MERLIN
Multilingvální platforma pro evropské referenční úrovně

Barbora Štindlová, Veronika Čurdová

Ústav jazykové a odborné přípravy
Univerzita Karlova v Praze

Korpusová lingvistika Praha 2014
Praha, 17. – 19. 9.

OBSAH

1. Projekt Merlin
2. Data
3. Anotace
4. Cílová hypotéza
5. Problémy
6. Závěry

OBSAH

1. PROJEKT MERLIN

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

  (Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context)

  - vychází ze SERRJ (referenční úrovně a deskriptory)
  - 3 jazyky (čeština, němčina, italština)

  http://www.merlin-platform.eu

  - Lifelong Learning Programme
    (nr. *518989-LLP-1-2011-1-DE-KA2-KA2MP)*
  - 01/2012 – 12/2014
  - **Technische Universität Dresden** (DE) *(koordinátor projektu)*
  - **EURAC** (IT), **Univerzita Karlova** (CZ), **Eberhard-Karls-Universität Tübingen** (DE), telc GmbH (DE), Berufsförderungsinstitut Oberösterreich (AT), European Center of Modern Languages – Council of Europe (AT) *(partneři)*

1. PROJEKT MERLIN: CÍLE

- vytvořit online platformu umožňující vyhledávat konkrétní jazykové rysy typické pro danou jazykovou úroveň (A1 – C1)
  - přispět k vymezení a ověření úrovní definovaných podle SERRJ
- vybudovat multilingvální žákovský korpus
  - s klasifikací chyb použitelnou pro všechny tři jazyky zároveň
- cílová skupina: učitelé, metodici, examinátoři, tvůrci výukových materiálů a testů, SLA/FLT odborníci

OBSAH

2. DATA: SBĚR

- zdroj
  - standardizované testy (telc, UNIcert, CCE)
    - psaná produkce
    - pro český subkorpus
      - zkoušky CCE organizované ÚJOP UK (psaní)
      - úrovně A2, B1 a B2 (vždy asi 150 textů)
  - metadata
    - věk, rod, L1, úroveň podle SERRJ, testovací instituce, datum, úkol

	čeština	němčina	italština	Celkem
A1	1	57	30	88
A2	49	199	294	542
A2+	112	107	94	313
B1	89	219	343	651
B1+	75	115	53	243
B2	72	219	2	293
B2+	9	73		82
C1	4	43		46
C2		4		4
Celkem (texty)	411	1,035	816	**2,262**
Celkem (slova)	64,488	125,927	92,359	**282,774**

merlin

2. DATA: ZPRACOVÁNÍ

- 1. přepis
  - sken rukopisu
  - XMLmind editor (in-line anotace)
    - detailní pravidla pro přepis
      - zachování původní podoby textu
      - vsuvky, škrty, nečitelné části, obrázky, …
      - anonymizace (místní a osobní jména)
- 2. konverze
  - PAULA (XML formát)

    http://purl.org/net/paula

0617.xml | 0819.xml

Copyright 2012 project Merlin, http://merlin-platform.eu

Transcriber: PJAKO

Author ID: 0617

Exercise

8.2.'07

Ahoj Aleno!

Děkuju za tvůj ~~dopis~~ [!]email e-mail[!]. Mám se dobře. Už nepracujem na diplomce, [!]pretože protože[!] mám moc čas, ale plánujem dovolena na [!]slovensku Slovensku[!]; v létě[??].

[!]Těšim Těším[!] se, že přijdeš do [!]Ďražďan Drážďan[!]. Chceš [!]navštevovat navštívit[!] mě doma [!]odpolodne odpoledne[!]? Kdy přijdeš? ~~[??]~~ Budu dojdu ~~[??]~~^ty^ na [!]nadráží nádraží[!].

