European Language Resource Coordination
Connecting Europe Facility

Jazyky a jazykové technologie v České republice

Tomáš Svoboda
Jan Hajič

Kolik jazyků máme v ČR?

- Jeden úřední jazyk (čeština)
- Skutečnost je složitější
  - Turistika
  - Obchod a mezinárodní komerční aktivity
  - Migrace
  - Regionální rozvoj, vývoj, státní správa, mez, spolupráce
    - Mobilita a doprava
    - Energetika, změna klimatu
    - Životní prostředí a
  - EU: přeshraniční spolupráce a mobilita
    - Zákony, vyhlášky, nařízení (pracovní, obchodní, sociální, turistika)
    - Zdravotní dokumentace, rychlá pomoc

European Language Resource Coordination
Connecting Europe Facility

Jazyky v ČR

- Čeština je výrazně převažující jazyk
- Češi mají dva cizí jazyky ve škole
  - Angličtina
  - Němčina, francouzština, ruština (další méně)
- Jazyky menšin:
  - Slovenština
  - Romština (Sinti)
  - Ruština, vietnamština, ukrajinština, polština, angličtina
  - Nářečí (Ostrava, Slezsko, Slovácko)

Jazyky sousedních zemí

Regionální spolupráce (Interreg aj.)

Region	Countries	Languages
Střední Evropa	Německo (část), Polsko, ČR, Rakousko, Slovensko, Maďarsko, Itálie (sever), Slovinsko, Chorvatsko	Němčina, polština, slovenština, maďarština, italština, slovinština, chorvatština
Dunajský	Německo (část), Bulharsko, Chorvatsko, ČR, Maďarsko, Rumunsko, Slovensko, Slovinsko	Němčina, bulharština, slovenština, maďarština, rumunština, slovinština, chorvatština

A další...
(např. NATO, OBSE, UNICEF, WHO, OSN…)

Vícejazyčnost v regionální spolupráci

- Přeshraniční

  - Rychlá pomoc, pomoc při nehodách, traumatologie
    - Zdravotní turistika
  - Právní předpisy a požadavky
    - osobní, ekonomická činnost, státní správa
  - Bezpečnost, policie, armáda, soudy, zastupitelství
  - Infrastrukturní projekty
    - Silnice a železnice
    - Energetika (ropa, plyn, přenos el. energie)
    - Pošta a komunikace
  - Finanční instituce a spolupráce

Migrace a zahraniční pracovníci

- Občané jiných zemí v ČR
  - 2009: 3,9 % (407 500) (srov.: Polsko 0,1 % - 39 500 lidí)
    - Z toho 1,4 % z EU, zbytek mimo EU
    - Nejvíce: Slovensko (82 tis.), Asie souhrnně, Ukrajina, i Německo (14 tis.)
- Migrace: tranzitní země
  - Dosud nevelká frekvence
  - Mimoevropské jazyky
- Zahraniční pracovní síla
  - Slovensko, Ukrajina, Rusko, Polsko
- Studenti a zahraniční učitelé
  - Slovensko
  - Rusko, Ukrajina, Vietnam, USA
  - Učitelé jazyků
  - Výzkumníci a vědci

ČR a jazykové technologie

- Jazykové technologie
  - Automatická (počítačová) analýza textu a mluvené řeči
    - Také syntéza (tzv. Text-To-Speech, nebo generevání textu)
  - Aplikace:
    - Kontrola pravopisu, kontrola gramatiky, automatické opravy, doplňování diakritiky
    - Automatický překlad (z a do češtiny)
    - Vyhledávání v textech (např. na internetu)
      - Bez ohledu na slovní tvary (tzv. tvaroslovná analýza)
    - „Big data“ analytika
      - Analýza sociálních sítí (polarita názorů)
      - Call centra (směrování hovorů, hodnocení operátorů)
    - Telefonické informační systémy (např. PID)
    - Kombinace „dolování“ (zajímavých) informací z textu a dat

Jazykové technologie v ČR: aktéři

- Výzkumné organizace
  - Univerzita Karlova v Praze, MFF (Ústav formální a aplikované lingvistiky) – vedoucí vývojové středisko v oblasti automatického překladu (v EU), podíl na vývoji světově nejrozšířenějšího systému „Moses“
  - Vysoké učení technické v Brně (Speech@FIT)
  - Fakulta aplikovaných věd ZČU Plzeň (Katedra kybernetiky, NTIS)
  - Ústav pro jazyk český AV ČR, v.v.i.
  - Masarykova univerzita v Brně (NLP Lab, Katedra poč. systémů a komunikací)
  - Technická univerzita Liberec (Katedra mechatroniky)
- Firmy
  - Phonexia (Brno), Lingea (Brno), Speechtek (Plzeň), Good Data (Praha), Geneea (Praha), Newton, ...
  - Jazykové technologie využívají (ve spolupráci s univerzitami a AV) např. ASPI/Kluwer, seznam.cz, centrum.cz, Google, ...; soudy, ČT
- Infrastruktura pro výzkum a vývoj
  - LINDAT/CLARIN, MFF UK (+3), součást Clarin ERIC (EU)

Jazykové technologie v ČR: stav

- Jazykové technologie
  - Neexistuje kritérium „funguje/nefunguje“: vždy funguje „v procentech“
  - Příklady:
    - Kontrola pravopisu: chybějící slova, nepozná správnost „jsme/jsem“ apod.
    - Převod slov na základní tvary pro vyhledávání: chybějící slova ve slovníku
    - Automatický překlad – nikdy nebude perfektní (každý zná Google Translate)
- Čeština – stav dostupných jazykových technologií
  - Popsáno v Bíle knize jazyků (cihlová obálka...)
    - Stav jazykových technologií, stav jazykových zdrojů
    - S výjimkou vyzrálosti a kvality zpracovaných textových korpusů (dat) < 5 bodů (na škále 1-5)
    - Automatický překlad souhrnně: nízká podpora
      - Hlavní důvod: nedostatek dat (přeložených textů)
- ***Kvalita závislá na dostupnosti jazykových dat (textů, audia)***

European Language Resource Coordination
Connecting Europe Facility

Odkazy

- Bílá kniha jazyků
  - http://www.meta-net.eu/whitepapers/volumes/czech
- Celoevropská koordinační síť jazykových technologií
  - http://www.meta-net.eu
- Infrastruktura pro distribuci dat a základních nástrojů
  - http://lindat.cz
- Nástroje pro překlad (open source)
  - http://www.statmt.org/moses
- Instituce:
  - ÚFAL MFF UK: http://ufal.mff.cuni.cz
  - Speech@FIT VUT Brno: http://speech.fit.vutbr.cz
  - ÚJČ AV ČR Praha: http://www.ujc.cas.cz
  - KKY ZČU Plzeň, odd. UI: http://www.kky.zcu.cz/cs/research-fields

Děkuji za pozornost!