ŽÁKOVSKÝ KORPUS MERLIN: JAZYKOVÉ ÚROVNĚ A TROJJAZYČNÁ CHYBOVÁ ANOTACE

Mgr. Barbora Štindlová, Ph. D., Mgr. Veronika Čurdová,
Mgr. Petra Klimešová, Mgr. Eva Levorová
ÚJOP UK, Praha

Práce s chybou, Poděbrady
17. – 18. 6. 2014

Education and Culture DG
Lifelong Learning Programme

OBSAH

1. Projekt Merlin
2. Sběr a zpracování dat
3. Anotace
4. Cílová hypotéza
5. Vybrané problémy anotace
6. Závěry

1. PROJEKT MERLIN

- Multilingvální platforma pro evropské referenční úrovně: Výzkum jazyka studentů v kontextu

http://www.merlin-platform.eu

  - vychází ze SERR (referenční úrovně a deskriptory)
  - pracuje se třemi jazyky z různých větví indoevropské jazykové rodiny: němčina (germánská), italština (románská) a čeština (slovanská)

1. PROJEKT MERLIN: CÍLE

- vytvořit online platformu umožňující vyhledávat konkrétní jazykové rysy typické pro danou jazykovou úroveň (A1 – C1)
  - přispět k vymezení úrovní definovaných podle SERR
- vybudovat multilingvální žákovský korpus
  - s klasifikací chyb použitelnou pro všechny tři jazyky zároveň
- výsledný produkt bude sloužit učitelům, examinátorům, metodikům, tvůrcům učebnic a studentům

2. **SBĚR A ZPRACOVÁNÍ DAT**

- zdroj dat
  - standardizované testy (telc, UNIcert, CCE)
  - pro český subkorpus
    - zkoušky CCE organizované ÚJOP UK (psaní)
    - úrovně A2, B1 a B2 (vždy asi 150 textů)

2. SBĚR A ZPRACOVÁNÍ DAT

- proces zpracování
  - 1. přepis do elektronické podoby
    - XMLmind editor
    - zachování původní podoby textu
    - transkripční manuál (vsuvky, škrty, nečitelné části, obrázky, anonymizace ..)
    - metadata
  - 2. anotace
    - návrh cílové hypotézy a klasifikace chyb na dvou úrovních

E:\Bara\projekty\Merlin\Czech_B1_Transcriptions_143\0605.xml

File Select Edit Search View Tools XML Window Options Help

merlin_document ▸ #comment

0605.xml

Copyright 2012 project Merlin, http://merlin-platform.eu
Transcriber: LZAHR
Author ID: 0605
Exercise
Milá Aleno!

Děkuji Ti mnohokrát za Tvůj email, dostala ~~jesem~~ jsem je před týdnem! Mám moc černé svědomí, že jsem Ti už dlouho nepsala! Nezlob na mne! ~~Neměla jsem~~ Měla jsem hodně prace poslední době. Učila jsem na českou zkouškou, však dnes mám volno :). Mimo to, děkuji za krásnou fotkou z dovolené, měla jsi moc radost!?!

Co [!]řicí řici[!] o mne život? Nic zajimavého nestalo, snad jen to, že plánuju dovolenou v létě s =Lukášem=. Chceme jet do Norska. To je hodně prace, ale hezký :).

Studium je dobré, v březnu musím psat sociologickou zkouškou, to je [!]těžké těžké[!]! ~~Jde~~ Ale Jede to! To je [!]vyborně výborné[!] novin~~k~~u, Alena ^bude^ blízko Drážďan v létě! Doufám, že můžeš pracovat ve firmě! Samozřejmý mám čas! Už se velmi [!]těším těším[!] na tvoji návštěvu. Srdečně chtěla bych Ti pozvat ~~o~~ k mně! Kdy přijdeš? Kde si dáme sraz? ~~Došla jsem D Došla bych ti radá!~~

Co se týká kurzu - to je dobrý nápad! Kdy přesně bude? A kolik stojí? Asi můžeš ~~tti~~ mi poslat inzerát?! Chtěl by ~~rád~~ mílé.

~~J~~ Dobře! Pro dnešek končím, je už pozdě večer, musím vstávat brzo zítra!

~~Všecho~~

~~Vš~~ Prejí Ti a tvým rodičům všechno nejlepší a [!]zvláštně zvláště[!] Ti hodně uspechů v 13. dubna! ~~Myslím na T~~

Už těším za na setkání!

Zatím Ahoj,

=Kateřina=

PS: Pozdravuj =Tomáše= od mne!

```
<merlin_document
    xsi:schemaLocation="http://www.eurac.edu/merlin file:
  <!--Copyright 2012 project Merlin, http:/ -->
    /merlin-platform.eu
  <transcriber> LZAHR </transcriber>
  <author_id> 0605 </author_id>
  <body>
    <exercise xml:space="preserve">
      Milá Aleno!
      <par/>
      Děkuji Ti mnohokrát za Tvůj email, dosta
      la
      <correction>
        <deletion> jesem </deletion>
      </correction>
       jsem je před týdnem! Mám moc černé svěd
      omí, že jsem Ti už dlouho nepsala! Nezlo
      b na mne!
      <correction>
        <deletion> Neměla jsem </deletion>
      </correction>
       Měla jsem hodně prace poslední době. Uč
      ila jsem na českou zkouškou, však dnes m
      ám volno
      <emoticon> :) </emoticon>
      . Mimo to, děkuji za krásnou fotkou z do
      volené, měla jsi moc radost!?!
      <par/>
      Co
      <error>
        <originalForm> řicí </originalForm>
```

