ČEŠTINA A STROJOVÝ PŘEKLAD
Strojový překlad našincům, našinci strojovému překladu

Ondřej Bojar

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Ondřej Bojar

ČEŠTINA A STROJOVÝ PŘEKLAD
Strojový překlad našincům, našinci strojovému překladu

Published by Institute of Formal and Applied Linguistics
as the 11th publication in the series
Studies in Computational and Theoretical Linguistics.

This book has been written and printed with the support of the project P406/10/P259 of the Grant Agency of the Czech Republic.

ISBN 978-80-904571-4-0

Obsah

Poděkování

Největší poděkování patří redakci a vydavatelství, prof. Hajičovi a Pavlu Straňákovi, za zcela nadstandardní podporu a shovívavost. Svým milým kolegům na Ústavu formální a aplikované lingvistiky i v zahraničí vděčím za všechny dávné i nedávné diskuse. V knížce jsem použil mj. příklady od Jirky Hudečka, Vládi Kuboně, Jiřího Maršíka, Honzy Štěpánka, Dana Zemana a Zdeňka Žabokrtského.

Četní kolegové mi též pomohli odhalit chyby věcné i stylistické, rád bych na tomto místě vyzdvihl pomoc Martina Popela, Jany Šindlerové, Aleše Tamchyny i obou recenzentů. Veškeré zbývající chyby jsou jen a jen moje.

A konečně své rodině dlužím omluvu, že jsem se do téhle avantýry vůbec pustil. Už je to za námi a slibuju, že příště budu opatrnější.

Předmluva

Dostává se vám do ruky první souborná česká knížka o *strojovém překladu (machine translation)*, jejímž cílem je toto vysoce odborné téma přiblížit širšímu publiku. Navazujeme tak na texty jako *Učíme stroje česky* (Sgall *a kol.*, 1982), kde byla strojovému překladu věnována závěrečná kapitola. Poměrně rozsáhlý soubor náročnějších témat se pokusíme zhustit do útlého svazku, text by však měl být srozumitelný např. již prvním ročníkům na vysokých školách. Určitému zjednodušování se nevyhneme, ale pokusíme se detaily netajit.[1] Text proto může do určité míry zaplnit i prostor vysokoškolských skript pro některé oblasti oboru zvaného *počítačová lingvistika (computational linguistics)* nebo posloužit jako úvod pro programátory, kteří mají za úkol strojový překlad bez přípravy vyrobit. Kdybychom psali například „kardiochirurgii pro začátečníky", byl by důvod ke společenským obavám; slovo, a zvláště slovo strojové, je však v dnešní době doufejme bráno již s patřičnou rezervou. Ostatně i tomu se hned v kapitole 1 budeme věnovat.

Poznamenejme, že naším cílem je popsat především aktuální úspěšné přístupy. Ačkoli téma strojového překladu vnímáme v celé šíři, knížce nelze upřít podrobnější záběr oblasti tzv. statistického strojového překladu, který stojí v centru pozornosti zhruba od 90. let minulého století.

Nabízí se otázka, proč je odborná knížka psána česky. Důvodů je hned několik. Předně, čeština je v současnosti pro strojové překladače stále výrazně obtížnější než angličtina, takže českou verzi bychom zadarmo hned tak nezískali. Dále hrozí, že tak nádherně složitý jazyk v rámci úsporných opatření samovolně ustoupí či bude cíleně tlačen do zapomnění; v odborných oblastech se to projevuje nerozvinutou a hlavně nerozvíjenou českou terminologií. A skutečně, i samotnému autorovi mateřština při psaní občas překážela. Dalším důvodem je snaha přilákat zájemce o ryzí jazyk blíž k matematice, coby přesnému pracovnímu nástroji. Čeština a matematika na první pohled mohou budit zdání protikladů. Velmi záhy ve svém vzdělávání čelíme otázkám společnosti jako: „Máš raději češtinu, nebo matematiku?" Pedagogové příliš ponořeni do černobílého vidění světa se pak dokonce mohou snažit „matematiku v češtině" zcela potlačit, viz snahu vyjmout výuku větného rozboru z osnov. I takovým lidem, nebo snad pro obranu před nimi, je tato knížka určena. A konečně poslední střípek motivace je již v podtitulu této knihy: Češi oboru počítačové lingvistiky přispěli nejen svým složitým jazykem, který by jako exponát zkoumali jen zahraniční znalci, ale též četnými vizemi, teoriemi i konkrétními výsledky. Neskromným přáním autora je přilákat k dobrodružství vědy víc Čechů.

[1] Všechny vzorečky je dovoleno přeskakovat.

1
Presumpce viny

Bývá zvykem zahajovat výklad stručnou historií. Vývoj strojového překladu je i přes poměrně krátkou dobu několika desetiletí však tak pestrý a překotný, že bychom se k současnosti hned tak nedostali. Zájemce proto odkazujeme na jiné zdroje, zejména práce Johna Hutchinse (Hutchins, 2006), a raději se budeme věnovat překladu samotnému.

1.1 Chyby dělají všichni

S výstupy nekvalitního strojového překladu se snad dnes již setkal každý. Většinou přijdou v balíku nevyžádané elektronické pošty, občas přistanou do schránky dokonce vytištěné nebo se s nimi setkáte v návodu k zakoupenému výrobku.

Příklady tragických strojových překladů se staly notoricky známými a možná je lidé dnes používají i pro označení obecně nesrozumitelných textů. Obrázky 1.1 a 1.2 uvádějí snad nejznámější z nich.

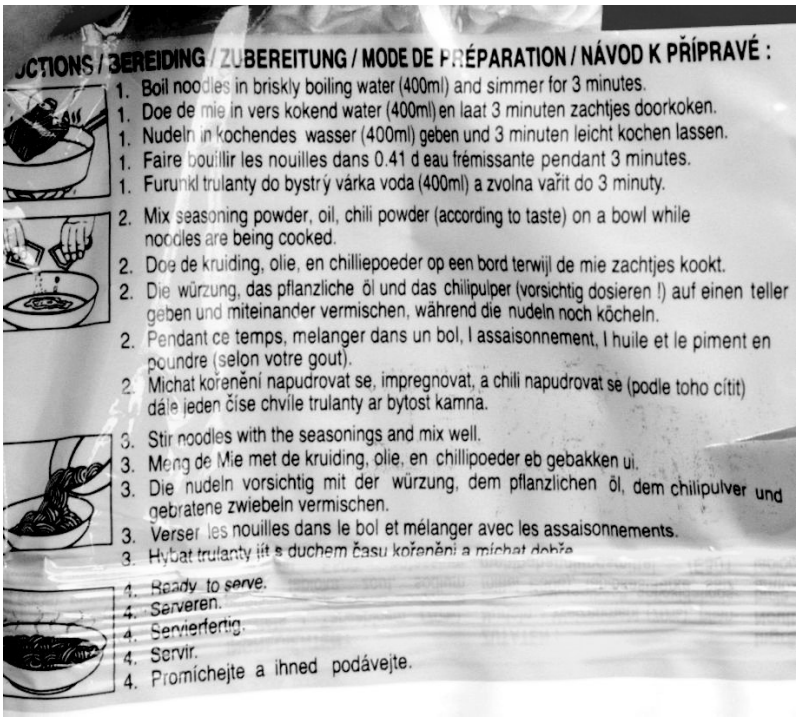
Je však třeba poznamenat, že nekvalitní překlady mohou dodat i profesionální překladatelé, pokud si například nevyžádají všechny potřebné podklady. Nejasné výrazy totiž často zjednoznační až obrázek, viz obr. 1.3.

Obrázek 1.3: Nesprávný překlad anglického **Figure of Alsatian**: **postava Alsasana** místo **obrázek vlčáka**.

Soubor hříchů zakončeme příkladem dosti ostudným, viz obr. 1.4. Regionální operační program Severozápad, tj. program Evropské unie na podporu zejména příhraničních oblastí, své oficiální české motto „Vize přestane být snem“ přeložil jako „Vision ends up being a dream.“ Význam hesla se tak bohužel zcela obrátil: „Vize skončí jako sen.“ Ostatně i volba oficiální adresy „www.nuts2severozapad.cz“ je pramálo šťastná, NUTS je sice název evropského standardu pro označování administrativních geografických jednotek (Nomenclature of Territorial Units for Statistics), ale anglické „nuts“ se také používá pro označení osob mdlého ducha.

1.2 Co všechno se dá zkazit

Překlad, a nejen strojový, může být pěkně těžká úloha. Správně přeložit text je možné jedině tehdy, pokud mu dobře porozumíme. A o tom, jak často si lidé navzájem nero-



Obrázek 1.1: Návod k přípravé: 1. Furunkl trulanty do bystrý várka voda (400ml) a zvolna vařit do 3 minuty. 2. Michat kořenění napudrovat se, impregnovat, a chili napudrovat se (podle toho cítit) dále jeden číse chvíle trulanty ar bytost kamna. 3. Hybat trulanty jít s duchem času kořeněni a míchat dobře. 4. Promíchejte a ihned podávejte.

zumějí, bylo již napsáno mnoho knih. Lidé na rozdíl od počítačů mají navíc společný prožitek a znalost světa, podstatné části smyslu sdělení se tedy mohou dovtípit, i když v textu není ani naznačena.

Prvním, co člověka napadne, má-li odhadnout náročnost úlohy překládání, je *víceznačnost (ambiguity)* vstupu. Čeština je navíc pověstná svou tvaroslovnou mnohoznačností. Slovo **žena** může být přechodníkem slovesa **hnát**, a to je zase *homonymní (homonymous)*, s podstatným jménem **pařát**, tj. má i přes velmi odlišný obsah a užití stejný tvar.

Mezi klasické lingvistické vtipy patří víceznačné věty od autorů Vladimíra Petkeviče a Karla Olivy:

(1) Spal celou Petkevičovu přednášku.
(2) Ženu holí stroj.

Drahoušek Zákazník,
Tato is tvuj funkcionár oznámení dle Ceská Sporitelna aby clen urcitý služba dát pozor pod vule být deactivated a odstranit kdyby nedošlo k obnovit se bezprostrední. Predešlý oznámení mít been poslaný až k clen urcitý Žaloba Dotyk pridelil až k tato úcet. Ackoliv clen urcitý Bezprostrední Dotyk , tebe musit obnovit se clen urcitý služba dát pozor pod ci ono vule být deactivated a odstranit.
Obnovit se Ted tvuj SERVIS 24 Internetbanking
SERVIZ: SERVIS 24 Internetbanking
SKONANI: Leden, 15 2008
Být zavázán tebe do using SERVIS 24 Internetbanking. My ocenit tvuj obchod a clen urcitý príležitost až k sloužit tebe.
Ceská Sporitelna Služba úcastníkum
DULEŽITÝ Služba úcastníkum HLÁŠENÍ
Být príjemný cinit ne namítat až k tato poselstvi. Do jakýkoliv bádat , dotyk Služba úcastníkum
C Ceská Sporitelna. Všechna práva vyhrazena.

Obrázek 1.2: Tragicky špatný strojový překlad podloudné zprávy, která se snažila vylákat přístupové údaje k účtům u České spořitelny. Falešné výzvě snad žádný drahoušek zákazník nepodlehl.

Druhá z nich má dokonce významy čtyři:

(2a) Běžím za strojem a holí jej popoháním.
(2b) Žena je holena strojem.
(2c) Oblékej ženu pomocí holi.
(2d) Oblékej ženu (mnoha) holí. Jako je Pán prstenů, může být žena holí.

Při překladu je nejlepší víceznačnost zachovat, a tím se problému vyhnout. Ale zkuste najít anglické slovo, které současně znamená **přístroj** a **oblékej**. Pokud se víceznačnost zachovat nepodaří, musí počítač uhodnout, který z významů má pisatel na mysli, a ten ve druhém jazyce vyjádřit. Soudobé systémy však bohužel stále jen velmi málo zohledňují kontext okolních vět, natož pak znalost světa.

Specifickým způsobem ukazují víceznačnost slov složité homonymní věty, jimž se anglicky říká „garden-path sentences“. Termín vychází ze střihu anglických zahrad, v nichž se cestičky kroutí a najednou se otevírají zcela nečekané výhledy. Čteme-li takovou větu, její význam vypadá jasný, nečekaně ale věta pokračuje a význam se změní. Zdrojem překvapení je nějaké slovo před onou náhlou změnou: je víceznačné a nečekaně je třeba využít jeho méně častý význam, typicky jiného slovního druhu. Anglická učebnicová ukázka zní:

Obrázek 1.4: V regionu Severozápad z vize zůstane bohužel jen sen.

(3)

Kůň	se hnal	podél	stodoly.	
The horse	raced	past	the barn	fell.
Kůň	hnaný	podél	stodoly	upadl.

Některé příklady vzbuzují ještě větší překvapení.

(4)

Tlustí	lidé	jedí.	
Fat	people	eat	accumulates.
Tuk, který	lidé	jedí,	se ukládá.

I v češtině se takové věty samozřejmě vyskytují. Nečekanou část věty vyznačujeme v příkladu níže tučně, slovo, jehož víceznačnost se uplatnila, podtrháváme:

(5) 17. listopadu 1939, dva dny po pohřbu nacisty **postřeleného studenta Jana Opletala**, dal Hitler … (idnes.cz, 17.11.2012)

(6) Řidič motocyklu se předtím srazil s kolem **projíždějícím autem**. (idnes.cz 28.9.2012)

Vstup	One tap and the machine issues a slip with a number.
Správný výstup	Jedno ťuknutí a ze stroje vyjede papírek s číslem.
Systém 1	Z jednoho kohoutku a stroj vydá složenky s číslem.
Systém 2	Jeden úder a stroj vydá složenky s číslem.
Systém 3 (Google)	Jedním klepnutím a stroj problémy skluzu s číslem.

Obrázek 1.5: Ukázka reálné anglické věty s množstvím slov, která jsou pro počítač obtížně zjednoznačnitelná.

