PLIN009 – Strojový překlad

Základní informace
Úvod do překladu
Úvod do strojového překladu
Nástin vývoje strojového překladu

Vít Baisa

jaro 2013

7. března 2013

Informace k výuce

- přednáška: čtvrtek 9.10–10.55 v G32
- konzultace: B203 (FI, 2. patro budovy B)
  úterý 9.00–10.30, čtvrtek 9.30–10.30
- po předchozí domluvě, možné i jindy
- email: `xbaisa@fi.muni.cz`
- `nlp.fi.muni.cz/~xbaisa/plin019`
- studijní materiály pouze tyto slajdy a stránky předmětu
- sledujte interaktivní osnovu v IS

Podmínky ukončení

- povinná cca 5minutová prezentace:
  - ~~zajímavý / zásadní článek z oblasti SP~~
  - systém SP – popis, ukázky a srovnání
- závěrečná písemná práce
  - ukázky otázek budou ukázány v průběhu semestru

Prezentace – požadavky, doporučení

- maximálně 5 minut
- následná diskuze v rámci IS (diskuzní fórum předmětu)
- 3–5 slajdů (PDF), nb a projektor k dispozici
- slajdy vložíte nejpozději 14 dní po prezentaci vloženy do IS
- prezentace nebude hodnocena
- na začátku každé hodiny 1–2 prezentace
- počínaje 3. týdnem výuky

Prezentace – struktura, obsah

Prezentace článku, studie

- bibliografická identifikace
- prezentace obsahu publikace
- vytažení nejdůležitějších „myšlenek“

INOVA.CZ

- Strukturální fond EU
- Evropský sociální fond (ESF)
- Operační program pro vzdělávání a konkurenceschopnost (OPVK)
- *Mezi bohemistikou a informatikou. Inovace vysokoškolské výuky češtiny v kontextu počítačového zpracování přirozeného jazyka* (INOVA.CZ)
- `www.projekt-inova.cz`
- informace o aktivitách v rámci projektu

Literatura

- John Hutchins – *Machine translation: past, present, future*
- John Hutchins – *An introduction to machine translation*
- Philipp Koehn – *Statistical Machine Translation*
- Sergei Nirenburg et al. – *Readings in Machine Translation*
- Jiří Levý – *Umění překladu*
- Jiří Levý – *České theorie překladu*
- další literatura a zdroje viz stránky předmětu

Překlad I

Překlad

Překlad je převod textu ze zdrojového jazyka do jazyka cílového.

Tlumočení

Tlumočení je ústní překlad mluveného jazyka.

Překlad I

Překlad

Překlad je převod textu ze zdrojového jazyka do jazyka cílového.

Tlumočení

Tlumočení je ústní překlad mluveného jazyka.

Překlad je jako žena: buď věrný, nebo hezký.

Překlad II

- odborný překlad $\times$ literární překlad
- přesná reprodukce $\times$ volná převodová parafráze

Maimonidés, 12. stol.

Pro překlad slova je rozhodující kontext.

Werner Winter

Každé slovo je element vytržený z celkového jazykového systému a jeho vztahy k jiným segmentům systému jsou v jednotlivých jazycích rozdílné.

Každý význam je element z celého systému segmentů, v něž mluvčí rozčleňuje skutečnost. V jazyce Mohave: otec ženy $\neq$ otec muže

Jaké vlastnosti zdroje mají být zachovány? – *J. Levý*

	odborný styl	publicistická a rétorická próza	umělecká próza a drama	volný verš	pravidelný verš	hudební text (libreto)	dabing
denotativní význam	i	i	i	i	i	i-v	i-v
konotativní význam	v	i-v	i	i	i	i	i
stylistické zařazení slova	i-v	i	i	i	i	i	i
větná stavba	v	i-v	i	i	i	i	i
opakování (rytmus, rým)	v	v	v	i-v	i	i	i-v
délka a výška samohlásek	v	v	v	i-v	i-v	i	i
způsob artikulace	v	v	v	i-v	i-v	i-v	i

