny pod jeden typ, tzn. pouze binární rozdělení na MRW/nonMRW (57). Tento přístup vypovídá o tom, že v kontextu NLP systémů na identifikaci metafory jsou ze slovních druhů autosémantika podstatně důležitější než např. tolik problematické a zároveň na metaforu bohaté předložky.

Abychom reflektovali tyto poznatky o potenciálních nevýhodách původního MIPVU pro účely komputační lingvistiky a zároveň plně využili jeho kvalit na poli korpusové lingvistiky (jedná se o jediný takto komplexní a všeobecně uznávaný manuál pro identifikaci metafory), rozhodli jsme se současně[36] s druhým kolem testování IAA opatřit všechny čtyři texty dodatečnou druhou vrstvou anotace s tagem NLE, a to následujícím způsobem:

- Po anotaci první vrstvy byly vyfiltrovány všechny případy MRW s jednomyslnou shodou všech tří anotátorů a po vzoru Shutova et al. (2018) byla vyřazena všechna synsémantika.
- Tyto případy byly trojicí anotátorů individuálně posouzeny následujícím způsobem (opět se jednalo o binární úlohu a k posouzení bylo možno použít zmíněnou dvojici slovníků): pokud k danému MRW existuje ekvivalent v doslovném jazyce, označit MRW číslem 1 a pokud neexistuje, označit MRW nulou (=NLE).

Nespornou výhodou oddělení vrstvy anotace metafory podle MIPVU a vrstvy s NLE tagem je fakt, že nedochází k narušení původního protokolu a jsou tak zachována statistická data o výskytu metafory v textu (která lze využít např. v komparativních korpusových studiích). Tato modifikace je navíc zcela nezávislá na jazyku a lze ji tedy využít v jakémkoli podobném výzkumu.

[36] Při anotaci celého korpusu by tento krok bylo nejvhodnější provést až po celkové adjudikaci datasetu, případně současně s ní.

Text		(…) běhali	po	zámku	a	přikládali.
Anotace metafory	Anotátor 1	1	0	0	0	1
	Anotátor 2	1	0	0	0	1
	Anotátor 3	1	0	0	0	1
NLE	Anotátor 1	**1**	-	-	-	**0**
	Anotátor 2	**1**	-	-	-	**0**
	Anotátor 3	**1**	-	-	-	**0**

Tabulka 17: Příklad anotace pomocí tagu NLE

Úryvek v tabulce č. 17 obsahuje dva případy MRW (slovesa *běhali* a *přikládali*), z nichž u jednoho bylo anotátory identifikováno, že pro něj existuje doslovný ekvivalent, a u druhého, že neexistuje:

Kontextuální význam slovesa *běhat* (podle SSJČ „rychle chodit v různých směrech; muset konat mnoho pochůzek při obstarávání, zařizování něčeho“) je evidentně v kontrastu s významem základním (SSJČ: „rychle jít, utíkat, spěchat“), nabízí se zde na základě údajů dostupných ze dvojice slovníků ekvivalentní, byť dnes již archaický výraz *šukat* (který ale díky románu *Babička* od Boženy Němcové stále přežívá v povědomí současného mluvčího) (SSJČ: „pohybovat se z místa na místo (těkat); vykonávat (zprav. drobné) práce“).

Kontextuální význam slovesa *přikládat* (podle SSJČ „přidáním něčeho zvětšit původní množství (zvl. topiva)“) je též v kontrastu s významem základním (SSJČ: „položit na něco n. blízko k něčemu“). V tomto případě ale nebyl nalezen žádný doslovný ekvivalent, a tak bylo jednotce přiřazeno číslo 0 (NLE). Jinými slovy, mluvčí zde nemá jinou možnost, jak daný význam vyjádřit, aniž by použil metaforického výrazu.

Tagem NLE byly dodatečně anotovány všechny čtyři úryvky a následně byla vypočtena IAA pomocí Fleissova κ. Jak lze vidět v následující tabulce, dosažené hodnoty κ dokládají velmi dobrou úroveň reprodukovatelnosti pro všechny čtyři žánry.