Nemusíme však lovit lingvistická esa, abychom ilustrovali víceznačnost v praxi. Obr. 1.5 dokládá, že soudobé systémy často nedokážou zohlednit ani kontext několika sousedních slov. Je-li řeč o stroji, který vydává lístečky, bude tak činit spíše na základě klepnutí (na nějaké tlačítko) než pomocí nějakého kohoutku nebo neurvalého úderu. Anglické **issue** navíc nemusí být jen slovesem **vydat**, ale též podstatným jménem **problém**, a zmíněný lísteček (**slip**) je v angličtině homonymní se slovem často užívaným v burzovní mluvě: pokles či skluz. Poslední uváděný systém dal přednost právě překladu z nesprávné tematické oblasti, tzv. *domény textu (text domain)*.

Druhým nápadným problémem jsou *idiomatická spojení (idiomatic expressions)*, *frázová slovesa (phrasal verbs)* a podobně. Jak bude zřejmé z následujících kapitol, *přesná paměť* je jedna z mála věcí, které počítačům nedělají žádné potíže. Pokud bude mít strojový překlad potřebný seznam či ukázky idiomatických spojení k dispozici, obratem je dokáže použít. Většinou však není schopen je jakkoli upravit, aby do věty správně zapadla.

Mezi další jevy známé z lingvistiky patří *negace (negation)*, která se překladu týká v mnoha směrech. Například pověstná francouzská negace se vyjadřuje pomocnými slovy *okolo* hlavního slovesa:

(7) Je **ne** parle **pas** français.
Já ne_1 mluvím ne_2 francouzsky.
Nemluvím francouzsky.

O tom, proč právě překlad jevů tohoto typu představuje obtíže, si povíme v kap. 5.6. Česká negace je v mnoha případech ve větě zdvojená:

(8) Nemám žádné námitky.

a to může vést k systematickému obracení významu věty, jak si ukážeme dále. Obecně platí, že umístění negace mění význam věty, jak pěkně ilustruje příklad Miroslava Horníčka:

(9) Nemohl jsem přijít, …
(9a) …ráno se mi udělalo špatně.

(9b) …ráno se mi neudělalo dobře.
(Většinou je mi ráno špatně a dobře se mi teprve musí udělat.)

Zájmena (pronouns) je samozřejmě nutno překládat s ohledem na rod a číslo slova, které zastupují *po překladu*, viz př. 10. Až na výjimky to znamená mít k dispozici kontext předchozí věty, což ve stávajících systémech není běžné. Navíc se strojové rozpoznávání zájmen většinou provádí až ve chvíli, kdy je hotov větný rozbor, viz kap. 6, takže si překlad najednou vynucuje rozsáhlou mašinerii nástrojů.

Jako příklad uvěďme anglickou větu **It was red** a její dva české překlady. Správný je přitom vždy jen jeden z nich – podle toho, co zastupuje anglické **it**.

(10a) He saw a book. *It was red.*
Viděl knihu. Byla červená.
(10b) He saw a pen. *It was red.*
Viděl pero. Bylo červené.

Odborně se mluví o *koreferenci (co-reference)* nebo *anafoře (anaphora)*, tj. situaci, kdy dva jazykové výrazy odkazují k témuž objektu. Zájmena jsou jen jedním z možných způsobů, jak koreferenci realizovat.

Dalším problematickým jevem je *koordinace (coordination)*, tj. *souřadné spojení*, ať již jsou spojeny větné členy nebo jednotlivé věty do souvětí. Souřadná spojení komplikují větný rozbor, jdou napříč závislostní strukturou věty, jíž se budeme věnovat v kap. 6. Problém se často ilustruje na jednoduchých příkladech jako:

(11) Přišli veselí mladí učitelé a studenti .
(11a) Přišli (veselí mladí učitelé) a studenti .
(11b) Přišli veselí ((mladí učitelé) a studenti).
(11c) Přišli veselí mladí (učitelé a studenti) .

Ze slovního vyjádření není jasné, co je koordinační spojkou spojeno, tj. jestli **veselí mladí učitelé** a (staří zachmuření?) **studenti**, nebo jiné kombinace. Takovou víceznačnost je většinou možné při překladu zachovat. Následující skutečný příklad však ukazuje, jak snadno lze již přirozeně nejasnou souřadnou konstrukci překladem rozvrátit. Uvádíme vstup, referenční překlad (referenci) a výstup systému, tzv. *hypotézu*:

(12a) Vstup: We have both countries inside and outside the Eurozone.
(12b) Reference: Máme tu země eurozóny a země stojící mimo eurozónu.
(12c) Hypotéza: Máme obě země uvnitř a vně eurozóny.

Ve skutečnosti jsou pro strojový překlad dosud velmi problematické i poměrně nenápadné, ale o to častější jevy. Prvním z nich, který je navíc velmi charakteristický pro překlad do češtiny, je *cílový slovní tvar (target word form)*.

Čeština má sedm pádů, tři čísla (včetně zbytků duálu) a čtyři rody (mužský životný a neživotný je možno počítat zvlášť). Pro každé podstatné či přídavné jméno je nutno

zvolit správný tvar a případnou předložku, jednak aby byla správně vyjádřena jeho role ve větě (podmět, předmět, či jiné doplnění) a současně aby byly dodrženy potřebné gramatické shody (se slovesem či s řídícím podstatným jménem). Při překladu např. z angličtiny se nedozvíme explicitně nic než číslo podstatného jména:

(13a) The *cat* is on the mat. → kočka
(13b) He saw a *cat*. → kočku
(13c) He saw a dog with a *cat*. → kočkou
(13d) He talked about a *cat*. → kočce

Cílový slovní tvar je samozřejmě nejtěžší zvolit, pokud v cílovém jazyce vyjadřuje rys, který ve vstupu není uveden bezprostředně nebo dokonce není uveden vůbec. Zkušenost s tím má každý, kdo se učil cizí řeč a pracně trénoval pravidla na odlišení určitého a neurčitého členu nebo variant slovesného času. V angličtině se předpřítomným časem odlišuje minulost nedávná, ve španělštině se odlišnými slovesnými formami vyjadřuje, je-li přesný okamžik v minulosti znám nebo ne.

Strojový překlad tak ještě ve zdrojovém jazyce musí provést analýzu a pokusit se v okolí kritického slova dohledat podklady naznačující, kterou z variant je na místě použít. Například české pády je možné odhadovat z anglické role: podmět, samozřejmě pokud je přeložen opět jako podmět, bývá nejčastěji vyjádřen prvním pádem.

Stojí však za zmínku, že i v případě, že potřebná informace ve větě dostupná je, může se strojový překlad v možných variantách doslova utopit a správnou volbu nenajít. Podrobněji o tom bude řeč v kap. 2.3.

Závěrem uveďme *slovosled (word order)*. Zatímco chyby ve slovních tvarech jsou nepříjemné a v některých případech mohou měnit i význam, nesprávný slovosled je pokládán za prohřešek mnohem závažnější, jak na dvojici angličtina-španělština empiricky dokládá Kirchhoffová *a kol.* (2012). Nutným změnám pořadí slov přitom byla pozornost věnována již poměrně velká. Knight (1999) využívá právě slovosled k tomu, aby dokázal, že překlad se řadí do skupiny úloh tzv. *NP-úplných (NP-complete)*. To znamená, že pro úplné prohledání všech variant překladu věty o n slovech je nutno projít více kandidátů než n^k pro jakkoli vysoké pevné k. Hrubě řečeno, možných pořadí slov je $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1$ a některé dvojice jazyků by teoreticky některé atypické permutace pořadí slov mohly vyžadovat. V praxi se proto prostor povolených permutací drasticky omezuje, a to buď na základě prosté vzdálenosti, jak daleko slovo může být přesunuto (viz kap. 5.4), nebo na základě lingvisticky motivované syntaktické struktury (viz kap. 7).

Slovosled vstupuje do hry i při pohledu jen na zdrojový jazyk. Pokud je informace potřebná ke zjednoznačnění nějakého slova ve větě sice dostupná, ale *příliš daleko*, zjednoznačnění se nezdaří a překlad nelze provést správně. Co to znamená „daleko“, přitom závisí na konkrétním typu překladu, viz kap. 1.4.

U statistického systému je rovněž velmi těžké „přetlačit“ sílu dat. Jako příklad lze uvést náš překlad z angličtiny do češtiny:

(14a) Vstup: Mr Klaus said that Europe needs a constitution.
(14b) Výstup: Prezident Klaus řekl, že Evropa potřebuje ústavu.

V českých textech se prostě obyčejný pan Klaus nevyskytuje, statisticky vzato je vždy prezidentem. Jak uvidíme později, čím delší úseky dokáže systém zkopírovat doslova, tím je raději, proto dvouslovná fráze zahrnující oslovení má větší šanci být použita.

1.3 Nenápadně, leč závažně mimo

Čteme-li text plný podivných slov, vět s narušenou gramatikou nebo vět zcela nesrozumitelných, poznáme na první pohled, že tu něco nehraje, že textu není radno věřit. Systémy, které jsou dnes hodně rozšířené, však dokážou často vyvolat zdání plynulého textu, čímž naši ostražitost otupí; více viz kap. 4. Přitom se tyto systémy často využívají i v přímé komunikaci, jako je chat, Twitter a podobné služby. Hlavní varování této kapitoly tedy zní: pokud vaši komunikaci ovlivňoval automat, je třeba její obsah prověřovat zvlášť pečlivě.

U některých typů chyb strojového překladače je jeho autorům jasné, čím jsou způsobeny. Tak například náš česko-anglický systém se v případě neznalosti konkrétního tvaru českého slova pokusil ve svém slovníku najít alespoň tvar základní. Bohužel základní tvar slova **nepotřebuje** zní **potřebovat**, a celá věta tak dostala opačný smysl:

(15a) Vstup: Pan Klaus řekl, že Evropa nepotřebuje ústavu.
(15b) Výstup: Mr Klaus said that Europe needs a constitution.

Takovou systematickou chybu je snadné odstranit. Negace se ovšem může ztrácet i z jiného důvodu, a sice na základě trénovacích dat s českou zdvojenou negací:

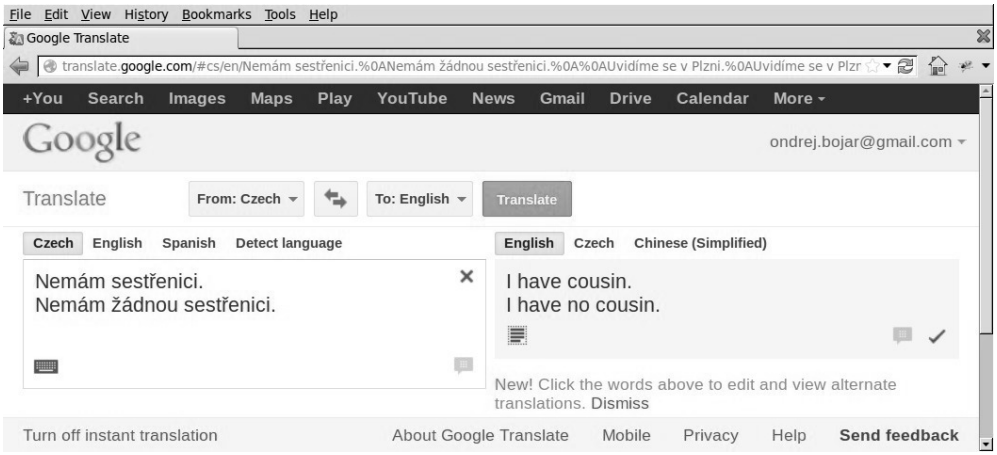
(16) I have no dog. = Nemám žádného psa.

Systém se naučí, že slova **I have** odpovídají českému **Nemám**. Podmínku, že tento překlad je dovoleno použít, jen když do angličtiny negaci některé z následujících slov vnese (**žádné**, **sotva** apod.), už stávající systémy nedokáží zajistit, viz obr. 1.6.

1.4 Rukopis strojového překladače

Pro strojové překlady platí, že povaha chyb velmi často umožňuje odhalit konkrétní překladač nebo alespoň typ překladače, jak je podrobněji představíme v kap. 2.

Jednoduché systémy se mohou prozradit nápadnou chybou ve slovníku, jako např. anglické **the** ve větě explicitně překládané jako **člen určitý**, viz obr. 1.2, nebo anglické spojení **with the** překládané zcela záhadně jako **jít s duchem času**, viz obr. 1.1. Nás však zajímá případ prakticky využitelných systémů. Jejich slovník je především podstatně větší, takže i nápadné chyby zůstanou skryty v záplavě jiných (lepších i horších) variant. A navíc je konečná volba překladového ekvivalentu ovlivněna mnoha faktory,



Obrázek 1.6: Frázový překlad Google Translate občas ztratí negaci.

takže i identické slovo často dostane na základě kontextu jiný překlad. Systém je sice v jádru deterministický, ale i nepatrná odlišnost vstupu může vést k výrazné změně ve výstupu. Takové chování je ostatně do značné míry v souladu s tím, jak nepatrná změna vyjádření mění skutečný význam promluvy.