Překlad (Levý)

- musí reprodukovat
  - slova originálu
  - ideje originálu
- se má dát číst jako originál
- má být čten jako překlad
- by měl
  - obrážet styl originálu
  - ukazovat styl překladatelův
  - být čten jako text náležející do doby
    - originálu
    - překladatelovy
- může k originálu něco přidávat nebo z něho vynechávat
- by neměl nikdy k originálu nic přidávat a vynechávat

Translatologie

- vědní obor zabývající se překladem textů mezi jazyky a sémiotickými systémy
- otázky přesnosti (věrnosti), přeložitelnosti
- překlad mezi kulturními oblastmi, obdobími
- větev deskriptivní (kritika a dějiny) $\times$ aplikovaná (praxe)
- 60.–70. léta vznik, lingvistická orientace
- 80. léta přiblížení literární teorii
- 90. léta obrat k překladateli jako jedinci

Překladatel

Co by měl překladatel znát (Levý):

- zdrojový jazyk
- cílový jazyk
- věcný obsah textu: dobové reálie, obor (u odborného překladu)

Levý o uměleckém překladu

Překlad má působit jako umělecké dílo.

Strojový překlad a umělecké překlad – Levý

Strojovému překladu jde nutně o atomizování věty na nejjednodušší srovnatelné jednotky; uměleckému naopak o převádění co nejvyšších celků.

Typy překladu podle Romana Jakobsona

- **mezijazykový** – převod mezi různými jazyky
- **vnitrojazykový** – převod v rámci jednoho jazyka, např. do jiného nářečí, do spisovné podoby apod.
- **meziznakový** – převod mezi různými znakovými systémy

Otázky překladu

- Je vůbec přesný překlad mezi jazyky možný?
- Jak se pozná, že w_1 je překladový ekvivalent slova w_2?
- anglické typy větru: airstream, breeze, crosswind, dust devil, easterly, gale, gust, headwind, jet stream, mistral, monsoon, prevailing wind, sandstorm, sea breeze, sirocco, southwester, tailwind, tornado, trade wind, turbulence, twister, typhoon, whirlwind, wind, windstorm, zephyr
- jak přeložit slova jako *alkáč*, *večerníček*, *telka*, *čoklbuřt*, *knížečka*, *ČSSD* . . . ?
- kód navajo – jazyk jako šifra

Jazykový relativismus I

- vlastnosti jazyka podstatně ovlivňují naše vnímání světa
- vlastnosti různých jazyků se výrazně liší
- jejich mluvčí tudíž žijí v různých, nepřevoditelných světech

Ludwig Wittgenstein

„Hranice mého jazyka znamenají hranice mého světa.“

Fritz Mauthner

Kdyby byl Aristoteles z kmene Dakotů, jeho logika by nabyla zcela odlišné podoby.

Jazykový relativismus II – dualismus

- **teorie matrice** (mould theories): jazyk a myšlení jsou totožné, myslíme jazykem
- **teorie pláště** (cloak theories): jazyk je na povrchu, za ním je složitá spleť myšlenek

Kam patří *jazykový relativismus*?

Sapir-Whorfova hypotéza

- historicky významná teorie psycholingvistiky
- 30. léta 20. století, Edward Sapir, původ v jazykovém relativismu
- srovnání pojmů v indiánských a indoevropských jazycích
- teorie rozpracována Benjaminem Lee Whorfem
- později kritizována, testovatelná podoba hypotézy (pojmy pro barvy) prokázala spíše opak

Strojový překlad I – definice

Strojový překlad

Obor počítačové lingvistiky zabývající se návrhem, implementací a aplikací automatických systémů (programů) pro překlad textů s minimálním zásahem člověka.

Např. používání elektronických slovníků při překladu nepatří do strojového překladu.