Kdy je prázdninové^ý^ kurz~~y~~ [!]češtiný češtiny[!] a jak dlouho [!]potrva potrvá[!] se? [!]Nevi~~e~~m Nevím[!] že můžu se dokonce ucházet o stipendium, ale [!]samostatne samostatně[!] bylo by [!]výborne výborné[!], kdybychom se ~~[??]~~ [!]videme viděli[!] často v létě. Kolik stojí tento kurz? [!]Mužeš Můžeš[!] poslat tento inzerát ku [!]mě mně[!]?

Teď, nevím, že~~[??]~~ budu [!]navštevovat navštěvovat[!] [!]tě~~b~~e tebe[!] v létě, [!]pretože protože[!] chcem pojet do Prievidzi s letadlem na letiště a tam chcem [!]letát létat[!] od Vysokou do Vysokou [!]tatra Tater[!]; a nazpět. Snad budu [!]letát létat[!] k tě~~b~~e. :-)

Srdečně tě zdravím.

=David=

```
<transcriber> PJAKO </transcriber>
<author_id> 0617 </author_id>
<body>
  <exercise xml:space="preserve">
    8.2.'07
    <par/>
    <greeting> Ahoj Aleno! </greeting>
    <par/>
    Děkuju za tvůj
    <correction>
      <deletion> dopis </deletion>
    </correction>
    <error>
      <originalForm> email </originalForm>
      <targetForm> e-mail </targetForm>
    </error>
    . Mám se dobře. Už nepracujem na diplomc
    e,
    <error>
      <originalForm> pretože </originalForm>
      <targetForm> protože </targetForm>
    </error>
    mám moc čas, ale plánujem dovolena na
    <error>
      <originalForm> slovensku </originalForm>
      <targetForm> Slovensku </targetForm>
    </error>
    <correction>
```

[1] element contains characters other than white space [cvc-complex-type.2.3]

Edit | Validity | INS

OBSAH

3. ANOTACE

- 2 perspektivy popisu žákovského jazyka
  - FLT: chybová anotace
  - SLA: lingvistické rysy mezijazyka

3. ANOTACE: INDIKÁTORY

- vymezení indikátorů, jež umožní popsat charakter žákovského jazyka
  - vhodné pro všechny tři jazyky (N, I, ČJ)
  - standardní i nestandardní formy
  - zvládnutelné pro anotátory

3. ANOTACE: INDIKÁTORY

- zdroje pro specifikaci

1. **škála SERRJ (deskriptory)**

- např. spojovací výrazy, lexikální variabilita, kolokace …
- problém měřitelnosti , resp. operacionalizace deskriptorů (např. *srozumitelnost, koherence …)*

2. **výzkum SLA a jazykového testování**

- rozsáhlá analýza odborné literatury
  - oblasti: ortografie, gramatika, slovní zásoba, koherence/koheze, sociolingvistická adekvátnost / pragmatika

3. dotazník (anketa mezi experty)

- např. *modální slovesa* (N, ČJ), *časová souslednost* (I), *apostrofy a diakritika* (I, ČJ), *slovosled v otázkách, vyjadřování zdvořilosti …*

4. empiricky vymezené indikátory

- učebnice a analýza produkce nerodilých mluvčích
- např. *záměna slovního druhu, reflexivní slovesa, klitika, dvojí negace, formálnost vyjadřování …*

3. ANOTACE: SCHÉMA

- výběr relevantních indikátorů pro Č, N a I a jejich transformace do anotačního schématu

  - společné rysy (např. *connector accuracy, subject – verb agreement, verb tense, collocations …)*

  - jazykově specifické rysy (např. Č: *dvojí negace, posesivní reflexíva;* I: *lexikalizovaná klitika,* N: *modální částice,* I/N: *členy…)*