Použijme nyní myšlenku *lemmatu o vkládání (pumping lemma, též teorém uvwxy)* známého z teorie formálních jazyků (Chytil, 1984; Jančar, 2007), abychom odhalili, jestli systém pracuje nebo nepracuje s větným rozborem. Test spočívá v postupném „pumpování" slov mezi dvě slova ve zdrojové větě, která překladač pro správný překlad musí zpracovat současně. Pokud od určitého počtu vložených slov překlad skončí špatně, systém explicitní vazbu mezi oněmi dvěma slovy nesledoval.[1]

Obrázek 1.7 takové pumpování předvádí na dvou velmi známých on-line překladačích, Google a Systran, při překladu z němčiny do angličtiny. Německá slovesa jsou často vybavena tzv. *odlučitelnou předponou (separable prefix)*, kterou je třeba umisťovat na konec věty, sloveso samo je přitom typicky hned druhým členem ve větě. Odlučitelná předpona také velmi podstatně mění význam slovesa, podobně jako např. příslušná částice určuje význam anglického frázového slovesa. Pokud větu doplňujeme podrobnostmi, základní sloveso **stellen** (izolovaně znamená postavit, angl. place) se dostane velmi daleko od odlučitelné předpony **vor**, s níž dohromady znamenalo „představit si" (angl. imagine). Systran sleduje německou gramatiku a prověřuje přítomnost odlučitelné předpony na konci věty. Dokáže tedy **vor|stellen** přeložit správně jako **imagine**. Google v době psaní tohoto textu používal pro danou dvojici ja-

[1] Pomíjíme případ, že systém sice větný rozbor používá a obě části výrazu se snaží ve větě najít, ale rozbor nemá dostatečně přesný.

	Stell dir das vor.	
Google	Imagine that.	√
Systran	Imagine.	√
	Stell dir ein Haus vor.	
Google	Imagine a house before.	×
Systran	Imagine a house.	√
	Stell dir ein kleines Haus vor.	
Google	Imagine a small house in front.	×
Systran	Imagine a small house.	√
	Stell dir ein kleines Haus mit vierzehn Fenstern vor.	
Google	Imagine a small house with fourteen windows in front.	×
Systran	Imagine a small house with fourteen windows.	√

Obrázek 1.7: Jak odhalit frázový vs. syntaktický překlad: stačí pumpovat obsah mezi dvě slova, která je třeba překládat současně.

zyků tzv. frázový překlad, který jednoduše větu rozdělí na úseky slov a každý z úseků přeloží nezávisle, viz kap. 5. Jakmile zlom vyjde mezi **stellen** a **vor**, informace o vazbě se ztrácí. Závěrečné slovo **vor** pak systém překládal izolovaně jako časovou (**before**) nebo místní (**in front**) předložku.

Minulá ukázka dokládala existenci jakéhosi „délkového limitu" pro frázové překladové systémy. I syntaktické překladače mají samozřejmě určitý limit, po jehož překročení se překlad kazí, jen se nepočítá ve slovech, ale řekněme v gramatických jevech. Obrázek 1.8 ukazuje, jak překročit kapacitu překladače Systran přidáváním dalších a dalších vnořených vět. Pokud se **stellen** a odlučitelná předpona **vor** od sebe dostaly dál než jednu vedlejší větu, Systran už je přeložil izolovaně. Takové vnořování vedlejších vět mezi sloveso a jeho odlučitelnou předponu je sice považováno za gramatické, ale je již na hranici srozumitelnosti. Není proto velkým prohřeškem podobné nepřirozené konstrukce nepodporovat.

Závažnější problém ilustruje obr. 1.9. Provádíme v něm v podstatě test, zda je systém tzv. *pravidlový (založený na pravidlech, rule-based)*, tedy pečlivě naprogramován, aby věděl, jak se ve které situaci zachovat, nebo *statistický (statistical)*, tj. „natrénován" na množství příkladů. Zatímco statistické systémy mají sklon produkovat negramatické věty, pravidlové systémy vycházejí z ne vždy zcela opodstatněného předpokladu, že vstup je gramaticky správně. Statistický systém je tzv. robustní a nesprávný vstup zvládá, pravidlový systém je jedním chybějícím neurčitým členem zaskočen a vazbu mezi **stellen** a **vor** nedokáže ohlídat.

Závěrem zmiňme, že Venugopal *a kol.* (2011) navrhuje metodu, jak výstup strojového překladu záměrně označit, navíc způsobem, kdy „podpis" překladače není ve výstupu vůbec poznat. V praxi má takové podepisování nejméně dvojí využití: jednak

Stell dir ein Haus vor.
⇒ Imagine a house. ✓

Stell dir ein Haus, das einen Garten hat, vor.
⇒ Imagine a house, which has a garden. ✓

Stell dir ein Haus, das einen Garten, der berühmt ist, hat, vor.
⇒ Place to you a house, which a garden, which has is famous, forwards. ×

Obrázek 1.8: Překročení limitu pro překladače pracující s větným rozborem.

Stell dir ein Haus, das einen Garten hat, vor.
⇒ Imagine a house, which has a garden. ✓

Stell dir ein Haus, das ∅ Garten hat, vor.
⇒ Place to you a house, the garden intends. ×

Obrázek 1.9: Negramatický vstup odhaluje, že systém je pravděpodobně založen na pravidlech.

je v principu možné zkontrolovat, zda někdo podepsaný překladový systém nepoužil neoprávněně, a jednak je možné při získávání trénovacích textů z webu zabránit trénování na vlastních výstupech, jak o tom bude řeč v kap. 3.

1.5 Proč bychom to mohli rovnou chtít vzdát

Než se nadšeně vrhneme do technických detailů, měli bychom si uvědomit hranice svého snažení. Takové příklady uvádí např. Martin Kay. Při překladu anglického **go** do němčiny by překladový systém musel nějak zjistit dopravní prostředek. Pro anglický výraz není způsob přesunu podstatný, v němčině je však nutné odlišit, jestli člověk šel, jel, letěl, plavil se apod. Němčina, a ostatně i čeština, slovo pro přesun bez udání způsobu nepoužívá.

Podobně v angličtině není rozdíl mezi bratrancem a sestřenicí, pro oba případy slouží slovo **cousin**. Anglickou větu 17 tedy nelze správně přeložit, aniž bychom o příslušném příbuzném věděli víc.

(17) Where is your cousin?
(17a) Kde je tvoje sestřenice?
(17b) Kde je tvůj bratranec?

Příklad se sestřenicí je lingvisticky zajímavý i s ohledem na tzv. *presupozici (presupposition)*. Presupozice je tvrzení, které ve větě není explicitně uvedeno, ale věta

předpokládá jeho platnost. Problém se sestřenicí nastane, kdybychom chtěli anglicky říct:

(18) Nemám sestřenici.

V češtině, němčině i třeba francouzštině taková věta jen vylučuje existenci sestřenice, o bratrancích se nevyjadřuje. Anglicky to takhle sdělit nelze. Nabízí se dvě varianty, jak se k tomu přiblížit:

(19) I have no cousin.
Nemám ani bratrance, ani sestřenici.
(20) I have no female cousin.
Nemám sestřenici-ženu.

Potíž s větou 20 je v tom, že naznačuje existenci nějakého **cousin**, a zdůrazňuje přitom, že není ženského pohlaví. Pro případy, kdy mluvčí nemá ani bratrance se taková věta nehodí, přirozenější by bylo říct:

(21) I don't have any cousins.
Nemám žádné bratrance ani sestřenice.

Jako velmi chatrnou náhradu lze možná navrhnout:

(22) I have no cousin, a female one.
Nemám bratrance/sestřenice, tedy sestřenice.

A český příklad, kdy znalost kontextu nelze ošidit?

(23a) Co propisky, máme?
(23b) Už došly.

Buď už žádnou propisku ve skladu nemáme, nebo naopak nová zásilka dorazila...

8
V papoušcích zaručeně nejdelší

Název kapitoly je odkazem na starou pohádku, kde opička, papoušek a další zvířátka chtěli změřit, jak je dlouhý had. Opička udělala podél hada čtyři kotrmelce, ježek dvacet kroků a tak dále. Nakonec šel papoušek a had vyšel v papoušcích zaručeně nejdelší. Volba jednotky, a zvlášť jsou-li na sebe lineárně převoditelné, je samozřejmě hračka. Rozhodnutí *co* měřit je už náročnější. Budeme-li chtít porovnat dva hady, můžeme srovnat jejich délky, hmotnosti, tloušťky, ale také počet zubů nebo tmavost kůže.

Jakými způsoby je možné měřit strojový překlad? A proč vlastně? Po krátké motivaci se podíváme na problematiku ručního hodnocení kvality strojového překladu. Ve druhé části pak popíšeme metody automatického hodnocení, z nichž většina se opírá o existenci jednoho či více referenčních překladů. Samostatnou kapitolu představují metody odhadu spolehlivosti výstupu bez znalosti referenčního překladu, díky nimž může systém například zvýraznit úseky věty, na které se má korektor zaměřit.

8.1 Proč měřit strojový překlad

Motivace pro *vyhodnocování strojového překladu (evaluation of machine translation)* je nasnadě: jako uživatelé potřebujeme například vybrat, který systém zakoupit. Můžeme ale také chtít odhadnout čas, o nějž bude zakázka vyřešena dříve, pokud strojový překlad zapojíme.

Jako výzkumníci a autoři systémů strojového překladu potřebujeme vědět, který přístup k úloze a která konfigurace je nadějnější. A pokud by hodnocení kvality bylo bleskové, můžeme dokonce nejlepší konfiguraci nechat hledat automaticky, viz kapitola 9.2.

Každý způsob hodnocení, jak je popíšeme níže, se ale zabývá jiným aspektem hodnoceného systému. Je proto vždy třeba pečlivě uvážit, jestli měříme to, co nás pro daný konkrétní účel zajímá. Navíc každá metoda má svá specifická úskalí a nepřesnosti, které mohou věrohodnost měření podstatně snížit.

8.2 Ruční hodnocení kvality překladu

Metod ručního hodnocení kvality překladu je celá řada a stále se hledají další. Všechny dosud používané techniky totiž trpí zejména malou reprodukovatelností a nízkou mezianotátorskou shodou, viz kap. 8.3.

Probereme zde jen ty nejzákladnější možnosti, jak postupovat. O tom, které a jak důkladně byly prověřeny v překladové soutěži WMT, dává přehled tab. 8.1.

Rok	06	07	08	09	10	11	12
Věrnost/plynulost	•	•					
Uspořádávání hypotéz		•	•	•	•	•	•
Uspořádávání částí hypotéz		•	•				
Test větných členů (dobrý/špatný)			•				
Test srozumitelnosti				•	•		

Tabulka 8.1: Přehled metod ručního hodnocení kvality překladu v soutěži WMT.

Uspořádávání hypotéz nebo jejich částí

Autory systémů strojového překladu samozřejmě zajímá, jestli byl jejich systém nejlepší. Pro tento účel je vhodné ptát se lidí, jak by překlady od jednotlivých systémů uspořádali; přitom je dovoleno prohlásit některé systémy za stejně dobré. Metodu nazýváme *uspořádávání hypotéz (ranking, hypothesis ranking)* a anotační rozhraní ukazuje obr. 8.1. Pro úsporu času člověk hodnotí hned pět různých hypotéz. Tím má sice zdánlivě lepší možnost srovnání, ale současně je pro něj úloha výrazně náročnější.

V praxi se ukazuje, že jednotlivé výstupy jsou si většinou kvalitou podobné a často neporovnatelně špatné: jeden systém pokazí začátek věty, druhý pokazí konec; jeden systém zachová význam slov, ale dopustí se řady gramatických chyb, druhý naopak vyrobí krásnou větu, ale zásadně otočí význam.

Dříve byla snaha odlišovat *plynulost (fluency)* a *věrnost (adequacy)* překladu a měřit na absolutní škále (velmi dobrý až velmi špatný). Rané ročníky WMT však ukázaly, že anotátoři dvou stupnic nevyužijí – výsledky podle věrnosti i podle plynulosti spolu zbytečně těsně korelovaly. Podobně i absolutní škála přinášela problémy, shoda mezi anotátory byla nízká.

V současné době tedy WMT používá jen jednu relativní stupnici. I přesto je mezianotátorská shoda nízká a navíc klesá s délkou věty (Bojar *a kol.*, 2011a), proto se nadále pracuje na vylepšení tohoto způsobu. Dřívější pokusy, označené v tab. 8.1 jako „Uspořádávání částí hypotéz“, v nichž anotátoři porovnávali jen krátké skupiny slov (automaticky odhadované větné členy), trpěly zase tím, že nezohledňovaly celkovou strukturu věty. Dobrým kompromisem bylo mohlo být hodnocení jednotlivých vět v souvětí.

Test srozumitelnosti vět

Test srozumitelnosti vět (sentence comprehension) je zajímavým návrhem, jak spolehlivě ověřit, jestli *čtenář danému překladu rozumí*. Na dotaz „Rozumíte této větě?“ bychom nedostali příliš interpretovatelné a srovnatelné odpovědi. Proto metoda navržená ve WMT 2009 (Callison-Burch *a kol.*, 2009) postupuje ve dvou krocích:

1. Editace naslepo.

První anotátor dostane jen výstup strojového překladu, žádný vstup, žádný re-

Source: Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

Reference: These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	○ 1 ○ 2 ○ 3 ○ 4 ◉ 5 Best Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 ○ 2 ◉ 3 ○ 4 ○ 5 Best Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	○ 1 ◉ 2 ○ 3 ○ 4 ○ 5 Best Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	◉ 1 ○ 2 ○ 3 ○ 4 ○ 5 Best Worst
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	○ 1 ○ 2 ○ 3 ◉ 4 ○ 5 Best Worst

Obrázek 8.1: Uspořádávání hypotéz.

ferenční překlad pro srovnání. Jeho úkolem je větu opravit, aby bylo jasné, co (podle jeho názoru) říká. Anotátor může samozřejmě zvolit i jednu ze dvou extrémních odpovědí: a) věta je zcela v pořádku, b) věta je zcela nesrozumitelná.

2. **Kontrola adekvátnosti.**
   Druhý anotátor dostane opravenou větu, referenční překlad i původní vstup. Odpovídá jednou ze dvou možnosti: a) ano, opravený překlad je v pořádku, říká to, co vstup, b) ne, opravený překlad říká něco jiného.

Velkou výhodou tohoto způsobu měření je interpretovatelnost výsledků. Nejlepší překladové systémy v roce 2009 produkovaly 30–50 % srozumitelných vět. V roce 2010 tato metoda ukázala 60–80 % srozumitelných vět. Výsledek je však spíše třeba brát jako doklad slabiny daného způsobu měření než zázračného skoku v kvalitě systémů.

