MERLIN
Illustrating European Reference Levels in Three Languages

Barbora Štindlová, Veronika Čurdová

Institute for Language and Preparatory Studies
Charles University in Prague

Grammar and Corpora 2014
Warsaw , June 25–27 2014

OUTLINE

1. Merlin project
2. Data
3. Annotation
4. Target hypothesis
5. Problems
6. Conclusions

OUTLINE

1. Merlin project
2. Data
3. Annotation scheme and process
4. Target hypothesis
5. Problems
6. Conclusions

1. MERLIN PROJECT

- **M**ultilingual Platform for the **E**uropean **R**eference **L**evels: **In**terlanguage Exploration in Context

  - related to CEFR (levels, descriptors)
  - 3 languages (Czech, German, Italian)

  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) *(Lead Partner)*
  - **EURAC** (IT), **Charles University** (CZ), **Eberhard-Karls-Universität Tübingen** (DE), telc GmbH (DE), Berufsförderungsinstitut Oberösterreich (AT), European Center of Modern Languages – Council of Europe (AT) *(associated partners)*

1. MERLIN PROJECT: AIMS

- to develop an open online platform with authentic learner data
  - to create a multilingual database (learner corpus)
  - to use annotation applicable to all three languages
- to illustrate CEFR levels and contribute to the validation of the concept
- target group: teachers, teacher trainers, test developers, SLA researchers

OUTLINE

merlin

1. Merlin project
2. Data
3. Annotation scheme and process
4. Target hypothesis
5. Problems
6. Conclusions

2. DATA: COLLECTION

- source of data
  - standardized tests (telc, CCE)
    - written production
  - metadata
    - age, gender, L1, CEFR level, test institution, date, task

	Czech	German	Italian	Total
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
Total (texts)	411	1,035	816	**2,262**
Total (words)	64,488	125,927	92,359	**282,774**

2. DATA: PREPARATION

merlin

- 1. transcription
  - scanning of hand-written texts
  - XMLmind editor
    - data transcribed according to detailed rules
    - inline annotation, marking insertions, deletions, ambiguous and unreadable tokens, emoticons, …
    - personal and place names anonymization, identification of foreign words and direct citations of the test prompts, …
- 2. conversion
  - PAULA (XML format)

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

OUTLINE

merlin

1. Merlin project
2. Data
3. Annotation
4. Target hypothesis
5. Problems
6. Conclusions

3. ANNOTATION

merlin

- description of learner language from two perspectives
  - FLT: error annotation
  - SLA: linguistic characteristics of the learner language

3. ANNOTATION: INDICATORS

- reflect indicators describing features and characteristics of learner language
  - indicators appropriate to all 3 languages
  - both standard and non-standard forms
  - manageable for annotation

3. ANNOTATION: INDICATORS

- several sources

1. **CEFR scales (descriptors)**

- e.g. connector accuracy, content jumps, collocation usage …
- problem of operationalization (e.g. *intelligibility of the text, level of text elaboration, coherence …*)

2. **SLA and language testing research**

- extensive literature review (areas: orthography, grammar, vocabulary, coherence/cohesion, sociolinguistic appropriateness/pragmatics)

3. ANNOTATION: INDICATORS

3. questionnaire study (expert interviews)

- e.g. *modal verbs* (D, CZ), *diacritics* (I, CZ), *sequence of tenses* (I) …

4. experience-based indicators

- textbooks and learner texts analyses
- e.g. *double negation, clitic usage, POS confusion, level of formality …*

3. ANNOTATION: SCHEME

- selection of relevant indicators for CZ, D, I
  and their transformation into the annotation scheme

  - common features (e.g. *connector accuracy, subject – verb agreement, verb tense, collocations …)*

  - language-specific features (e.g. CZ: *double negation, possessive reflexive pronouns,* I: *lexicalised clitics,* D: *modal particles,* I/D: *articles …)*

merlin

			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.]		