Strojový překlad II – předmět zájmu

Uvažujeme pouze odborné texty:

- webové stránky
- technické manuály
- vědecké dokumenty
- prospekty, katalogy
- právnické texty
- obecně texty z omezených domén

Nuance na různých jazykových vrstvách v umělecké literatuře jsou mimo schopnosti současných nástrojů NLP.

Strojový překlad III

Ve skutečnosti je výstup z SP vždy revidován. Mluví se o *před-překladu* resp. o *post-editaci*.

Ta je někdy nutná i u člověka, ovšem systémy SP dělají zcela rozdílné chyby.

Pro člověka jsou typické chyby:

- špatné předložky (*I am in school*)
- chybějící členy (*I saw man*)
- špatný čas (*Uviděl jsem – I was seeing*), . . .

Pro počítač jsou typické zejména chyby významové: *Kiss me, honey.*

Metody zlepšení kvality strojového překladu

- omezení vstupu na:
  - podjazyk (oznamovací věty)
  - doménu (informatika)
  - typ dokumentu (patentové dokumenty)
- pre-processing textu (např. ruční syntaktická analýza)

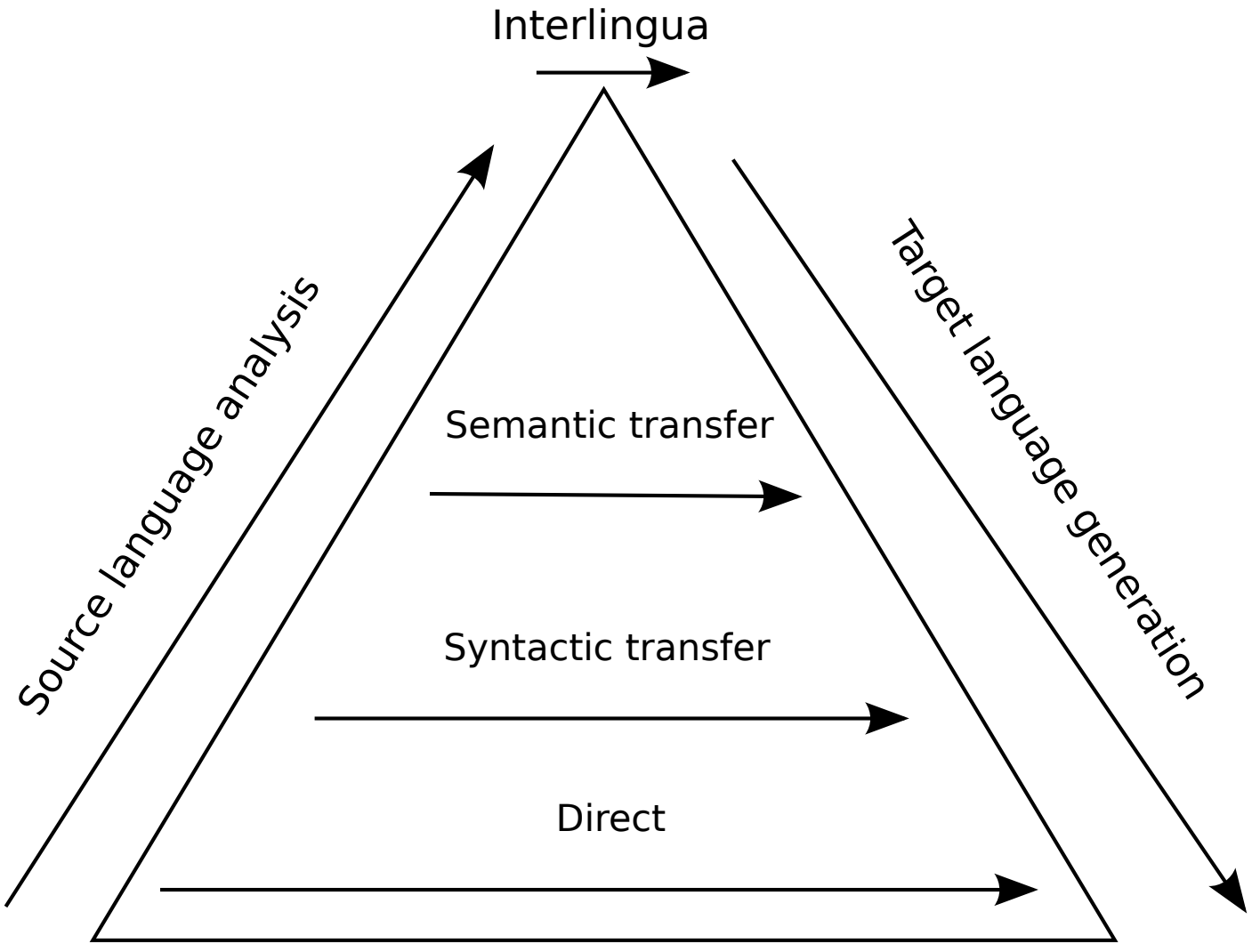
Základní pojmy

- **přesnost** (accuracy, precision)
- **srozumitelnost** (intelligibility)
- **plynulost** (fluency)
- **zdrojový** (výchozí) jazyk (source language, SL)
- **cílový jazyk** (target language, TL)
- **korpus** (corpus, corpora)
- **víceznačnost** (ambiguity)

Klasifikace podle přístupu (approach)

- pravidlový (znalostní) strojový překlad
  rule-based, knowledge-based – RBMT, KBMT
  - transferový
  - interlingua
- statistický strojový překlad
  statistical machine translation – SMT
- hybridní strojový překlad
  hybrid machine translation – HMT, HyTran

Vauquoisův trojúhelník



Klasifikace podle interakce s uživatelem

- (ruční překlad)
- ruční překlad s pomocí počítače
  machine-aided human translation – MAHT
- automatický překlad s interagujícím překladatelem
  human-aided machine translation – HAMT
- plně automatický překlad
  fully automated high-quality (M)T – FAHQT