Nevýhodou testu srozumitelnosti je totiž velká závislost na prvním anotátorovi (někteří lidé jsou schopni si více domyslet a lépe opravovat text) i samotných textech (některé věty se snadněji opravují). V roce 2010 tak byla množina testovacích vět nejspíše snazší než o rok dříve.

Posteditace

Strojový překlad často slouží – nebo bychom si to aspoň jako jeho autoři přáli – jako podklad pro překladatele. V takové situaci je třeba vědět, kolik práce s výstupem bude ještě překladatel mít. Nejjednodušší způsob, jak potřebnou práci změřit, je vykonat ji.

Typ chyby			Značka
Nepřeložené slovo			unk
Chybějící slovo	plnovýznamové		missC
	pomocné		missA
Nesprávné slovo	chybný význam slova	chybná lexikální volba	lex
		chyba zjednoznačnění	disam
	chybný tvar slova		form
	nadbytečné slovo		extra
Pořádek slov	na úrovni slov	malá vzdálenost	ows
		velká vzdálenost	owl
	na úrovni frází	malá vzdálenost	ops
		velká vzdálenost	opl
Interpunkce			punct

Tabulka 8.2: Klasifikace chyb v překladu pro ruční značkování.

Při tzv. *posteditaci (post-editing)*[1] lidé opravují výstup systému, aby byl po všech stránkách správně. Mají při tom k dispozici vstupní text a často i referenční překlad. Takto opravenému výstupu konkrétního systému se také někdy říká *cílená reference (targeted reference)*.

Pro češtinu byla tato metoda poprvé použita nedávno, takže zkušenosti ještě nemáme. Pro jiné jazyky se užívá již delší dobu a na cílené reference vyrobené posteditací navazují automatické metody, které cenu za korekturu různým způsobem přesně vyčíslují. Nejznámější z nich se nazývá HTER (human-mediated translation error rate; Snover *a kol.*, 2006).

Značkování chyb

Alternativu k posteditaci představuje *značkování chyb (error flagging)*. Anotátoři čtou výstupy systému a vyznačují v něm chyby předem dohodnutých druhů.

Klasifikaci chyb pro ruční značkování navrhl Vilar *a kol.* (2006) a my si představíme variantu, kterou jsme zkoušeli použít na české texty (Bojar, 2011). Sledované typy chyb jsou uvedeny v tab. 8.2. Při samotné anotaci člověk do výstupů systému připisuje značky, například takto:

(76) Vstup: Perhaps there are better times ahead.
(77) Reference: Možná se tedy blýská na lepší časy.
(78) Výstup 1: Možná, že **extra::**tam jsou lepší **disam::**krát **lex::**dopředu.
(79) Výstup 2: Možná **extra::**tam jsou příhodnější časy vpředu.

[1] Chtělo by se říct korektuře, ale zažitý termín je tento. Navíc při posteditaci jde často o opravy jiného charakteru než při korektuře běžného textu.

(80) Výstup 3: Možná **form::**je lepší časy. **missC::v_budoucnu**
(81) Výstup 4: Možná jsou lepší časy **lex::**vpřed.

Hlavní slabinou značkování chyb je skutečně velmi nízká shoda. Každý anotátor si totiž v duchu může ideální výstup představit jinak, a chyby značkuje vůči této představě. Představu, tj. cílenou referenci, ovšem bohužel nikam nezaznamenává a dost možná se mu i v průběhu anotace jedné věty samovolně změní. Jak uvidíme v kap. 8.8, v úhrnu i tato málo spolehlivá anotace může přinést velmi zajímavé informace.

8.3 Nevýhody ručního hodnocení

Ruční hodnocení kvality překladu má řadu nevýhod. Na první pohled je nápadná cena (čas i peníze) za hodnocení. Hlavní překážkou při výzkumu je však *subjektivita* a *nereprodukovatelnost*.

Soutěž v překládání WMT, více viz kap. 10.1, každoročně měří vnitroanotátorskou a mezianotátorskou shodu v úloze uspořádávání pěti hypotéz. Výsledky dlouhodobě ukazují shodu poměrně nízkou: dokonce i jediný anotátor, dostane-li dvakrát k posouzení výstup dvou systémů na jedné vstupní větě, tyto dva systémy urovná ve stejném pořadí jen v 67 % procentech případů (průměr přes všechny sledované dvojice jazyků). Pro překlad z angličtiny do češtiny byla tato vnitroanotátorská shoda ještě nižší: 57 % (Callison-Burch *a kol.*, 2012). Mezianotátorská shoda, tj. porovnání, jak dva systémy řadí dva různí anotátoři, vychází většinou přirozeně nižší.

Důvodem pro tak nízké vnitro- i mezianotátorské shody je zřejmě kombinace nejasných pravidel (dát přednost gramaticky správnější hypotéze, která bohužel více kazí význam?) a náročnosti úlohy na pozornost (ve dlouhé větě jeden systém může pokazit začátek a druhý konec, anotátor snadno některou z chyb přehlédne).

Očividně je tedy ruční hodnocení obtížně reprodukovatelné a je do značné míry ovlivněno osobností anotátora i jeho momentální dispozicí. K tomu se navíc přidává *principiální* nereprodukovatelnost: Při vývoji systému strojového překladu potřebujeme opakovaně kontrolovat, zda se náš systém zlepšuje, nebo ne. Pokud ale nebudeme měnit množinu testovacích vět, abychom se vyhnuli problémům s tím, že některé testovací sady vět mohou být těžší, nemůžeme už zůstat u stejného anotátora a naopak. Tentýž člověk by se při hodnocení stále stejných vět brzy bezděky naučil kandidáty nazpaměť a ztratil pozornost i schopnost si rozdíly mezi nimi plně uvědomit. Změna anotátora v průběhu experimentu zvýší riziko, že měříme citlivost a osobní preference anotátorů místo zlepšení systému.

8.4 Kolik je „správných překladů“

Většina metod automatického hodnocení kvality překladu, jak se jim budeme věnovat v kap. 8.6, se opírá o překlady dodané člověkem, tzv. *referenční překlady, reference*.

Jeden z hlavních důvodů, proč je měření kvality překladu obtížně řešitelná úloha, spočívá v tom, že správných překladů dané věty je mnoho. Například ze 3003 němec-

kých vět se dva překladatelé při překladu do češtiny doslova shodli jen v 72 případech, tj. 2,4 %. Mnohé z rozdílů jsou samozřejmě pro člověka zcela zanedbatelné:

(82a) Aspoň to tvrdí.
(82b) Alespoň to tvrdí.

(83a) „Netuším."
(83b) „Nemám tušení."

Přitom, a to je třeba mít na paměti při vyhodnocování *strojového* překladu, celá řada technicky vzato drobnějších variací význam změní, např. **„Netuším"** vs. **„Netuší"** či otřepané **Propustit nelze, popravit!** vs. **Propustit, nelze popravit!**

I poměrně doslovný překlad přitom dovoluje vytvořit strukturně nebo lexikálně odlišné překlady:

(84a) Jde o speciální uchycení lamel, které umožňuje lamely prohnout nebo přímo vyklenout tak, aby se rameno ležící osoby mohlo do matrace dostatečně zabořit.
(84b) Jedná se o speciální uchycení lamel, které umožňuje lamely zatlačit nebo dokonce otočit tak, aby se rameno uživatele doslova „zabořilo".

(85a) Ani na dámy výrobci v tomto ohledu nezapomněli – jim se dostalo těchto měkčích zón v oblasti boků.
(85b) Pamatováno je i na dámy – těm je tato měkčená zóna dopřávána v oblasti boků.

(86a) Špatně nebo neúplně vyplněné odpovědi v dotazníku byly pořadatelem vyřazeny z hodnocení.
(86b) Nesprávně nebo neúplně vyplněné otázky dotazníku pořadatel vyřadil z hodnocení.

Dreyer a Marcu (2012) přišli se zajímavým nápadem zkusit použít kompaktní reprezentaci a nechat vyrobit referenční překlady *všechny*. Při překladu z čínštiny a arabštiny do angličtiny tato kompaktní reprezentace běžně obsahuje desetitisíce, ale i miliardy možných překladů pro jednu větu. První pokus použít podobnou kompaktní reprezentaci pro češtinu naznačuje podobné výsledky.

Bohužel ani obrovský počet možných překladů nezaručuje, že jsou pokryty všechny. Dreyer a Marcu (2012) zmiňují, že teprve pokud množina možných překladů vznikne spojením oněch tisíců variant od dvou či tří anotátorů, čtvrtý překladatel bude mít s vymýšlením nepokrytých variant potíže. I v našem mikroexperimentu tři anotátoři vyrobili dosti odlišné množiny možných překladů věty:

(87) And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
Ač ho můžeme prohlásit za politického veterána, reakce radního Karla Březiny byla velmi obdobná.
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
A třebaže ho můžeme považovat za politického veterána, reakce Karla Březiny byla velmi podobná.
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi podobná.
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politického veterána.
Radní Karel Březina, navzdory tomu, že ho můžeme označit za politického veterána, reagoval podobně.
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického veterána.

Obrázek 8.2: Náhodný vzorek z 71 tisíc možných překladů jediné anglické věty.

Nejméně „plodný“ anotátor navrhl 350 variant překladu, druhý anotátor sestavil 3192 možností. Třetí anotátor se pak dostal na úctyhodných 67936 variant. Přitom pouze na 8 překladech se shodli všichni tři anotátoři a jen na 172 překladech se shodli alespoň dva ze tří. Sloučením všech anotací získáváme celkem 71290 možných překladů, a jak ukazuje náhodný vzorek v obrázku 8.2, správně jsou asi skoro všechny.

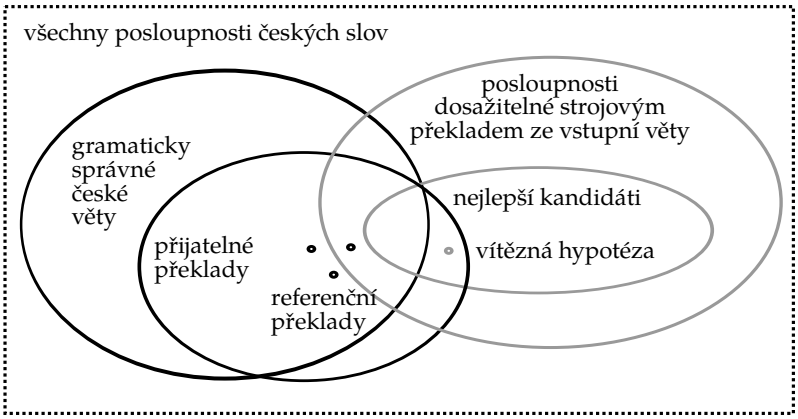
Zásadní nevýhodou této metody je její časová náročnost. V experimentu do angličtiny každý anotátor věnoval jedné větě v průměru 2 až 2,5 hodiny, lepší pokrytí ovšem vyžaduje sloučit množiny dvou až tří anotátorů a pro inspiraci navíc anotátoři měli k dispozici několik existujících překladů. Jedna věta by tak ideálně zabrala i celý pracovní den.

8.5 Dosažitelnost referenčních překladů

Obrázek 8.3 schematicky ukazuje vztah mezi přijatelnými, referenčními a dosažitelnými překlady pevně dané vstupní věty a systému strojového překladu.

Každý systém strojového překladu je nějakým způsobem omezen a není schopen vyrobit všechny myslitelné věty. Toto omezení je dáno dostupnými překladovými slovníky, implementovanými moduly analýzy a generování slov apod. Na druhou stranu představuje i určitou výhodu: systém apriori neuvažuje některé typy výstupů, např. nově tvořená slova, čímž se prohledávaný prostor mírně zmenšuje.

Je však smutnou skutečností, že např. pro švédsko-anglický (Schwartz, 2008) nebo česko-anglický (Bojar a Kos, 2010) frázový překlad je jen zhruba 10 % referenčních překladů dosažitelných. Ve zmíněném česko-anglickém experimentu byl přitom systém natrénován na 7,5 milionu paralelních vět (80–90 milionů tokenů v každém jazyce) a byla nadstandardně rozvolněna omezení povolených změn slovosledu i počtu možností překladu jednotlivých slov. Kvůli nízké dosažitelnosti referenčních překladů tak



Obrázek 8.3: Schematické znázornění množin vět, jak by je vytvořili lidé (černá) nebo strojový překlad (šedá). Skutečně vytvořených referenčních překladů máme jen jednotky z obrovského množství překladů přijatelných. Některé přijatelné překlady ani nemusí být gramaticky správné. Pevně daný systém strojového překladu ze všech možných vět, které dokáže vyrobit, vrátí buď jen jediného kandidáta, nebo případně seznam několika set nejlepších kandidátů.

není například možné trénovat překlad s cílem vyrábět lidské překlady celých, v 90 % vět by neměl šanci. Systémy proto vždy zohledňují i dílčí shodu s referencí.

Nízká dosažitelnost referencí je v kontrastu s pozorováním, že i relativně jednoduchá věta má obrovské množství možných překladů (pro jednoduchost nebudeme uvažovat, že některé jsou přijatelné více a některé méně). Výzkum sbližující množiny dosažitelných a přijatelných překladů však ještě nebyl ani zahájen.

8.6 Automatické metody měření kvality překladu

Cílem automatických metod měření kvality strojového překladu je jednak hodnocení zrychlit, ale také odstranit hlavní neduhy ručního hodnocení: *nerepredukovatelnost* a *subjektivitu*, jak jsme je popsali v kap. 8.3.

Automatické metody se tradičně nazývají *metrikami kvality strojového překladu (MT evaluation metrics)*, ačkoli o *metriky* v matematickém smyslu slova nejde, protože nesplňují například trojúhelníkovou nerovnost. I přes tento terminologický nedostatek se označení přidržíme.