Text	Fleissovo κ
Academic	0,97
Fiction	0,90
News	0,90
Spoken	1
Fleissovo κ celkem	**0,94**

Tabulka 18: Výsledek měření IAA pro NLE tag

Poměrně zásadní problém při aplikaci NLE tagu spatřujeme v tom, že dostupné slovníky pro český jazyk se vyznačují vysokou mírou výskytu archaismů a že dostatečně nereflektují současný stav jazyka. Jelikož při analýze a hledání doslovných významů nelze spoléhat na individuální intuice a práce se slovníkem je tedy nutností,

nezbývá než znovu zdůraznit, že možnost pracovat se slovníky založenými na korpusu by zde byla obrovským přínosem.

3.7 Shrnutí

Při přímé aplikaci MIPVU v nemodifikované podobě na texty v českém jazyce podle očekávání vyvstala řada problémů podobného rázu, jaké reportovali i výzkumníci v případě ruštiny (Bardzylova et al., 2013) a litevštiny (Urbonaitė, 2015).

Na základě důkladné analýzy chyb po prvním kole testování spolehlivosti bylo navrhnuto několik úprav protokolu, které by měly zaručovat jeho lepší aplikovatelnost na český jazyk. Po zapracování těchto modifikací proběhlo druhé kolo testování spolehlivosti, tentokrát na žánrově a velikostně srovnatelném vzorku textů jako u (Steen et al., 2010) a (Bardzylova et al., 2013), a došlo k rapidnímu zlepšení dosažené hodnoty IAA, která dokonce mírně překonává pomyslný benchmark nastavený autory MIPVU. Statistické testy tedy potvrdily, že navrhované modifikace protokolu pro český jazyk zajišťují srovnatelnou míru reprodukovatelnosti, jíž disponuje původní MIPVU pro angličtinu.

Jako nadstavba byl představen a otestován tag NLE (*No Literal Equivalent*), pomocí nějž je možné z dat anotovaných podle MIPVU vyfiltrovat ty případy MRW, které jsou obzvláště relevantní pro systémy automatické identifikace (nelexikalizované) metafory.

Dalším logickým krokem a zároveň námětem pro budoucí práci je předkládanou verzi MIPVU modifikovanou pro český jazyk aplikovat na dostatečně velký korpus, který by následně mohl být využíván nejen pro účely komputační lingvistiky.

Závěr

V práci byla charakterizována lingvistické anotace jako vědecká disciplína a byla ilustrována její role v kontextu NLP a komputační lingvistiky. Byly nastíněny aktuální trendy v anotaci sémantických jevů a nabyté teoretické poznatky byly aplikovány na konkrétním anotačním projektu, jenž byl oproti anglickému originálu modifikován pro použití na českém jazyce.

Blízká budoucnost se pravděpodobně ponese právě ve znamení modifikace prověřených anotačních protokolů, jež spolehlivě fungují v angličtině a dalších světových jazycích, na míru menších a typologicky odlišných jazyků, jež v mnohých případech nedisponují tak rozsáhlou a propracovanou lingvistickou infrastrukturou (korpusy, aktuální slovníky atp.). Jak bylo vidět na příkladu modifikace protokolu MIPVU pro češtinu, tento proces bude s velkou pravděpodobností během na dlouhou trať, protože i v případě českého jazyka, jenž patří mezi jazyky s dlouhou a soustavně pěstovanou lingvistickou tradicí, jsme se setkali s poměrně zásadními překážkami. Alternativou je vývoj jazykově nezávislých protokolů, který je ale minimálně v oblasti sémantiky značně problematický, ne-li zcela nemožný.

Další oblastí, která v oblasti lingvistické anotace může do budoucna přinést velmi zajímavé výsledky, je crowdsourcing. Příklady z druhé kapitoly (entailment, sarkasmus) jsou důkazem toho, že vhodný design HITů a propracované a jasně formulované anotační instrukce mohou i v oblasti anotace sémantických jevů směle konkurovat klasickým manuálním metodám anotace při nižších nákladech a podstatně větší rychlosti.