Edit

Validity

INS

3. ANOTACE

- anotační schéma
  - reflektuje indikátory popisující aspekty žákovského jazyka
- cílová hypotéza
  - stanovena ve dvou kolech (na základě pravidel užívaných v korpusu FALKO – HU Berlín)

3. ANOTACE: INDIKÁTORY

- vymezení indikátorů, jimiž lze popsat charakter žákovského jazyka (standardní i nestandardní formy)
- vhodné pro všechny tři jazyky (N, I, ČJ)
- základ pro anotaci (ale i analýzu) dat

3. ANOTACE: INDIKÁTORY

- více zdrojů
  1. využití deskriptorů, s nimiž pracuje SERR
  2. výzkum odborné literatury zaměřené na osvojování jazyka a metody hodnocení žáků cizích jazyků
  3. empiricky založené indikátory (analýza učebnic, ankety mezi studenty, vyučujícími a zkoušejícími)
  4. indikátory získané analýzou samotných žákovských projevů

3. ANOTACE: INDIKÁTORY

1. indikátory podle SERR
   - např. *pozdravy/rozloučení, kolokace, lexikální variabilita …*
   - problém měřitelnosti , resp. operacionalizace deskriptorů (např. *srozumitelnost, koherence …)*

2. deduktivně vymezené indikátory
   - rozsáhlá analýza odborné literatury
   - ortografie, gramatika, slovní zásoba, koherence/koheze, sociolingvistická adekvátnost

3. ANOTACE: INDIKÁTORY

3. empiricky vymezené indikátory
   - analýza učebnic, anketa
   - např. *modální slovesa* (N, ČJ), *časová souslednost* (I), *apostrofy a diakritika* (I, ČJ), *slovosled v otázkách, vyjadřování zdvořilosti …*

4. induktivně vymezené indikátory
   - lingvistická analýza produkce nerodilých mluvčích
   - např. *záměna slovního druhu, reflexivní slovesa, klitika …*

3. ANOTACE: ANOTAČNÍ SCHÉMA

- výběr relevantních indikátorů a jejich transformace do anotačního schématu
  - rysy společné i jazykově specifické
- kombinace dvou přístupů
  - značkování formální odlišnosti od cílové hypotézy (*chybějící element, přebývající element, chybně spojené elementy* ap.)
  - hierarchicky strukturovaná klasifikace lingvistická (*chyba ortografická, gramatická, lexikální* a jejich podrobnější klasifikace dle slovnědruhové i větněčlenské platnosti)
- detailní dokumentace a manuál s příklady pro anotátory

3. ANOTACE: PROCES ANOTACE

- digitalizace
  - transkripce, in-line anotace…
  - anotační schéma
- automatická anotace
  - tokenizace, lematizace, POS …
- manuální anotace
  - dvoufázová v souvislosti s cílovou hypotézou
- statistika

4. CÍLOVÁ HYPOTÉZA

- cílová hypotéza (*target hypothesis, **TH***)
  - rekonstrukce studentova projevu s minimálními zásahy
  - základem pro anotaci (resp. chybovou anotaci, *EA)*

- MERLIN > 2 cílové hypotézy
  - TH1: minimální (ortograficky a gramaticky ’správná‘ věta)
  - TH2: změny na sémantické a pragmatické rovině (lexikální, koheze a koherence textu apod.)

UKÁZKA JEDNOTLIVÝCH ROVIN

tok	TH1	EA1	EA1	TH2	EA2
Je	Je			Je	
profesorka	profesorka			profesorka	
z	z			z	
Německa	Německa			Německa	
a	a			a	
učí	učí			učí	
němčtinu	němečtinu	O_Graph_act_Ad	O_Graph_act_Ad	němčinu	V_Wordform_deriv
na	na			na	
Karlové	Karlově	O_Graph_act_ch		Karlově	
Univerzitě	univerzitě	O_Capit		univerzitě	
v	v			v	
Praze	Praze			Praze	
.	.			.	

Přišti semestru budu psát diplomovou práce.

TH1 + EA1:

Příští (O_Graph_act_O, O_Graph_act_O) *semestr* (G_Morphol_case_wrong) *budu psát diplomovou práci* (G_Morphol_case_wrong).

- O_Graph_act_O
  - ortografie : grafém : chybí diakritika
- G_Morphol_case_wrong
  - gramatika : flexe : chyba v pádu

5. VYBRANÉ PROBLÉMY ANOTACE

- minimální zásahy do textu

*Vzala si své **oblíbenější** botičky.*

*Myslíš, že **budeš končit** sraz v 5 hodin?*

***Zvadríš** všechny a hlavně Petra.*

***Sle** fotky?*

- ***kratke*** *kalhoty /* ***dětí*** *nemluví /* ***mužů*** *přijít*
  - ? ortografie n. flexe
- *Chtěl bych tě pozvat ke mně* ***doma.***
  - ? TH1 (valence) n. TH2 (lexikum)
- ***kdyby bychom, že jsi se***
  - ? ortografie n. morfologie

- *Možná mluví **o čem** udělají...*
  - ? spojka (korelativum) n. valence *(udělají)*
- Jsou má dovoleny.
- Můžu **docházet** pro tebe.
- Prázdninový kurz češtiny **to** je dobrý nápad.
- Ze cloveku práce ne dost pozitivu vsehno bude udělat spatně počasí v jeho důse.

Děkuji za pozornost !

Lifelong Learning Programme
(nr. *518989-LLP-1-2011-1-DE-KA2-KA2MP)*