			Abbreviation(s) of the tag	LANGUAGE SPECIFITY? (ITA/GER/CZ	TARGET LANGUAGE MODIFICATI TO BE SPECIFIED (omission/ad Y/N?	DESCRIPTION OF THE TAG	SPAN OF THE TAG	EXAMPLES	SOURCE (CEFR, inductive, user based ...)	Identical or very similar TAG IN ANOTHER AS (Source, Name, Link)
GRAMMAR	**negation**	negation general	**G_Neg_neggen_Pos** **G_Neg_neggen_O** **G_Neg_neggen_Ch** **G_Neg_neggen_Ad**	partly	Y	Error tag This tag is to be used in case of a) wrong placement of negation expressions (tag: G_Neg_neggen_Pos) b) missing part of negation (tag: G_Neg_neggen_O) c) wrong use of negation words (GER: nicht, nein, kein; ITA: no, non; CZE: ne, žádný (tag: G_Neg_neggen_Ch) d) redundant/wrongly added negation word (G_Neg_neggen_Ad)	a) _pos: 1 token; 2 tokens for CZE (neg.word and verb, if neg. word wrongly distributed) b) _o: 1 or more token - the POS to be negated; c) _ch: 1 token - the negation word	a) *[mám ne] kávu, *[půjdu neráno], *Ich habe Hunger [keinen]; *Io credo [non]. b) *On [ne] velký; *Io [mangio] nè carne nè pesce {Io non mangio né carne né pesce}; *Luca non va a scuola perché ne [ha] voglia; *Non è né chiaro [scuro] {Non è né chiaro né scuro}. c) *Bohužel, nemám [ne] čas; *Ich habe [nicht] Zeit. *Er wird dort arbeiten [nein]; *[Non], viene più tardi. d) *Man kann nicht auf solchen grössen Teil seiner Personlichkeit zu verzichten, ohne seine psychische Gesundheit [nicht] zu schaden.	RB	
		double negation	**G_Neg_negdoub**	CZ	Y	Error tag In Czech all negated pronouns require a negated verb form. This tag includes the missing negation particle "ne" at the verb.	1 token (verb, existing part of the double negation)	*[mám] žádný čas {nemám žádný čas} *nikdo [volal] → {nikdo nevolal}	textbook	
	Verb Valency (obligatory arguments)	complement number	**G_Valency_complnumb_O** **G_Valency_complnumb_Ad**			Error tag Definition: Verb valency refers to the number of arguments controlled by a verbal predicate. Verb valency includes all obligatory arguments, including the subject of the verb. A complement can be realized as adjective phrase (Die Sitzung dauerte	wrong argument (usually 1 token) or whole clause including punctuation mark (if argument is missing)	a) *Já vstávám v 5, [ale vstáváš v 8.] {...ale ty vstáváš v 8.} *Ich liebe {dich}. *Er hat uns nicht gesagt, ob {er} kommen will. *[Spero che possa aiutarLa.]		

3. ANOTACE: SCHÉMA

- 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
    - hierarchická (3 úrovně: jazyková oblast, suboblast a specifický rys)
- detailní anotační manuál s příklady

3. ANOTACE: POSTUP

- digitalizace
  - přepis
    - *XMLmind* editor
  - konverze
    - *PAULA*
- anotace
  - manuální: 2kolová (TH1 and TH2)
    - *MMAX2* a *Falko Excel AddIn*'s
  - automatická: tokenizace, lemmatizace, POS …
    - *UIMA*
- prohlížeč (vizualizace)
  - *ANNIS*

OBSAH

4. CÍLOVÁ HYPOTÉZA (TH)

- rekonstrukce studentova projevu
  - základ pro identifikací nestandardní jevů (pro anotaci)

- MERLIN > 2 cílové hypotézy
  - korpus FALKO (minimální x rozšířená TH)

  - TH1: správnost
    - minimální (?) změny v ortografii, morfologii, syntaxi
    - ortograficky a gramaticky ’správná‘ věta
  - TH2: přiměřenost
    - změny na sémantické a pragmatické rovině
    - lexikum, koheze a koherence textu apod.

tok	TH1	EA1	EA1	TH2	EA2
Tibor *Tibor*	Tibor			Tibor	
je *is*	je			je	
z *from*	z			z	
Madarsku *Hungary*	Maďarska	O_Graph_graphgen_O	G_Morphol_case_wrong	Maďarska	
a *and*	a			a	
studuje *he studies*	studuje			studuje	
v *in*	v			v	
praze *Prague*	Praze	O_Capit		Praze	
na *at*	na			na	
filozofske *philosophical*	filozofské	O_Graph_graphgen_O		filozofické	V_semdenot_word/fs_1
fakulté *faculty*	fakultě	O_Graph_graphgen_Ch		fakultě	
.	.			.	