Neméně důležité je i soustavné prohlubování spolupráce napříč anotátorskou komunitou, sdílení nejnovějších poznatků a dodržování nastavených anotačních a kódovacích standardů a publikování výsledků práce pod svobodnými licencemi.

Použité zdroje:

Ahn, L. V., & Dabbish, L. (2008). Designing Games With a Purpose. *Communications of the ACM,51*(8), 58-67. doi:10.1145/1378704.1378719

Artstein, R. (2017). Inter-annotator agreement. In *Handbook of linguistic annotation* (pp. 297-313). Springer, Dordrecht.

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics,34*(4), 555-596.

Badryzlova Y., Lyashevskaya O. (2017). Metaphor Shifts in Constructions: the Russian Metaphor Corpus. *AAAI Spring Symposium Series,* 127-130.

Badryzlova, Y., Shekhtman, N., Isaeva, Y., & Kerimov, R. (2013). Annotating a Russian corpus of conceptual metaphor: a bottom-up approach.

Bloomfield, L. (1933). *Language*. New York: H. Holt.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., & Gorrell, G. (2013). GATE Teamware: A web-based, collaborative text annotation framework. *Language Resources and Evaluation,47*(4), 1007-1029. doi:10.1007/s10579-013-9215-6

Bowman, S.R., Angeli, G., Potts, C., & Manning, C.D. (2015). A large annotated corpus for learning natural language inference. *EMNLP*.

Castro, S., Cubero, M., Garat, D., & Moncecchi, G. (2017). HUMOR: A Crowd-Annotated Spanish Corpus for Humor Analysis. *SocialNLP@ACL*.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,20*(1), 37-46. doi:10.1177/001316446002000104

Čermák, F. a Schmiedtová, V. (2004). Český národní korpus – základní charakteristika a širší souvislosti. *Library Revue*, *15*, 152–168.

Day, D.S., McHenry, C., Kozierok, R., & Riek, L.D. (2004). Callisto: A Configurable Annotation Workbench. *LREC*.

DBPedia. (n.d.). Retrieved February 14, 2019, from https://wiki.dbpedia.org/about

Design Principles. (n.d.). Retrieved from https://tei-c.org/Vault/ED/edp01.htm

Dipper, S., Götze, M., & Stede, M. (2004). *Simple Annotation Tools for Complex Annotation Tasks: an Evaluation*.

Estellés-Arolas, E., & González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, *38*(2), 189-200.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin,76*(5), 378-382. doi:10.1037/h0031619

Granger, S. (2009). *International Corpus of Learner English: Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š. and Žabokrtský, Z. (2018). *Prague Dependency Treebank 3.5*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN PID: http://hdl.handle.net/11234/1-2621.

Hladká, B., Mírovský, J., & Schlesinger, P. (2009). Play the Language: Play Coreference. *ACL/IJCNLP*.

Chmelař P., Hellebrand D., Hrušecký M., & Bartík V. (2011). Nalezení slovních kořenů v češtině. *Sborník příspěvků 10. ročníku konference Znalosti 2011*, 66-77.

Ide, N., & Suderman, K. (2014). The Linguistic Annotation Framework: a standard for annotation interchange and merging. *Language Resources and Evaluation, 48*, 395-418.

Ide, N., Calzolari, N., Eckle-Kohler, J., Gibbon, D., Hellman, S., Lee, K., Nivre, J., & Romary, L. (2017). Community Standards for Linguistically-Annotated Resources. *Handbook of Linguistic Annotation,* 113-165.

Ide, N., Chiarcos, C., Stede, M., & Cassidy, S. (2016). *Designing annotation schemes : From model to representation.*

Jurgens, D., & Navigli, R. (2014). It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *TACL*, 2, 449-464.

Karlík, P (2017). ENTAILMENT. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Khodak, M., Saunshi, N., & Vodrahalli, K. (2017). A Large Self-Annotated Corpus for Sarcasm. *CoRR, abs/1704.05579*.