HAMT a MAHT někdy souhrnně označovány jako CAT – computer-aided translation.

Klasifikace podle směru a četnosti překladu

Podle četnosti:

- dvojjazyčné systémy (bilingual)
- vícejazyčné systémy (multilingual)

Podle směru:

- jednosměrné (unidirectional)
- obousměrné (bidirectional)

Systémy strojového překladu

Apertium (RBMT, open-source), **Babelfish** (Yahoo), **Caitra** (CAT systém), **ČESILKO** (česko-slovenský překlad), **EuroTra** (ambiciózní projekt EK), **Google Translate**, **Logos** (OpenLogos, jeden z nejstarších MT systémů), **METEO** (překlad předpovědí, angličtina, francouzština), **Moses** (open-source MT systém), **Pangloss** (example-based MT), **Rosetta** (obsahuje logickou analýzu), **Systran** (jeden z nejstarších MT systémů), **Trados** (překladová paměť, CAT systém), **Verbmobil** (překlad řeč↔řeč mezi němčinou, angličtinou a japonštinou), . . .

Konference, workshopy

- ACL – Annual meetings of the Association for Computational Linguistics
- NIST – National Institute of Standards and Technology
- Translating and the Computer (Londýn)
- RANLP – Recent Advances in Natural Language Processing
- MT Summit
- The Xth Conference of the Association for Machine Translation in the Americas
- LREC – Language Resources and Evaluation Conferences
- `www.wikicfp.com`

(Elektronické) informační zdroje

- odkazy na stránkách předmětu
- MT Archive
- `www.statmt.org`
- ACL Anthology
- Translation Journal

Instituce

- IAMT – International Association for Machine Translation:
  - EAMT – European Association for Machine Translation
  - AMTA – The Association for MT in the Americas
  - AAMT – The Asian-Pacific Association for MT
- META-NET – sdružuje evropská MT pracoviště
- British Computer Society Natural Language Translation Group
- UK MFF ÚFAL
- Obec překladatelů (překlady krásné literatury)
- Jednota tlumočníků a překladatelů
- Ústav translatologie, FF UK