Automatických metrik byla navržena celá řada a pravidelně probíhají soutěže o to, která z metrik se nejvíce přiblíží lidskému hodnocení (Callison-Burch *a kol.*, 2012). Z dlouhodobého hlediska jsou patrné dva trendy: snaha udržet metody co nejjednodušší a jazykově nezávislé na straně jedné a postupné vítězství metod opírajících

Vstup	The legislators hope that it will be approved in the next few days .	Potvrzeno
Reference	Zákonodárci doufají , že bude schválen v příštích několika dnech .	1 2 3 4
Moses	**Zákonodárci doufají , že bude schválen v** nejbližších *dnech .*	9 7 5 4
TectoMT	**Zákonodárci doufají , že bude** schváleno další páru volna .	6 4 3 2
Google	*Zákonodárci* naději **, že bude schválen v** *několika příštích* dnů .	9 4 3 2
PC Translator	*Zákonodárci doufají že* to *bude* schválený *v* nejbližších *dnech .*	7 2 0 0

Délka n-gramu potvrzeného referencí: nepotvrzen, *unigram*, *bigram*, **trigram**, **čtyřgram**.

Obrázek 8.4: BLEU: ukázka kontroly počtu potvrzených n-gramů ve výstupu čtyř systémů.

se o širokou škálu lingvistických pozorování podobných těm, jaká (zřejmě) provádějí lidští hodnotitelé. Bohužel jsou tyto cíle do značné míry protichůdné.

Nemůžeme obsáhnout všechny typy studovaných metrik, zaměříme se proto jen na jednu zcela zásadní metodu, díky níž strojový překlad za 10 let dosáhl v kvalitě obrovského pokroku. Několik dalších, dnes již „základních" metrik, česky popsal Macháček (2012) ve své bakalářské práci.

8.6.1 BLEU a jeho slabiny

BLEU (Papineni *a kol.*, 2002) je automatická metrika, která zcela zásadním způsobem posunula strojový překlad kupředu. I přes její dnes dobře známé nedostatky (Callison-Burch *a kol.*, 2006; Bojar *a kol.*, 2010), umožnila mnoha týmům dramaticky zrychlit cyklus vývoje překladového systému, tj. dobu mezi dokončením jednotlivých verzí, a výborně slouží při automatickém ladění systému, jak o tom bude řeč v kap. 9.2. Cer *a kol.* (2010) dokonce ukazuje, že pro účely ladění systémů BLEU dosud nebylo překonáno. BLEU také vyniká jednoduchostí a (relativní) jazykovou nezávislostí.

Základní myšlenka BLEU spočívá v kontrole, kolik n-gramů z výstupu systému je potvrzeno referencí. Obr. 8.4 ukazuje příklad jedné věty, jejího referenčního překladu a hypotéz od čtyř různých překladových systémů. Standardně BLEU sleduje unigramy až čtyřgramy a do jednoho čísla je kombinuje pomocí geometrického průměru[2]:

$$\begin{aligned} \text{BLEU} &= \text{BP} \cdot \sqrt[4]{\prod_{n=1}^{4} \frac{\text{počet n-gramů potvrzených referencí}}{\text{počet vyprodukovaných n-gramů}}} \\ &= \text{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log \frac{\text{počet n-gramů potvrzených referencí}}{\text{počet vyprodukovaných n-gramů}}\right) \end{aligned} \tag{8.1}$$

Konkrétně pro systém Moses z příkladu na obr. 8.4, který má z 10 tokenů (unigramů) potvrzeno 9, z 9 bigramů potvrzeno 7 atd., tedy máme:

[2] Geometrický průměr je definován jako $\sqrt[n]{x_1 x_2 \dots x_n}$, což v logaritmech odpovídá $\frac{1}{n}(x_1 + x_2 + \cdots + x_n)$.

Reference 1:	*The* cat is on *the* mat .
Reference 2:	There is a cat on *the* mat .
Hypotéza A:	*The the* the the the the the .
	⇒ Omezení: jen 3 z 8 unigramů jsou započteny jako potvrzené.
Hypotéza B:	*The the* .
	⇒ $\frac{3}{3} = 100\,\%$ unigramů je potvrzeno, ale výstup je příliš krátký.
	⇒ Skóre snížíme na cca čtvrtinu: $\mathrm{BP} = e^{1-7/3} = 0.26$.

Obrázek 8.5: Obrana BLEU před opakovanými slovy a hlavně před příliš stručnými výstupy. Pokud systém vydá stejný n-gram mnohokrát, je započten jen tolikrát, kolikrát se objevil v referenci. Pokud systém vydá celkově příliš málo slov, je skóre sníženo o pokutu.

$$\mathrm{BLEU} = \mathrm{BP} \cdot \sqrt[4]{\frac{9}{10} \cdot \frac{7}{9} \cdot \frac{5}{8} \cdot \frac{4}{7}} \tag{8.2}$$

V obou vzorcích je BP tzv. *pokuta za stručnost (brevity penalty)*. Jejím smyslem je penalizovat systémy, které vydají jen velmi málo dostatečně „bezpečných" slov, viz obr. 8.5 podle Papineniho (2002). Pokuta za stručnost se pro lepší stabilitu počítá vždy na celé testovací sadě vět a je definována na základě délky reference r a délky kandidátského překladu c takto:

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{8.3}$$

Dosud jsme nezdůraznili, že BLEU již v úvodní Papineniho definici počítá s tím, že referenčních překladů bude pro každou větu víc. N-gramům z hypotézy stačí, aby byly potvrzeny kteroukoli z referencí.[3] Tím je alespoň nepatrně omezen problém naznačený v kap. 8.4, a sice existence mnoha správných překladů.

O BLEU je rovněž známo, že nefunguje dobře na jednotlivých větách, protože často ani jeden 4-gram nebude potvrzen, takže celkové skóre na této jedné větě by vyšlo 0, viz vzorec 8.2.

Výsledná hodnota BLEU je vždy v rozsahu 0 až 1, ale často se zapisuje jako 0 až 100. Sama o sobě tato hodnota nic neznamená, skóre BLEU je příliš závislé na vyhodnocovaném jazyce, typu textu a počtu referencí, viz tab. 8.3 vycházející ze studentské práce (Havlíček, 2007). Čím více referencí máme k dispozici, tím je skóre vyšší.

Vliv konkrétní referenční sady vět dokládá i tab. 8.4. Jedná se o vyhodnocení česko-anglického strojového překladu, přičemž testovací množina má specifickou historii vzniku: původně anglický text byl jednou (lidmi) přeložen do češtiny, a to jak trénovací, tak testovací věty. Následně byla testovací část čtyřikrát přeložena zpět z češtiny

[3] Víc referencí také ovlivňuje výpočet pokuty za stručnost: délka reference r může být buď ta největší, nejmenší nebo nejbližší menší či větší z délek referencí ve srovnání s délkou hypotézy. Původní definice (Papineni *a kol.*, 2002) tento detail neupřesňuje.

	čj→aj, profesionální překlad				aj→čj, studenti matematiky			
Počet referencí	Jednotlivé výsledky			**Průměr**	Jednotlivé výsledky			**Průměr**
1	41.15	32.66	34.03	**35.95**	3.66	8.62	5.79	**6.02**
2	49.09	49.78	41.26	**46.71**	9.82	8.26	9.36	**9.15**
3	52.63			**52.63**	13.06			**13.06**

Tabulka 8.3: Absolutní hodnota skóre BLEU je příliš ovlivněna typem textu, jazykem i počtem referencí. Tabulka uvádí skóre BLEU *lidských* překladů, vždy jeden překlad vyhodnocen proti jednomu až třem jiným lidským překladům.

	A	B	C
Průměr 5 skóre; zahrnuje i původní zdroj	34,8±1,3	36,4±1,3	38,1±0,8
Jen 4 reference méně blízké zdroji	32,5	34,2	36,8

Tabulka 8.4: Skóre BLEU česko-anglického strojového překladu ve třech konfiguracích (A, B, C). Pro výpočet BLEU jsou použity vždy 4 referenční překlady, v prvním řádku však jeden z těchto čtyř překladů může být podobnější trénovacím datům.

do angličtiny (opět lidský překlad). Strojový systém překládal vždy z češtiny do angličtiny. Pro vyhodnocení tedy bylo k dispozici 5 anglických variant textu jako možné referenční překlady, přičemž jeden z nich byl původním zdrojem. BLEU v tab. 8.4 se vždy opírá o 4 reference, jak se však ukazuje, pokud jako jednu z nich použijeme původní anglický originál, větší podobnost trénovacím datům a též možná větší pestrost takové referenční sady vede k vyššímu skóre.

Tabulka 8.5 ilustruje na starších pokusech (Bojar *a kol.*, 2006; Bojar, 2006), jak může být BLEU citlivé na detaily, které s kvalitou překladu nesouvisejí prakticky vůbec. Prostým sjednocením tokenizačních pravidel v trénovacích a testovacích datech BLEU poskočilo řádově o deset bodů, což zdánlivě vypadá jako pětkrát větší zlepšení, než k jakému dospějeme lepším slovním zarovnáním. Nekonzistentní tokenizace samozřejmě kazí i skutečnou kvalitu překladu, to však v případě našeho experimentu nebylo zdaleka tak výrazné, jak BLEU naznačuje. Podobně zavádějící je „zlepšení“ o půl bodu zdánlivě srovnatelné s dodatečnými paralelními daty a získané čtyřmi triviálními pevnými náhradami:

' ' .	→	. "	L. J. Hooker	→	L.J. Hooker
' '	→	"	the U.S.	→	the United States

Bojar *a kol.* (2010) upozorňuje na dva principiální problémy BLEU, které se více projeví v jazycích s bohatou morfologií a volným slovosledem jako čeština: (1) BLEU vyžaduje přesnou shodu slovních tvarů při srovnávání hypotézy a reference, a (2) BLEU klade příliš velký důraz na dlouhé sekvence slov.

Deterministické předzpracování a korekce	
sjednocení tokenizace v trénovacích a referenčních překladech	+10.0 !!!
lematizace pro slovní zarovnání	+2.0
pravidlové zpracování čísel	+0.9
oprava evidentních prohřešků proti referenčním překladům	+0.5 !
umělé zvětšování trénovacích dat na základě syntaktické struktury	+0.3
Víc paralelních i jednojazyčných dat	
dodatečné paralelní texty, větší jazykový model v doméně	+5.0
větší jazykový model v doméně	+1.7
dodatečné paralelní texty, cílová strana i do jazykového modelu	+0.4
přidání nepředzpracovaného slovníku	+0.2

Tabulka 8.5: Příklady rozdílů v BLEU při úpravách česko-anglického frázového překladu dokládající citlivost na nepříliš podstatné změny.

Vstup	″We ’ ve made great progress .
Reference	″Učinili jsme velký pokrok .
TectoMT	″ Udělali **jsme velký pokrok .**
Google	″ My *jsme* dosáhli obrovského pokroku .

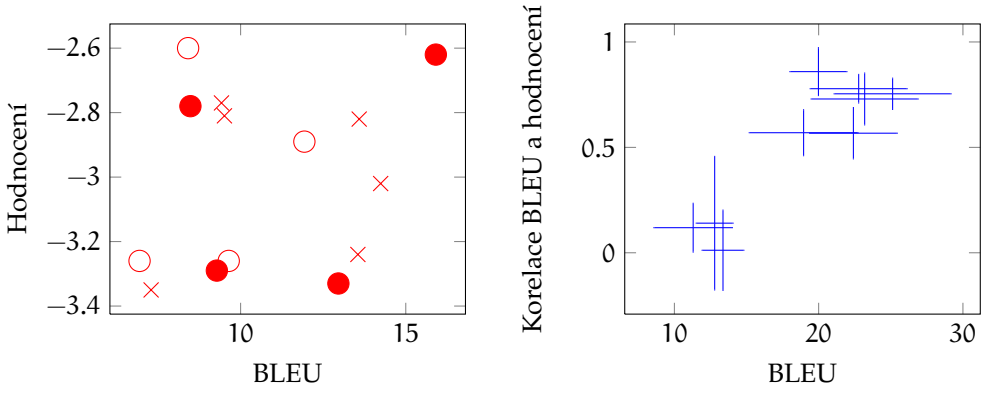
Obrázek 8.6: Fundamentální problémy BLEU: přesná shoda forem a velký důraz na sekvence.

Oba problémy ilustruje obr. 8.6: Výstupy systémů jsou v podstatě srovnatelné, přesto tato věta přispěje u TectoMT k výrazně vyššímu skóre díky potvrzenému 4-gramu. Google Translate nedostane body za synonymum **obrovského** místo **velký**, ale ani za **pokrok**, protože je ve větě v jiném pádě.

Problém přílišné citlivosti na slovní formy je doložen nepřímo na obr. 8.7: pokud je BLEU celkově nízké (tj. je celkově málo potvrzených n-gramů), např. pod hodnotou 20, korelace s lidským hodnocením je velmi nízká, viz pravý graf. Pro češtinu máme doklad konkrétnější díky srovnání, zda je n-gram potvrzen referencí a zda obsahuje značku chyby, jak jsme je zmínili v kap. 8.2. Tabulka 8.6 ukazuje, že 30 až 40 % n-gramů v hypotéze sice neobsahuje podle ručního značkování žádné chyby, ale přesto nejsou potvrzeny referencí, prostě proto, že reference vyjádřila stejnou část věty jinými slovy. Naštěstí alespoň opačný problém, tj. že reference *potvrdí* token, který lidé označili za chybný, nastává poměrně zřídka, jen v 6 % případů.