Kroupová, L. et al. (2005). *Slovník spisovné češtiny pro školu a veřejnost: s Dodatkem Ministerstva školství, mládeže a tělovýchovy České republiky*. Praha: Academia.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., & Zasina, A.J. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. *LREC*.

Lakoff, G. (1987). The Death of Dead Metaphor. *Metaphor and Symbolic Activity,2*(2), 143-147. doi:10.1207/s15327868ms0202_5

Lakoff, G., Johnson, M. (1980). Metaphors We Live By. University of Chicago Press

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Little, W., Onions, C. T., Fowler, H. W., Coulson, J., & Friedrichsen, G. W. (1973). *The Shorter Oxford English Dictionary on Historical Principles*. Oxford: Clarendon Press.

Macmillan Education Ltd. (2002). *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan Education.

Markert, K., & Nissim, M. (2002). Towards a Corpus Annotated for Metonymies: the Case of Location Names. *LREC*.

Mpouli, S. (2017). Annotating similes in literary texts. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Nekula, M. (2017). IRONIE. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Nevěřilová, N. (2017): COMMON SENSE. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Pavlas, D., Vrabeľ, O., & Kozmér, J. (2018). Applying MIPVU Metaphor Identification Procedure on Czech.

pojmy:anotace. (2014, Nov 24). In Příručka ČNK. Retrieved 23:28, April 22, 2019, from https://wiki.korpus.cz/doku.php?id=pojmy:anotace&rev=1416826135

Pragglejaz Group. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol,22*(1), 1-39. doi:10.1080/10926480709336752

Ptáček, T., Habernal, I., & Hong, J. (2014). Sarcasm Detection on Czech and English Twitter. *COLING*.

Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for Machine Learning*.

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data Knowl. Eng., 74*, 1-12.

Shutova, E. (2015). Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics, 41*, 579-623.

Shutova, E., Beigman Klebanov, B., & Leong, C.W. (2018). A Report on the 2018 VUA Metaphor Detection Shared Task. *Proceedings of the Workshop on Figurative Language Processing*, 56–66.

Schmid, H. (2008). Tokenizing and Part-of-Speech Tagging. In *Corpus Linguistics: An International Handbook* (Vol. 1, pp. 527-551). Walter De Gruyter.

Snow, R., Oconnor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast---but is it good? *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP 08*. doi:10.3115/1613715.1613751

Steen, G. (2017). Identifying metaphors in language. In Semino, E., Demjén, Z. (Eds.) *The Routledge Handbook of Metaphor and Language* 73-87. London: Routledge.

Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). A method for linguistic metaphor identification. From MIP to MIPVU.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). Brat: A Web-based Tool for NLP-Assisted Text Annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics,*102-107.

Stubbs, Amber. (2011). *MAE and MAI: Lightweight Annotation and Adjudication Tools.* 129-133.

Štícha, F. et al. (2013). *Akademická gramatika spisovné češtiny*. Praha: Academia.

TEI: Text Encoding Initiative. (n.d.). Retrieved from https://www.tei-c.org/

Urbonaitė, J. (2015). Metaphor identification procedure MIPVU : an attempt to apply it to Lithuanian.

Veale, T., Klebanov, B. B., & Shutova, E. (2016). *Metaphor: A computational perspective*. San Rafael, CA: Morgan & Claypool.

Veselková, J. et al. (1986). *Mluvnice češtiny. 2, Tvarosloví*. Praha: Academia.

Veselovská, K (2017). POSTOJOVÁ ANALÝZA. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Veselovská, K., Hajič, J., & Šindlerová, J. (2012). Creating annotated resources for polarity classification in Czech. *KONVENS*.

Wave (noun). (n.d.). In *Macmillan English Dictionary Online*. Macmillan Education. Retrieved April 13, 2019, from https://www.macmillandictionary.com/dictionary/british/wave_1.

Williams, A., Nangia, N., & Bowman, S.R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *NAACL-HLT*.

Yimam, S.M., Gurevych, I., Castilho, R.E., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. *ACL*.