Motivace pro strojový překlad po 2. světové válce

- období informačního boomu
  - 1922 – pravidelné rozhlasové vysílání BBC
  - 1923 – rozhlasové vysílání v ČR
  - 1936 – pravidelné televizní vysílání BBC
  - 1953 – začíná TV vysílání v ČR
- rozvoj počítačů
  - nultá generace – Z1–3, Colossus, ABC, Mark I,II
  - první generace – ENIAC, MANIAC

V roce 1947 měla RAM kapacitu 100 čísel a sčítání dvou čísel trvalo 1/8 sekundy!

Ranné názory na strojový překlad

- překlad je často opakovaná činnost – věřilo se, že bude tuto proceduru možné počítačem napodobit
- úspěchy použití počítačů v kryptografii: vhodné i pro strojový překlad?

Warren Weaver

When I look at an article in Russian, I say: This is really written in English, but it has been coded in some strange symbols.
I will now proceed to decode.

První impulsy

V roce 1950 rozesílá Weaver memorandum 200 adresátům, ve kterém nastiňuje některé problémy strojového překladu.

- víceznačnost jako častý jev
- průnik logiky a jazyka
- souvislosti s kryptografií
- univerzální vlastnosti jazyka

Zájem o strojový překlad podnícen na několika pracovištích. Do té doby pouze na University of London vedené A. Boothem. Zejména na MIT, University of Washington, University of California, Harvard, Georgetown, . . .

Témata a první výměny zkušeností

- morfologická, syntaktická analýza
- reprezentace významu a znalostí
- tvorba a práce se slovníky
- 1952 – první veřejná konference na MIT
- 1954 – předvedení systému pro strojový překlad

Georgetown experiment

První funkční prototyp strojového překladu.

- 50 vět (zřejmě pečlivě vybraných)
- spolupráce s IBM
- slovník obsahoval 250 slov
- překlad z ruštiny do angličtiny
- gramatika pro ruštinu obsahovala 6 pravidel

Demonstrace systému vyvolala nadšení. MT bylo očividně možné. Následně odstartovalo mnoho nových projektů, hlavně v USA a Rusku.

Vývoj v 50. letech

- MT oblast podnítila rozvoj a výzkum na poli
  - teoretické lingvistiky (Chomsky)
  - počítačové lingvistiky
  - umělé inteligence (60. léta)
- s větším pokrytím kvalita strojového překladu klesala
- i nejlepší systémy (GAT, Georgetown, RE$\rightarrow$EN) poskytovaly nepoužitelný výstup

Zklamání ze slabých výsledků

- i přes nevalné výsledky přetrvával optimismus
- Yehoshua Bar-Hillel píše v roce 1959 kritiku stavu strojového překladu
- tvrdí, že počítače nejsou schopné provádět lexikální desambiguaci
- fully automated high-quality translation (FAHQT) podle Bar-Hillela stěží dosažitelné

Yehoshua Bar-Hillel – příklad pro desambiguaci

Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.

Výdaje na projekty strojového překladu se začaly snižovat.

ALPAC report

- Automatic Language Processing Advisory Commitee
- organizace pod U.S. National Academy of Science
- analýzy a vyhodnocení kvality a použitelnosti systémů SP
- doporučila omezit výdaje na podporu strojového překladu
- negativní dopad na strojový překlad jako vědeckou oblast
- chyba spočívala zřejmě v silném podceňování složitosti porozumění přirozenému jazyku
- vývoj strojového překladu v Evropě a Japonsku pokračoval nepřerušeně dál
- celých 15 let trvalo než SP v USA znovu získal vážnost a původní postavení

Renesance strojového překladu – první velké úspěchy

TAUM-METEO

- překlad z angličtiny do francouzštiny
- od r. 1977 používán pro překlad předpovědí počasí
- vyvinut na University of Montreal