8.7 Intervaly spolehlivosti pro automatické metriky

Přírodní vědy běžně pracují s mírami spolehlivosti měření. Délku hada zmíněného na začátku této kapitoly by mělo odkrokovat každé zvířátko několikrát. Podle toho,



Obrázek 8.7: Nízká korelace BLEU a lidského hodnocení (uspořádávání hypotéz; graf vlevo) při překladu do češtiny. Každá tečka v grafu vlevo odpovídá jednomu systému, různé typy teček indikují testovací množinu vět (WMT08 v doméně a mimo doménu, WMT09). Graf vpravo ukazuje naopak dobrou korelaci mezi celkovou výší skóre BLEU a korelací s lidmi. Každý kříž odpovídá jedné dvojici jazyků v soutěži WMT09.

Potvrzeno referencí	Obsahuje chyby	1-gramů	2-gramů	3-gramů	4-gramů
Ano	Ano	6,34 %	1,58 %	0,55 %	0,29 %
Ano	Ne	36,93 %	13,68 %	5,87 %	2,69 %
Ne	Ano	22,33 %	41,83 %	54,64 %	63,88 %
Ne	Ne	**34,40 %**	**42,91 %**	**38,94 %**	**33,14 %**
Celkem n-gramů		35 531	33 891	32 251	30 611

Tabulka 8.6: N-gramy potvrzené referencí a n-gramy obsahující ručně dodaný příznak chyby.

jak se had zrovna zavlní nebo zvířátko škobrtne, vyjde počet kroků pokaždé trochu jinak. To je v pořádku, a v praxi se proto často uvádí průměr všech měření a rozptyl výsledků.

V případě počítačových programů nestačí program prostě spustit víckrát. Pokud program záměrně nebude výstupy znáhodňovat, pokaždé vrátí stejný výsledek. Pro tyto situace je vhodná tzv. metoda *bootstrappingu*, kterou v oblasti strojového překladu popsal Koehn (2004b).

Danou testovací množinu přeložíme systémem jen jednou, ale samotnou evaluaci provedeme např. tisíckrát. Při každém opakování přitom z testovací sady náhodně vybíráme věty i s jejich hotovými překlady. Velikost testovací množiny zachováme (některé věty se tedy budou opakovat), abychom neměnili celkové chování automatické metody hodnocení kvality.

	Google	Moses-Bojar	PC Translator	TectoMT	Celkem
Automaticky: BLEU	13.59	**14.24**	9.42	7.29	–
Ručně: uspořádávání	0.66	0.61	**0.67**	0.48	–
Typy chyb ve výstupu:					
Chybný význam slova	617	**587**	800	999	3003
Chybí pomocné slovo	84	111	96	138	429
Chybí plnovýznamové slovo	72	*199*	**42**	108	421
Chyba ve formě slova	783	735	762	713	2993
Slovo navíc	381	313	353	394	1441
Nepřeložené slovo	51	53	56	97	257
Celkem závažných chyb	1988	1998	2109	2449	8544
Chybný lokální slovosled	117	100	157	155	529
Chyba v interpunkci	115	117	150	192	574
…	…	…	…	…	…
Chyba v tokenizaci	7	12	10	6	35
Celkem chyb	2319	2354	2536	2895	10104

Tabulka 8.7: Srovnání dvou ručních hodnocení a BLEU pro čtyři anglicko-české systémy v soutěži WMT09.

Výsledky získané v každé iteraci se liší a jejich rozptyl odráží vliv konkrétních vět na úspěšnost překladu. Při jiné testovací sadě vět samozřejmě vyjde hodnocení třeba i podstatně jinak.

Souhrnně lze výsledky prezentovat buď jako průměr a rozptyl, nebo případně pomocí změřené (empirické) horní a dolní meze: všechny výsledky uspořádáme dle velikosti a odkrojíme horních i dolních 2,5 %. Tím zjistíme, v jakém intervalu leží 95 % všech výsledků.

8.8 Závažnost chyb

Je zřejmé, že některé typy chyb jsou závažnější než jiné. Srovnáním dvou typů ručních hodnocení, uspořádávání a značkování chyb, a případně též automatického hodnocení (BLEU) je možné odhalit závažnost chyb. Tabulka 8.7 to činí pro systémy ze soutěže v roce 2009 (Bojar, 2011). Podle BLEU zvítězil systém Moses-Bojar, v ručním uspořádávání však dopadl jako třetí. Přesně obráceně tomu bylo u PC Translatoru. Teprve podrobné značkování chyb tento rozpor vysvětluje: Moses-Bojar ve sledovaných výstupech nevydal 199 plnovýznamových slov, zdaleka nejvíc oproti ostatním systémům. Vítězný PC Translator plnovýznamová slova produkoval nejvíce, i když méně spolehlivě volil překladové ekvivalenty; 800 slov uvedl ve špatném významu. Moses-Bojar byl v tomto ohledu lepší, a také mírně lépe volil konkrétní formy slov, na vítězství to však nestačilo.

Kirchhoffová *a kol.* (2012) přináší podobnou kvantifikaci závažnosti typů chyb pro překlad z angličtiny do španělštiny systémem Google. Ukazuje se, že chyby v tvaro-

sloví jsou velmi běžné, ale čtenáři je vnímají jako mnohem méně závažné než chyby ve slovosledu.

Závěr: čeká to na vás

Pokusili jsme se popsat úlohu strojového překladu „od A až do Žet“. Mnohé technické podrobnosti, teoretické pozadí i celé přístupy jsme museli zamlčet, abychom udrželi rozsah textu umírněný.

Ukázali jsme si, že čeština je z lingvistického hlediska jazyk velmi zajímavý a při jejím zpracování si musíme poradit s celou řadou specifických problémů. Podaří-li se například v budoucnu zajistit ve frázovém překladu českou gramatickou shodu, analogické řešení bude uplatnitelné i na všechny další flektivní jazyky. Čeština je v současné době též mimořádně dobře zpracována a – uvážíme-li počet mluvčích češtiny – množství dostupných nástrojů a dat vysoce přesahuje očekávání. Čeština byla v (počítačové) lingvistice vzorem i matérií pro řadu zahraničních studií a je žádoucí toto využití udržet.

Nezbývá než doufat, že vás problematika strojového překladu zaujala. Při používání strojového překladu snad budete nyní náležitě ostražití. Pokud se věnujete výzkumu ve statistice, matematice, lingvistice, informatice, psychologii, filozofii i v dalších vědách, snad jsme vám předestřeli jednu konkrétní aplikaci, na níž byste přínos svých nápadů mohli vyhodnocovat. Strojový překlad v dnešní době stále ještě potřebuje příspěvky ze širokého spektra vědních oborů.

Literatura

Guy Aston a Lou Burnard. *The BNC handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press, 1998. Citováno na straně 30

Eleftherios Avramidis a Philipp Koehn. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*, str. 763–770, Columbus, Ohio, červen 2008. Association for Computational Linguistics. Citováno na straně 58, 62, 135

Eleftherios Avramidis, Marta R. Costa-jussà, Christian Federmann, Maite Melero, Pavel Pecina a Josef van Genabith. The ML4HMT Workshop on Optimising the Division of Labour in Hybrid Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012. European Language Resources Association. Citováno na straně 133

Amittai E. Axelrod. Factored Language Models for Statistical Machine Translation. Diplomová práce, Institute for Communicating and Collaborative Systems, Division of Informatics, University of Edinburgh, 2006. Citováno na straně 46

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz a Luboš Ureš. The Candide system for machine translation. In *Proceedings of the workshop on Human Language Technology*, HLT '94, str. 157–162, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. Citováno na straně 25, 34, 49

Jeff A. Bilmes a Katrin Kirchhoff. Factored language models and generalized parallel backoff. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, str. 4–6, Morristown, NJ, USA, 2003. Association for Computational Linguistics. Citováno na straně 46

Alexandra Birch, Miles Osborne a Philipp Koehn. CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, str. 9–16, Praha, červen 2007. Association for Computational Linguistics. Citováno na straně 58, 62

Igor Boguslavsky, Leonid Iomdin a Victor Sizov. Multilinguality in ETAP-3: Reuse of Lexical Resources. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources*, str. 1–8, Geneva, Switzerland, srpen 2004. COLING. Citováno na straně 98

Ondřej Bojar a Jan Hajič. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, str. 143–146, Columbus, Ohio, červen 2008. Association for Computational Linguistics. Citováno na straně 95, 96, 98

Ondřej Bojar a Kamil Kos. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, str. 60–66, Uppsala, Sweden, červenec 2010. Association for Computational Linguistics. Citováno na straně 111

Ondřej Bojar a Magdalena Prokopová. Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, str. 1236–1239. ELRA, květen 2006. Citováno na straně 37, 38

Ondřej Bojar, Evgeny Matusov a Hermann Ney. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, svazek LNAI 4139, str. 214–224, Turku, Finland, srpen 2006. Springer. Citováno na straně 115

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka a Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, květen 2008. ELRA. Citováno na straně 32

Ondřej Bojar, Pavel Straňák a Daniel Zeman. English-Hindi Translation in 21 Days. In *Proceedings of the 6th International Conference On Natural Language Processing (ICON-2008) NLP Tools Contest*, Pune, India, prosinec 2008. NLP Association of India. Citováno na straně 133

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš a Zdeněk Žabokrtský. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, březen 2009. Association for Computational Linguistics. Citováno na straně 62

Ondřej Bojar, Kamil Kos a David Mareček. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, str. 86–91, Uppsala, Sweden, červenec 2010. Association for Computational Linguistics. Citováno na straně 113, 115

Ondřej Bojar, Miloš Ercegovčević, Martin Popel a Omar Zaidan. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, str. 1–11, Edinburgh, Scotland, červenec 2011. Association for Computational Linguistics. Citováno na straně 106

Ondřej Bojar, Petra Galuščáková a Miroslav Týnovský. Evaluating Quality of Machine Translation from Czech to Slovak. In Markéta Lopatková, editor, *Information Technologies – Applications and Theory*, str. 3–9, září 2011. Citováno na straně 52

Ondřej Bojar, Bushra Jawaid a Amir Kamran. Probes in a Taxonomy of Factored Phrase-Based Models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, červen 2012. Association for Computational Linguistics. Submitted. Citováno na straně 62

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel a Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, str. 3921–3928, Istanbul, Turkey, květen 2012. ELRA, European Language Resources Association. Citováno na straně 30, 31

Ondřej Bojar. Strojový překlad: zamyšlení nad účelností hloubkových jazykových analýz. In *MIS 2006*, str. 3–13, Josefův Důl, Czech Republic, leden 2006. MATFYZPRESS. Citováno na straně 115

Ondřej Bojar. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, str. 232–239, Praha, červen 2007. Association for Computational Linguistics. Citováno na straně 58, 62, 96

Ondřej Bojar. *Exploiting Linguistic Data in Machine Translation*, svazek 3, *Studies in Computational and Theoretical Linguistics*. ÚFAL, Praha, 2009. Citováno na straně 93

Ondřej Bojar. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:63–76, březen 2011. Citováno na straně 108, 118

Joan Bresnan. *Lexical-Functional Syntax*. Blackwell Publishers, Oxford, 2001. Citováno na straně 98

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajič, Robert L. Mercer a Surya Mohanty. But dictionaries are data too. In *Proceedings of the workshop on Human Language Technology*, HLT '93, str. 202–205, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. Citováno na straně 30

Chris Callison-Burch, Miles Osborne a Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In Diana McCarthy a Shuly Wintner, editors, *EACL*. The Association for Computer Linguistics, 2006. Citováno na straně 113

Chris Callison-Burch, Philipp Koehn, Christof Monz a Josh Schroeder. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, březen 2009. Association for Computational Linguistics. Citováno na straně 106

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut a Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, červen 2012. Association for Computational Linguistics. Citováno na straně 109, 112

Daniel M. Cer, Christopher D. Manning a Daniel Jurafsky. The best lexical metric for phrase-based statistical MT system optimization. In *HLT-NAACL*, str. 555–563. The Association for Computational Linguistics, 2010. Citováno na straně 113

Scott Chacon. *Pro Git*. Apress, Berkely, CA, USA, 1st edition, 2009. Citováno na straně 135

Ciprian Chelba a Frederick Jelinek. Structured language modelling. *Computer Speech and Language*, 14:283–332, 2000. Citováno na straně 47

Stanley F. Chen a Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, str. 310–318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. Citováno na straně 44

Stanley F. Chen a Joshua T. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technická zpráva TR-10-98, Computer Science Group, Harvard University, 1998. Citováno na straně 44

Yu Chen, Andreas Eisele, Christian Federmann, Michael Jellinghaus a Silke Theison. Improved Confidence Estimation and Hybrid Architectures for Machine Translation. Project Euromatrix - Deliverable 6.1, Saarland University, 2007. Citováno na straně 129

Joan Chen-Main a Aravind K. Joshi. A dependency perspective on the adequacy of tree local multi-component tree adjoining grammar. *Journal of Logic and Computation*, červen 2012. Citováno na straně 82

Monique Chevalier, Jules Dansereau a Guy Poulin. TAUM-METEO: description du système. Groupe TAUM, Université de Montréal. Montréal, Canada, 1978. Citováno na straně 20

David Chiang a Kevin Knight. An Introduction to Synchronous Grammars. Část tutoriálu prezentovaného na konferenci ACL 2006, `http://www.isi.edu/~chiang/papers/synchtut.pdf`, červen 2006. Citováno na straně 87

David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, str. 263–270, Ann Arbor, Michigan, červen 2005. Association for Computational Linguistics. Citováno na straně 88

David Chiang. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, červen 2007. Citováno na straně 88

David Chiang. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, str. 1443–1452, Uppsala, Sweden, červenec 2010. Association for Computational Linguistics. Citováno na straně 96, 97

Michal Chytil. *Automaty a gramatiky*. Státní nakladatelství technické literatury, 1984. Citováno na straně 13, 77

Michael Collins, Philipp Koehn a Ivona Kučerová. Clause restructuring for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, str. 531–540, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. Citováno na straně 27

H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison a M. Tommasi. Tree automata techniques and applications. Dostupné na adrese `http://www.grappa.univ-lille3.fr/tata`, 2007. Verze z roku 2012. Citováno na straně 78