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*

OBSAH

5. PROBLÉMY: ČEŠTINA

TH1

? cílová hypotéza

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

nejoblíbenější

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

? skončíš /
ukončíš /
končí$_{3sg}$ /
skončí$_{3sg}$

3. ***Jsou** **má** dovolenu.

?? Jsou na dovolené. X Má dovolenou.

4. ***Zvadríš** všechny a hlavně Petra.

Zdravíš
(pozdravuj /
pozdravíš)

5. ***Sle** fotky?

? Pošlu / pošleš/
pošle / šli …

6. ***kratke*** *kalhoty*	7. ***dětí*** *nemluví*	8. ***mužů*** *přijít*
krátké	*děti*	*můžu*

? ortografie nebo flexe

9. ***kdyby bychom*** *koupili*

kdybychom

10. *že* ***jsi se*** *učil*

ses

? ortografie n. morfologie

11. *Možná mluví* ***o čem*** *udělají...*

... o tom$_{Loc}$, co$_{Acc}$ ....

? spojka *(„korelativum“)* n. valence *(mluví/udělají)*

???

12. Ze cloveku práce ne dost pozitivu vsehno

Že člověku práce ne dost ?pozitivní ?všechno

bude udělat spatně počasí v jeho důse.

?udělá špatné počasí v jeho ?duši

5. PROBLÉMY: ČEŠTINA

TH2

13. Srdečně *tě* *zdravím,* ***Tvoje*** *Dana.*

V_Txt_grfw

tagy bez opravy

14. Chtěl bych *tě* *pozvat* *ke mně* ***doma***

domů

? TH1 (valence) n. TH2 (lexikální ch.)

15. Kdy *je* ***ho?***

[on] / to

16. ***Docházím*** *na* *nádraží,* *samozřejmě.*

Dojdu

??? vid

17. *se svými*	***dítaty***		18. *rád*	***psám***		*19. já*	***být***	*strach*
	dětmi		*rád*	*píšu*			*mám*	
	? dítmi			*? psu*			*? jsem*	
	? dětaty			*? píšám*				

??? oddělené značkování TH1 a TH2

OBSAH

1. Projekt Merlin
2. Data
3. Anotace
4. Cílová hypotéza
5. Problémy
6. Závěry

MERLIN

- detekce jazykových rysů, které odpovídají úrovni znalosti jazyka na příslušné referenční úrovni
- zahrnuje 3 L2
- robustní výběr anotovaných jevů
- spolehlivost anotace je zcela zásadní!
  - proškolení anotátorů
  - přehledná a podrobná dokumentace
  - nutnost kontroly
    - supervize
    - dvojí anotace a IAA

REFERENCES

- Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., Vettori, Ch. (2014). The MERLIN corpus: Learner Language and the CEFR. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)}, May 26-31, 2014. ELRA, Reykjavik.

- Reznicek, M., Lüdeling, A., Krummes, C., and Schwantuschke, F., (2012). Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0. http://purl.org/net/falko-maual.pdf.
- Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In Díaz-Negrillo, A., Ballier, N., and Thompson, P., editors, Automatic Treatment and Analysis of Learner Corpus Data, pp. 101–123. John Benjamins, Amsterdam.

- Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., Abel, A., and Hana, J. (2013). MERLIN: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni.

- Wisniewski, K. (2013). The empirical validity of the CEFR fluency scale: the A2 level description. In Galaczi, E. D. and Weir, C. J., editors, Exploring Language Frameworks: Proceedings of the ALTE Krakow Conference, Studies in Language Testing, pp. 253–272. Cambridge University Press, Cambridge.

Děkuji za pozornost!

Barbora Štindlová
za MERLIN-Team