Systran

- velmi populární překladový systém
- využíván v projektu Apollo a Sojuz (od r. 1975)
- od r. 1976 oficiální MT systém používaný Evropským hospodářským společenstvím

Renesance II

80. léta

- vývoj zejména pravidlových systémů s použitím interlinguy
- první daty řízené systémy (Example-based MT)
- rozmach komerčních MT systémů

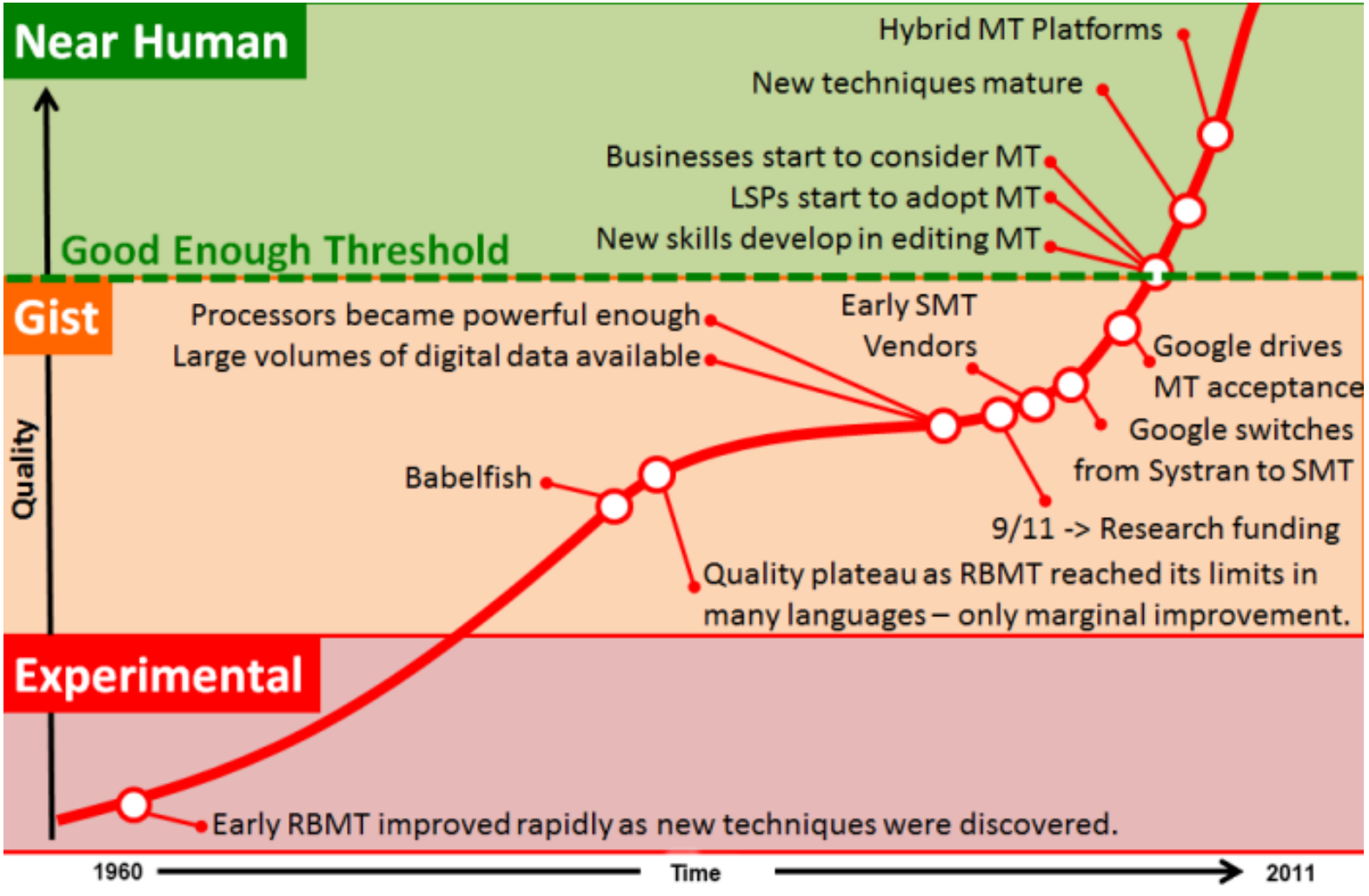
90. léta

- výzkum statistického překladu (IBM)
- pravidlové systémy stále dominují

po roce 2000

- statistické systémy převládají
- kvalita pravidlových systémů je zvyšována statistickými metodami (hybridní metody)
- přidávání dalších jazykových párů