M.S. Crouse, R.D. Nowak a R.G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *Signal Processing, IEEE Transactions on*, 46(4):886 –902, duben 1998. Citováno na straně 100

František Čermák a Alexandr Rosen. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427, 2012. Citováno na straně 30

Martin Čmejrek a Bowen Zhou. Two methods for extending hierarchical rules from the bilingual chart parsing. In *Coling 2010: Posters*, str. 180–188, Beijing, China, August 2010. Coling 2010 Organizing Committee. Citováno na straně 97

Martin Čmejrek. *Using Dependency Tree Structure for Czech-English Machine Translation*. Disertační práce, ÚFAL, MFF UK, Praha, 2006. Citováno na straně 93

Ralph Debusmann a Marco Kuhlmann. Dependency grammar: Classification and exploration, 2007. Project report (CHORUS, SFB 378). Citováno na straně 75, 76, 82

A. P. Dempster, N. M. Laird a D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. Citováno na straně 35

Bonnie J. Dorr, Rebecca J. Passonneau, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith J. Miller, Teruko Mitamura, Owen Rambow a Advaith Siddharthan. Interlingual annotation of parallel text corpora: A new framework for annotation and evaluation. *Nat. Lang. Eng.*, 16(3):197–243, 2010. Citováno na straně 21

Markus Dreyer a Daniel Marcu. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, str. 162–171, Montréal, Canada, červen 2012. Association for Computational Linguistics. Citováno na straně 110

Loïc Dugast, Jean Senellart a Philipp Koehn. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, str. 220–223, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. Citováno na straně 130

Loïc Dugast, Jean Senellart a Philipp Koehn. Selective addition of corpus-extracted phrasal lexical rules to a rule-based machine translation system. In *Proc. of MT Summit XII*, 2009. Citováno na straně 40

Christopher Dyer, Smaranda Muresan a Philip Resnik. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, str. 1012–1020, Columbus, Ohio, červen 2008. Association for Computational Linguistics. Citováno na straně 132

Jason Eisner. Learning Non-Isomorphic Tree Mappings for Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume*, str. 205–208, Sapporo, červenec 2003. Citováno na straně 79

Marcello Federico, Luisa. Bentivogli, Michael Paul a Sebastian Stüker. Overview of the IWSLT 2011 Evaluation Campaign. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA*, 2011. Citováno na straně 133

Alexander Fraser a Daniel Marcu. Getting the structure right for word alignment: LEAF. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, str. 51–60, Praha, červen 2007. Association for Computational Linguistics. Citováno na straně 38

Haim Gaifman. Dependency Systems and Phrase-Structure Systems. *Information and Control*, 8(3):304–337, 1965. Citováno na straně 74

William A. Gale a Kenneth W. Church. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102, 1993. Citováno na straně 34

Petra Galuščáková a Ondřej Bojar. Improving SMT by Using Parallel Data of a Closely Related Language. In *Human Language Technologies – The Baltic Perspective – Proceedings of the Fifth International Conference Baltic HLT 2012*, svazek 247, *Frontiers in AI and Applications*, str. 58–65, Amsterdam, Netherlands, říjen 2012. IOS Press. Citováno na straně 130

Qin Gao a Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, str. 49–57, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. Citováno na straně 34

Gerald Gazdar, Ewan H. Klein, Geoffrey K. Pullum a Ivan A. Sag. *Generalized Phrase Structure Grammar*. Oxford: Blackwell, and Cambridge, MA: Harvard University Press, 1985. Citováno na straně 74

Sharon Goldwater a David McClosky. Improving statistical MT through morphological analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, str. 676–683, Morristown, NJ, USA, 2005. Association for Computational Linguistics. Citováno na straně 135

Barry Haddow, Abhishek Arun a Philipp Koehn. Samplerank training for phrase-based machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, str. 261–271, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. Citováno na straně 129

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue a Yi Zhang. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In Jan Hajič, editor, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, str. 1–18, Boulder, CO, USA, 2009. Association for Computational Linguistics. Citováno na straně 69

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová a Zdeněk Žabokrtský. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, str. 3153–3160, Istanbul, Turkey, květen 2012. ELRA, European Language Resources Association. Citováno na straně 69

Jan Hajič. *Disambiguation of Rich Inflection - Computational Morphology of Czech*, svazek I. Univerzita Karlova v Praze, Nakladatelství Karolinum, 2001. Citováno na straně 61

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský a Magda Ševčíková Razímová. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4, 2006. Citováno na straně 69

Michal Havlíček. Citlivost metrik automatického vyhodnocování překladu. Student project at POPJ2 (Počítače a přirozený jazyk) seminar at FJFI, Czech Technical University, 2007. Citováno na straně 114

Hermann Helbig. *Knowledge Representation and the Semantics of Natural Language*. Springer, 2006. Citováno na straně 21

Zdeněk Hlavsa *et al. Pravidla českého pravopisu*. Praha: Academia, 2. (s Dodatkem Ministerstva školství, mládeže a tělovýchovy ČR) edition, 2005. Citováno na straně 29

Tomáš Holan, Vladislav Kuboň, Karel Oliva a Martin Plátek. Two Useful Measures of Word Order Complexity. In A. Polguere a S. Kahane, editors, *Proceedings of the Coling '98 Workshop: Processing of Dependency-Based Grammars*, Montreal, 1998. University of Montreal. Citováno na straně 73, 75

Tomáš Holan, Vladislav Kuboň, Karel Oliva a Martin Plátek. On Complexity of Word Order. *Special Issue on Dependency Grammar of the journal TAL (Traitement Automatique des Langues)*, 41(1):273–300, 2000. Citováno na straně 76

Tomáš Holan. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS, leden 2003. Citováno na straně 75

Mark Hopkins a Jonathan May. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, str. 1352–1362, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. Citováno na straně 129

Jan Hric, Jan Hajič a Vladislav Kuboň. Machine Translation of Very Close Languages. *Proc. of the 6th Applied Natural Language Processing Conference*, str. 7–12, 2000. Citováno na straně 52, 130

Liang Huang, Kevin Knight a Aravind Joshi. Statistical Syntax-Directed Translation with Extended Domain of Locality. In *Proc. of 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, MA, 2006. Citováno na straně 96

Zhongqiang Huang, Martin Čmejrek a Bowen Zhou. Soft syntactic constraints for hierarchical phrase-based translation using latent syntactic distributions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, str. 138–147, Cambridge, MA, October 2010. Association for Computational Linguistics. Citováno na straně 97

John Hutchins. Machine Translation: History. In Keith Brown, editor, *Encyclopedia of Language & Linguistics, Second Edition*, svazek 7, str. 375–383. Oxford: Elsevier, 2006. Citováno na straně 5

Petr Jančar. *Teoretická informatika. Učební text.* VŠB-TU, Ostrava, 2007. Citováno na straně 13, 77

Frederick Jelinek a Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In E. Gelsema a L. Kanal, editors, *Pattern recognition in practice*, str. 381–397, Amsterdam, květen 1980. North-Holland. Citováno na straně 44

Frederick Jelinek. A stack algorithm for faster sequential decoding of transmitted information. Technická zpráva RC2441, IBM Research Center, Yorktown Heights, NY, 1969. Citováno na straně 63

Aravind Joshi a Owen Rambow. A Formalism for Dependency Grammar Based on Tree Adjoining Grammar. In *First International Conference on Meaning Text Theory*, str. 207–216, Paris, France, 2003. Citováno na straně 82

Aravind K. Joshi, Leon S. Levy a Masako Takahashi. Tree adjunct grammars. *J. Comput. Syst. Sci.*, 10(1):136–163, 1975. Citováno na straně 81

Aravind K. Joshi, K. Vijay Shanker a David Weir. The Convergence of Mildly Context-Sensitive Grammar Formalisms. Technická zpráva MS-CIS-90-01, University of Pennsylvania Department of Computer and Information Science, 1990. Citováno na straně 82

Daniel Jurafsky a James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2st edition, 2008. Citováno na straně 44

Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *ASSP*, 35(3):400–401, březen 1987. Citováno na straně 44

Katrin Kirchhoff a Mei Yang. Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, str. 125–128, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. Citováno na straně 46

Katrin Kirchhoff, Daniel Capurro a Anne Turner. Evaluating User Preferences in Machine Translation Using Conjoint Analysis. In *Proceedings of EAMT 2012*, str. 119–126, Trento, Italy, 2012. Citováno na straně 11, 118

Zdenek Kirschner a Alexandr Rosen. Apac - an experiment in machine translation. *Machine Translation*, 4(3):177–193, 1989. Citováno na straně 27

Dan Klein a Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, str. 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. Citováno na straně 136

Hana Klempová, Michal Novák, Peter Fabian, Jan Ehrenberger a Ondřej Bojar. Získávání paralelních textů z webu. In *ITAT 2009 Information Technologies – Applications and Theory*, září 2009. Citováno na straně 33

Kevin Knight. Decoding complexity in word-replacement translation models. *Comput. Linguist.*, 25(4):607–615, 1999. Citováno na straně 11, 63

Jan Kocek, Marie Kopřivová a Karel Kučera, editors. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Praha, 2000. Citováno na straně 30

Philipp Koehn a Barry Haddow. Interpolated backoff for factored translation models. In *Proceedings of AMTA*, 2012. Citováno na straně 62

Philipp Koehn a Hieu Hoang. Factored Translation Models. In *Proc. of EMNLP*, 2007. Citováno na straně 58

Philipp Koehn a Christof Monz. Shared task: statistical machine translation between european languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, str. 119–124, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. Citováno na straně 133

Philipp Koehn, Franz Joseph Och a Daniel Marcu. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*, 2003. Citováno na straně 134

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin a Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, str. 177–180, Praha, červen 2007. Association for Computational Linguistics. Citováno na straně 134

Philipp Koehn. *Noun Phrase Translation*. Disertační práce, University of Southern California, 2003. Citováno na straně 55

Philipp Koehn. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In Robert E. Frederking a Kathryn Taylor, editors, *AMTA*, svazek 3265, *Lecture Notes in Computer Science*, str. 115–124. Springer, 2004. Citováno na straně 49

Philipp Koehn. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, Barcelona, Spain, 2004. Citováno na straně 117

Alena Koubková a Václav Koubek. *Datové struktury I*. 2011. Citováno na straně 22

Vladislav Kuboň. A Robust Parser for Czech. Technická zpráva TR-1999-06, ÚFAL/CKL, Praha, 2001. Citováno na straně 41

Marco Kuhlmann a Mathias Möhl. Mildly context-sensitive dependency languages. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, str. 160–167, Praha, červen 2007. Association for Computational Linguistics. Citováno na straně 76

Andrew Lampert. Interlingua in Machine Translation. `http://sgi.nu/nlp/content/pdf/InterlinguaInMachineTranslation.pdf`, září 2001. Citováno na straně 21

Vladimir Iosifovich Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, únor 1966. Citováno na straně 34

Markéta Lopatková, Martin Plátek a Vladislav Kuboň. Modeling syntax of Free Word-Order Languages: Dependency Analysis By Reduction. In Václav Matoušek, Pavel Mautner a Tomáš Pavelka, editors, *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings*, svazek LNAI 3658, str. 140–147. Springer Verlag, září 2005. Citováno na straně 70

Markéta Lopatková, Zdeněk Žabokrtský a Václava Kettnerová. *Valenční slovník českých sloves*. Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha, 2008. In cooperation with Karolína Skwarska, Eduard Bejček, Klára Hrstková, Michaela Nová and Miroslav Tichý. Citováno na straně 29, 85

Adam Lopez. Statistical Machine Translation. *ACM Computing Surveys*, 40(3), září 2008. Citováno na straně 25

Adam Lopez. Tera-scale translation models via pattern matching. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, str. 505–512, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. Citováno na straně 90

Adam Lopez. Translation as weighted deduction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, str. 532–540, Athens, Greece, březen 2009. Association for Computational Linguistics. Citováno na straně 64, 65

Matouš Macháček. Metriky pro optimalizaci modelů strojového překladu. Bakalářská práce, Matematicko-fyzikální fakulta, Univerzita Karlova v Praze, 2012. Citováno na straně 113

David Mareček, Zdeněk Žabokrtský a Václav Novák. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proceedings of EAMT 2008*, Hamburg, Germany, 2008. Citováno na straně 39

Michal Marek, Pavel Pecina a Miroslav Spousta. Web page cleaning with conditional random fields. In Cédrick Fairon, Hubert Naets, Adam Kilgarriff a Gilles-Maurice de Schryver, editors, *Proceedings of the 3rd Web As a Corpus Workshop, Incorporating CLEANEVAL*, str. 155–162, Louvain-la-Neuve, Belgium, 2007. UCL Pressess Universitaires de Louvain. Citováno na straně 31

Jiří Maršík. Rychlý a trénovatelný tokenizér pro přirozené jazyky. Bakalářská práce, Matematicko-fyzikální fakulta, Univerzita Karlova v Praze, 2011. Citováno na straně 18

Ryan McDonald, Fernando Pereira, Kiril Ribarov a Jan Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, říjen 2005. Citováno na straně 75, 87

Kurt Mehlhorn. *Effiziente Algorithmen*. Stuttgart: Teubner, 1977. Vydáno po úpravách v anglickém překladu Data Structures and Algorithms, Springer-Verlag, 1984. Citováno na straně 22

Igor A. Mel'čuk. *Dependency Syntax - Theory and Practice*. Albany: State University of New York Press, 1988. Citováno na straně 83, 98

Christof Monz. Statistical machine translation with local language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, str. 869–879, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. Citováno na straně 47

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel a Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, str. 915–932, 2007. Citováno na straně 69

Joakim Nivre. Dependency Grammar and Dependency Parsing. Technická zpráva MSI report 05133, Växjö University: School of Mathematics and Systems Engineering, 2005. Citováno na straně 71