- combination of
  - linguistic classification
    - hierarchical (3 levels: linguistic field, subfield and specific phnomenon)
  - target modification
- detailed annotation manual (examples)

3. ANNOTATION: WORKFLOW

merlin

- digitalization
  - transcription
    - *XMLmind* editor
  - conversion
    - *PAULA*
- annotation
  - manual: 2 rounds (TH1 and TH2)
    - *MMAX2* and *Falko Excel AddIn*'s
  - automatic: tokenization, lemmatization, POS …
    - *UIMA*
- searching and statistics
  - visualisation
    - *ANNIS*

OUTLINE

merlin

1. Merlin project
2. Data
3. Annotation
4. Target hypothesis
5. Problems
6. Conclusions

4. TARGET HYPOTHESIS

- (some kind of) reconstructed learner production
  - base for an error identification (annotation)

- MERLIN > 2 target hypothesis
  - FALKO corpus inspiration (minimal x extended TH)

  - TH1: linguistic correctness
    - minimal changes in orthography, morphology, syntax
  - TH2: linguistic appropriateness
    - lexical, semantic, pragmatic aspects

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.
Next semester I will write thesis

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
  - *orthography : grapheme : missing diacritics*
- G_Morphol_case_wrong
  - *Grammar : inflection : case error*

OUTLINE

merlin

1. Merlin project
2. Data
3. Annotation
4. Target hypothesis
5. Problems
6. Conclusions

5. PROBLEMS: CZECH

? target hypothesis

1. Vzala si	*své*	***oblíbenější***	*botičky.*
she took	*her*	*more favourite*	*shoes*
		nejoblíbenější	

2. Myslíš,	*že*	***budeš končit***	*sraz*	*v*	*5*	*hodin?*
you think	*that*	*you will finish*	*meeting (reunion)*	at	5	o´clock?
		? skončíš / ukončíš / končí$_{3sg}$ / skončí$_{3sg}$				

3. *Jsou	má	dovolenu.
*they are	he has	a holiday

?? Jsou na dovolené. X Má dovolenou.

4. *Zvadríš	všechny	a	hlavně	Petra.
you say hallo	(to) everyone	and	especially	(to) Peter

Zdravíš (pozdravuj / pozdravíš)

5. *Sle	fotky?
?send	pictures?

? Pošlu / pošleš/ pošle / šli …

6. ***kratke***	*kalhoty*	7. ***dětí***	*nemluví*	8. ***mužů***	*přijít*
short	*trousers*	*children$_{Gpl.}$*	*do not talk*	*men$_{Gpl}$ (I can)*	*to come*
krátké		*děti*		*můžu*	

? orthography or inflection

9. Chtěl bych	*tě*	*pozvat*	*ke mně*	***doma***
I would like	*you*	*to invite*	*to me*	*home*
				domů

? TH1 (valency) or TH2 (lexical)

10. **kdyby bychom**	*koupili*		10. že	***jsi se***	*učil*
If we (would)	*buy*		*that*	*you were*	*learning*
kdybychom				*ses*	

? orthography or morphology

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

Maybe they talk about what they will do

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

? conjunction *(„korelativum“)* or valency *(mluví/udělají)*

???

Ze cloveku práce ne dost pozitivu vsehno

that a person$_{3sg}$ work$_{1sg}$ not enough ?positive ?everything

Že člověku práce ne dost ?pozitivní ?všechno

bude udělat spatně počasí v jeho důse.

will make badly weather in his ?soul

?udělá špatné počasí v jeho ?duši

OUTLINE

merlin

1. Merlin project
2. Data
3. Annotation
4. Target hypothesis
5. Problems
6. Conclusions

merlin

- one of the only few corpora related to CEFR
  - detection of language features that match learners´proficiency on all reference levels
- covers 3 languages as L2
- selections of annotated features is solidly grounded
- highly controlled annotation is crucial!
  - annotators training
  - documentation
  - annotation check (double annotation and IAA) needed

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!

Thank you for your attention!

Barbora Štindlová
for the MERLIN-Team