Příliš positivní prognóza



Strojový překlad v současnosti I

- výpočetní technika a datové struktury dovolují práci s miliardami slovy
- Google 1PB sort, rok 2008
  - bilión 100bytových záznamů
  - 6 hodin
  - 4 000 počítačů
  - 48 000 disků
- vývoj MT systému dostupné komukoli
- roste počet paralelních korpusů
- přibývají jazykové zdroje pro minoritní jazyky
- kvalita překladu neustále (byť pomalu) stoupá

Strojový překlad v současnosti II

- SMT rulezz
- intenzivní sběr paralelních dat
- vývoj systémů vzhledem k hodnotícím metrikám
- USA: zájem o angličtinu jako TL
- EU: překlad mezi 23 úředními jazyky EU (EuroMatrix):
  angličtina, bulharština, čeština, dánština, estonština, finština, francouzština, irština, italština, litevština, lotyština, maďarština, maltština, němčina, nizozemština, polština, portugalština, rumunština, řečtina, slovenština, slovinština, španělština a švédština.

Strojový překlad v současnosti III

- korporace (Microsoft) zaměřeny na En jako SL
- velké páry (En↔Sp, En↔Fr): velmi dobrý překlad
- SMT obohacována syntaxí
- Google Translate jako gold standard
- morfologicky bohaté jazyky jsou opomíjeny
- En-* a *-En páry převažují

Motivace pro strojový překlad ve 21. století

- překlad webových stránek pro pochopení obsahu (gisting)
- metody pro výrazné urychlení překladatelské práce (překladové paměti)
- extrakce a vyhledávání informací mezi jazyky (cross-lingual IR)
- instantní překlad elektronické komunikace (ICQ)
- překlad na mobilních zařízeních

Lexikální výběr

Výběr správného překladového ekvivalentu:

- homonymie: *slaď, pila, baby, ženu*
- polysémie: *run, bank, klíč, kohout*
- synonymie: *kluk, chlapec, hoch; dívka, holka, děvče*

Slovosled I

Word order	English equivalent	Proportion of languages		Example languages
SOV	"I you love."	45%		Hindi, Japanese, Latin
SVO	"I love you."	42%		English, Mandarin, Russian
VSO	"Love I you."	9%		Hebrew, Irish, Zapotec
VOS	"Love you I."	3%		Baure, Fijian, Malagasy
OVS	"You love I."	1%		Apalai, Hixkaryana, Tamil
OSV	"You I love."	0%		Jamamadi, Warao, Xavante

Slovosled II – volný slovosled

Čím více morfologicky bohatší, tím volnější slovosled.

Katka **snědla** kousek koláče.

- Kati megevett egy szelet tortát → Katie eating a piece of cake
- Egy szelet tortát Kati evett meg → Katie ate a piece of cake
- Kati egy szelet tortát evett meg → Katie ate a piece of cake
- Egy szelet tortát evett meg Kati → Katie ate a piece of cake
- Megevett egy szelet tortát Kati → Katie eating a piece of cake
- Megevett Kati egy szelet tortát → Katie ate a piece of cake

Závěr

- strojový překlad patří mezi AI-complete problémy
- máme k dispozici obrovskou výpočetní sílu
- tržní potenciál je větší než kdy dřív
- je stále co zlepšovat
- statistické metody se zdají vhodnější (rychlé, levné)
- nové nápady jsou vítány! (BP, DP)