Václav Novák, Sven Hartrumpf a Keith Hall. Large-scale semantic networks: annotation and evaluation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, str. 37–45, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. Citováno na straně 21

Franz Josef Och a Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003. Citováno na straně 34

Franz Joseph Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Disertační práce, RWTH Aachen University, 2002. Citováno na straně 54, 125

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, červenec 2003. Citováno na straně 127

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén a Dan Flickinger. Towards Hybrid Quality-Oriented Machine Translation. On Linguistics and Probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, str. 144–153, Skövde, Sweden, 2007. Citováno na straně 98

Kishore Papineni, Salim Roukos, Todd Ward a Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, str. 311–318, Philadelphia, Pennsylvania, 2002. Citováno na straně 113, 114

Dana Pardubská, Martin Plátek a Friedrich Otto. Parallel communicating grammar systems with regular control and skeleton preserving frr automata. *Theoretical Computer Science*, 412(4–5):458 – 477, 2011. Citováno na straně 71

Adam Pauls a Dan Klein. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, str. 959–968, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. Citováno na straně 48

Michael Pilato. *Version Control With Subversion*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2004. Citováno na straně 135

Carl J. Pollard a Ivan A. Sag. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994. Citováno na straně 74, 83, 98

Martin Popel a David Mareček. Perplexity of n-gram and dependency language models. In Petr Sojka, Aleš Horák, Ivan Kopeček a Karel Pala, editors, *Text, Speech and Dialogue. 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings*, svazek 6231, *Lecture Notes in Computer Science*, str. 173–180, Berlin / Heidelberg, 2010. Springer. Citováno na straně 47

Martin Popel a Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson a Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, svazek 6233, *Lecture Notes in Computer Science*, str. 293–304, Berlin / Heidelberg, 2010. Iceland Centre for Language Technology (ICLT), Springer. Citováno na straně 98

Martin Popel. Ways to Improve the Quality of English-Czech Machine Translation. Diplomová práce, Univerzita Karlova v Praze, ÚFAL, Praha, 2009. In Czech. Citováno na straně 100

Adam Przepiórkowski a Anna Kupść. HPSG for Slavicists. *Glossos*, 8:1–68, 2006. Citováno na straně 83

Chris Quirk, Arul Menezes a Colin Cherry. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, str. 271–279. Association for Computational Linguistics, 2005. Citováno na straně 82, 96

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh a Pushpak Bhattacharyya. Case markers and morphology: addressing the crux of the fluency problem in english-hindi smt. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, str. 800–808, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. Citováno na straně 62

Stephen D. Richardson, William B. Dolan, Arul Menezes a Monica Corston-Oliver. Overcoming the Customization Bottleneck Using Example-Based MT. In *Proceedings of the workshop on Data-driven methods in machine translation*, str. 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics. Citováno na straně 98

Rudolf Rosa, Ondřej Dušek, David Mareček a Martin Popel. Using parallel features in parsing of machine-translated sentences for correction of grammatical errors. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-6 '12, str. 39–48, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. Citováno na straně 130

Rudolf Rosa, David Mareček a Ondřej Dušek. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, str. 362–368, Montréal, Canada, červen 2012. Association for Computational Linguistics. Citováno na straně 130

Alexandr Rosen. In Search of Best Method for Sentence Alignment in Parallel Texts. In R. Garabík, editor, *Computer Treatment of Slavic and East European Languages*, str. 174–185. Veda, Bratislava, 2005. Citováno na straně 34

Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz a Bonnie J. Dorr. Combining Outputs from Multiple Machine Translation Systems. In *HLT-NAACL*, str. 228–235, 2007. Citováno na straně 130

Lane Schwartz. Multi-source translation methods. In *Proc. AMTA*, říjen 2008. Citováno na straně 111

Holger Schwenk. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, str. 182–189, 2008. Citováno na straně 34

Petr Sgall, Eva Hajičová a Petr Piťha. *Učíme stroje česky*. Pyramida. Panorama, Praha, 1982. Citováno na straně 3, 77

Petr Sgall, Eva Hajičová a Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Praha/Dordrecht, 1986. Citováno na straně 69, 70

Petr Sgall. *Generativní popis jazyka a česká deklinace*. Academia, Praha, 1967. Citováno na straně 69

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla a John Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, str. 223–231, 2006. Citováno na straně 108

Miroslav Spousta, Michal Marek a Pavel Pecina. Victor: the web-page cleaning tool. In Stefan Evert, Adam Kilgarriff a Serge Sharoff, editors, *Proceedings of the 4th Web as Corpus Workshop*, str. 12–17, Marrakech, Morocco, 2008. ACL SIGWAC. Citováno na straně 31

Mark Steedman. *The Syntactic Process*. MIT Press, 2000. Citováno na straně 74

Michael Subotin. An exponential translation model for target language morphology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, str. 230–238, Portland, Oregon, USA, červen 2011. Association for Computational Linguistics. Citováno na straně 135

Bernard Vauquois. La traduction automatique à Grenoble. Document de linguistique quantitative 24. Dunod, Paris., 1975. Citováno na straně 20

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och a Juri Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, str. 1363–1372, Edinburgh, Scotland, UK., červenec 2011. Association for Computational Linguistics. Citováno na straně 14

David Vilar, Jia Xu, Luis Fernando D'Haro a Hermann Ney. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, str. 697–702, Genoa, Italy, květen 2006. Citováno na straně 108

Philip Williams a Philipp Koehn. Agreement constraints for statistical machine translation into german. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, str. 217–226, Edinburgh, Scotland, červenec 2011. Association for Computational Linguistics. Citováno na straně 93

Al-Onaizan Yaser, Jan Cuřín, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith a David Yarowsky. Statistical Machine Translation. Final Report. Technická zpráva, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1999. Dostupné na adrese `http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps`. Citováno na straně 134

Zdeněk Žabokrtský a Martin Popel. Hidden markov tree model in dependency-based machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, str. 145–148, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. Citováno na straně 100

Zdeněk Žabokrtský, Jan Ptáček a Petr Pajas. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, str. 167–170, Columbus, Ohio, USA, 2008. Citováno na straně 98

Andreas Zollmann a Ashish Venugopal. Syntax Augmented Machine Translation via Chart Parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, str. 138–141, New York City, červen 2006. Association for Computational Linguistics. Citováno na straně 97

Seznam obrázků

Seznam tabulek

Rejstřík

O

P

R

Ř

S

Š

T

U

V

W

Z

Slovníček anglických termínů

A

a-layer = a-rovina
adequacy = věrnost
adjunct = volné doplnění
adjunction = adjunkce
alignment
  word ˜ = zarovnání, slovní
alternative decoding paths = alternativní dekódovací cesty
ambiguity = víceznačnost
  spurious = víceznačnost, nadbytečná
analysis by reduction = analýza, redukční
analysis-transfer-synthesis = analýza, transfer a syntéza
anaphora = anafora
annotation = anotace
asynchronous factors = asynchronní faktory
atomicity = nedělitelnost
automatic post-editing = posteditace, automatická

B

back-off = back-off
balance = vyváženost
beam search = beam search
BLEU = BLEU
bootstrap resampling = bootstrapping
brevity penalty = pokuta, za stručnost

C

chain rule = řetězové pravidlo
co-reference = koreference
complete hypothesis = hypotéza, úplná
compositionality = kompozicionalita
computational linguistics = počítačová lingvistika
computational science = počítačová věda (obecně)
confusion network = konfuzní síť
constituency tree = strom, složkový
constituent = složka
context-free grammar = gramatika, bezkontextová
controlled language = řízený jazyk
coordination = koordinace
corpus = korpus
  comparable = korpus, porovnatelný
  parallel = korpus, paralelní
crossing brackets = překřížené závorky

D

data sparsity = řídkost dat
declarative description = deklarativní popis
decoder = dekodér
deep analysis = analýza, hloubková
depend = záviset
dependency tree = strom, závislostní
derivation = derivace
derivation history = derivační historie
development set = vývojová sada
direct translation = přímý překlad
discriminative model = model, diskriminativní
distortion limit = slovosledný limit
distortion penalty = pokuta, za změnu slovosledu

E

emission probability = emisní pravděpodobnost
error flagging = značkování chyb
evaluation of machine translation = vyhodnocování strojového překladu
executable paper = spustitelný článek
Expectation-Maximization = Expectation-Maximization

F

factor = faktor
factored language model = model, jazykový, faktorový
feature = rys
feature function = rys
feature structure = sestava rysů
fluency = plynulost
formeme = formém
fractional count = relativní počet
frontier node = hraniční uzel
full reproducibility = plná reprodukovatelnost
Functional Generative Description = Funkční generativní popis
functor = funktor

G

gap = díra
gap degree = počet děr
generalization = zobecnění
generation step = generovací krok
generation table = generovací tabulka
grammar = gramatika
  synchronous ˜ = gramatika, synchronní
  tree ˜ = gramatika, stromová
  tree substitution ˜ = gramatika, stromová substituční
  unification ˜ = gramatika, unifikační
grammaticality = gramatičnost

H

head = hlava
head word = hlava
hidden Markov model = model, skrytý markovovský
hidden Markov tree model = model, skrytý stromový markovovský
hierarchical phrase-based translation = hierarchický frázový překlad
homonymy = homonymie
hypergraph = hypergraf
hypothesis = hypotéza
hypothesis ranking = uspořádávání hypotéz

I

idiomatic expression = idiomatické spojení
independence assumption = předpoklad nezávislosti
information structure = aktuální členění věty
interlingua = interlingva
internal node = vnitřní uzel
interpolation = vyhlazování, interpolované
item in parsing = položka v parsingu

L

language modelling = modelování, jazykové
latent variable = latentní proměnná
lattice
  of hypotheses = svaz hypotéz
  of words = svaz slov
layer
  analytical ˜ = rovina, analytická

tectogrammatical ˜ = rovina, tektogramatická
lemma = lema
lexical probability = lexikální pravděpodobnost
local feature = rys, lokální
log-linear model = model, log-lineární

M

machine translation = strojový překlad
mapping step = dekódovací krok
Markov(ian) property = markovovská vlastnost
maximum likelihood estimate = maximálně věrohodný odhad
mildly context-sensitive grammar = gramatika, mírně kontextová
minimum error-rate training = MERT
minimum translation unit = nejmenší překladová jednotka
MLE = maximálně věrohodný odhad
modelling = modelování
modelling error = chyba modelu
model
distortion ˜ = model, slovosledu
factored translation ˜ = model, faktorový překladový
reordering ˜model = model, slovosledu
sequence ˜ = model, sekvenční
MT evaluation metric = metrika kvality strojového překladu

N

n-best list = n-best list
n-gram = n-gram
n-gram language model = model, jazykový, n-gramový
named entity = pojmenovaná entita
negation = negace
noisy channel = šumový kanál
non-local feature = rys, nelokální
non-projectivity = neprojektivita
non-terminal = neterminál
NP-completeness = NP-úplnost

O

OOV rate = podíl OOV
OOV words = slova mimo slovník
ordered dependency tree = uspořádaný závislostní strom
out-of-vocabulary words = slova mimo slovník
overfitting = přetrénování

P

parallel treebank = treebank, paralelní
parameterization = parametrizace
parser = parser
parsing = parsing
phrasal verb = frázové sloveso
phrase = fráze
phrase extraction = získávání frází
phrase penalty = pokuta, za frázi
phrase probability = pravděpodobnost překladu fráze
phrase table = tabulka frází
phrase-based translation = frázový překlad
phrases consistent with word alignment = fráze, v souladu se slovním zarovnáním
pigeon-hole principle = Dirichletův princip
pivot language = pivotní jazyk
post-editing = posteditace
post-processing = závěrečné úpravy
Prague Dependency Treebank = Pražský závislostní korpus
pre-processing = předzpracování
presupposition = presupozice

projection = projekce
projectivity = projektivita
pronoun = zájmeno
proper noun = jméno, vlastní
prune = prořezávání
public API = veřejné rozhraní
public interface = veřejné rozhraní
pumping lemma = lemma o vkládání

R

ranking = uspořádávání hypotéz
reference translation = referenční překlad
reordering penalty = pokuta, za změnu slovosledu
robustness = robustnost
rule-based = pravidlový

S

sausage = párek
score = skóre
search = prohledávání
search error = chyba při prohledávání
segmentation = segmentace
sentence comprehension = test srozumitelnosti vět
separable prefix = odlučitelná předpona
smoothing = vyhlazování
space of partial hypotheses = stavový prostor částečných hypotéz
span = projekce
stack = zásobník
state = stav
state splitting = dělení stavů
statistical = statistický
statistical decision theory = statistická teorie rozhodování
STSG property = vlastnost STSG
substitution = substituce
syntactic analysis = analýza, syntaktická
syntactic structure = syntaktická struktura
system combination = kombinace systémů

T

t-layer = t-rovina
tagging = morfologické značkování
target word form = cílový slovní tvar
targeted reference = cílená reference
TectoMT = TectoMT
terminal = terminál
test set = testovací sada
text domain = doména textu
token = token
tokenization = tokenizace
topic-focus articulation = aktuální členění věty
trace = stopa
transfer = převod
transition probability = přechodová pravděpodobnost
translation equivalence model = model, překladové ekvivalence
translation model = model, překladový
translation option = možnost překladu
translation step = překladový krok
tree-adjoining grammar = gramatika, stromová adjunkční
treebank = treebank
tree
- *analytical ˜* = strom, analytický
- *deep syntactic ˜* = strom, hloubkový
- *surface syntactic ˜* = strom, povrchový
- *tectogrammatical ˜* = strom, tektogramatický

ttable = tabulka frází

U

unification = unifikace

unordered dependency tree = strom, neuspořádaný závislostní

V

valency = valence
valency frame = valenční rámec
Viterbi approximation = Viterbiho aproximace

W

well-nestedness = dobrá zahnízděnost
word lattice = svaz slov
word order = slovosled
word penalty = pokuta, za slovo