Towards learner corpora of English, Finnish and Swedish (+ some other languages) in Finland

Ari Huhta
Centre for Applied Language Studies
University of Jyväskylä, Finland
ari.huhta@jyu.fi

CLARIN L2 Workshop
Göteborg
December 6-8, 2017

Goals

- Learn about annotation, error tagging, etc conventions (or decisions within CLARIN)
- Learn about automated tools for annotation etc
- Share information with other CLARIN members about the learning 'corpora' we have → possible collaboration in the future?
- Share with others our experiences about rating / CEFR-linking of the scripts / performances

Outline / key points

- Research projects & language examinations or tests → learner performances → learner corpora
  - mostly at the University of Jyväskylä
  - mostly **writing** performances / scripts; transcribed & anonymised
  - mostly **unannotated** 'corpora'
  - variation & similarity in the writing tasks across 'corpora'
  - rated against **CEFR** scales (or a version of them) → linguistic characteristics of CEFR levels
  - mostly **teenaged** learners
  - mostly **Finnish-speaking** Finns or **immigrants**
  - other **information about the learners** who wrote the texts

- Fin-Clarin & SLATE collaboration
- Linking with the CEFR levels
- Future plans
- Questions and issues

Where do the learner performances come from?

Research projects at the U. of Jyväskylä

- CEFLING 2007-09
- TOPLING 2010-13
- DIALUKI 2010-13 (and beyond)
- PhD student projects

National Certificates examination corpus (National Agency for Education & U. of Jyväskylä)

National educational evaluation study (Karvi – the Finnish Education Evaluation Centre) 2011-13 study)

CEFLING - Combining Second Language Acquisition and Testing Approaches to Writing

2007-2009 funded by the Academy of Finland

- **writing**, 5 different tasks (messages, narrative, descriptive/argumentative)
  - English word derivation
- **cross-sectional**
- **7th**, 8th, 9th graders (aged 12-15)
- English as FL (Finnish L1), ca. 580 scripts
- Finnish as L2 (different L1s), , ca. 1 200 scripts
- Background questionnaire
- Part of the Finnish L2 corpus has been annotated

CEFLING writing tasks 1
(original task instruction was in Finnish)

Note to a friend

You've set up a meeting with your English-speaking friend at a café. However, something has come up and you have other things to do. Send an email message to your friend.

- Explain why you can't come.
- Suggest a new time and place.

Remember to **begin** and **end** the message in appropriately. Write **in English / Finnish** in clear characters in the space below.

CEFLING writing tasks 3

Message to an internet store

Your parents have ordered a PC game for you from a British internet store. When you get the game you notice that it doesn't work properly. You get upset and decide to write an email message to the internet store. In the message, say

- who you are
- what your parents ordered
- why you're unhappy (mention at least two defects/problems)
- how you would like them to take care of the matter
- give your contact information

Remember to **begin** and **end** the message appropriately. Write **in English / Finnish** in clear characters in the space below.

CEFLING writing tasks 5

Narrative

Tell about the scariest / funniest / greatest experience in your life. Choose one.

- Tell what happened (what, where, when, and so on).
- Tell why the experience was scary / funny / great.

Write **in English / Finnish** in clear characters in the space below (continues on the reverse side).

TOPLING - Paths in Second Language Acquisition

2010-13 funded by the Academy of Finland

- **writing**; similar tasks to CEFLING, 3 times over a period of 2 years
- **longitudinal** and other (e.g. keystroke logging)
- primary, lower secondary, upper secondary (also university)
- English as FL (Finnish L1), ca. 3 400 scripts
- Finnish as L2 (various L1s), ca. 2 550 scripts
- Swedish as L2/FL (Finnish L1), ca. 2 450 scripts

DIALUKI – Diagnosing Reading and Writing in a Second or Foreign Language

2010-13 funded by the Academy of Finland

- **writing** 1-3 tasks, etc (reading, vocabulary, dictation; motivation; cognitive skills; background variables)
- **cross-sectional** (4th, 8th, 11th grade)
- English (Finnish L1), ca. 1 400 scripts (**longitudinal** several hundred)
- Finnish (Russian L1), ca. 400 scipts (longitudinal ca. 200)
- The longitudinal part for English still continues
- Automated analysis tools (Coh-Metrix & L2 Syntactic Complexity Analyzer) used in some sub-studies

The Finnish National Certificates of Language Proficiency

National Agency for Education & U. of Jyväskylä

- High-stakes **language examination** in 9 languages, see http://ykitesti.solki.jyu.fi/en/
- Started in 1994, by now ca. 110 000 test takers
- A portion of speaking and writing performances are stored in the NC corpus, see http://yki-korpus.jyu.fi/
- Can be linked with test takers’ test results and background variables
- Several hundred performances (in most languages)

An example of a PhD student corpus at U. of Jyväskylä

- Ghulam Khushik's PhD study 2015 –
- **writing, cross-sectional** (& longitudinal)
- lower & upper secondary students
- English as FL (Sindhi L1, in Pakistan) & EFL (Finnish L1 from DIALUKI)
- 6 writing tasks from CEFLING and TOPLING
- Automated analysis tools: Coh-Metrix & L2 Syntactic Complexity Analyzer

→ Differences between CEFR levels; characteristics of levels A1 – B1; differences & similarities between Finnish vs Sindhi L1 speakers of English as FL

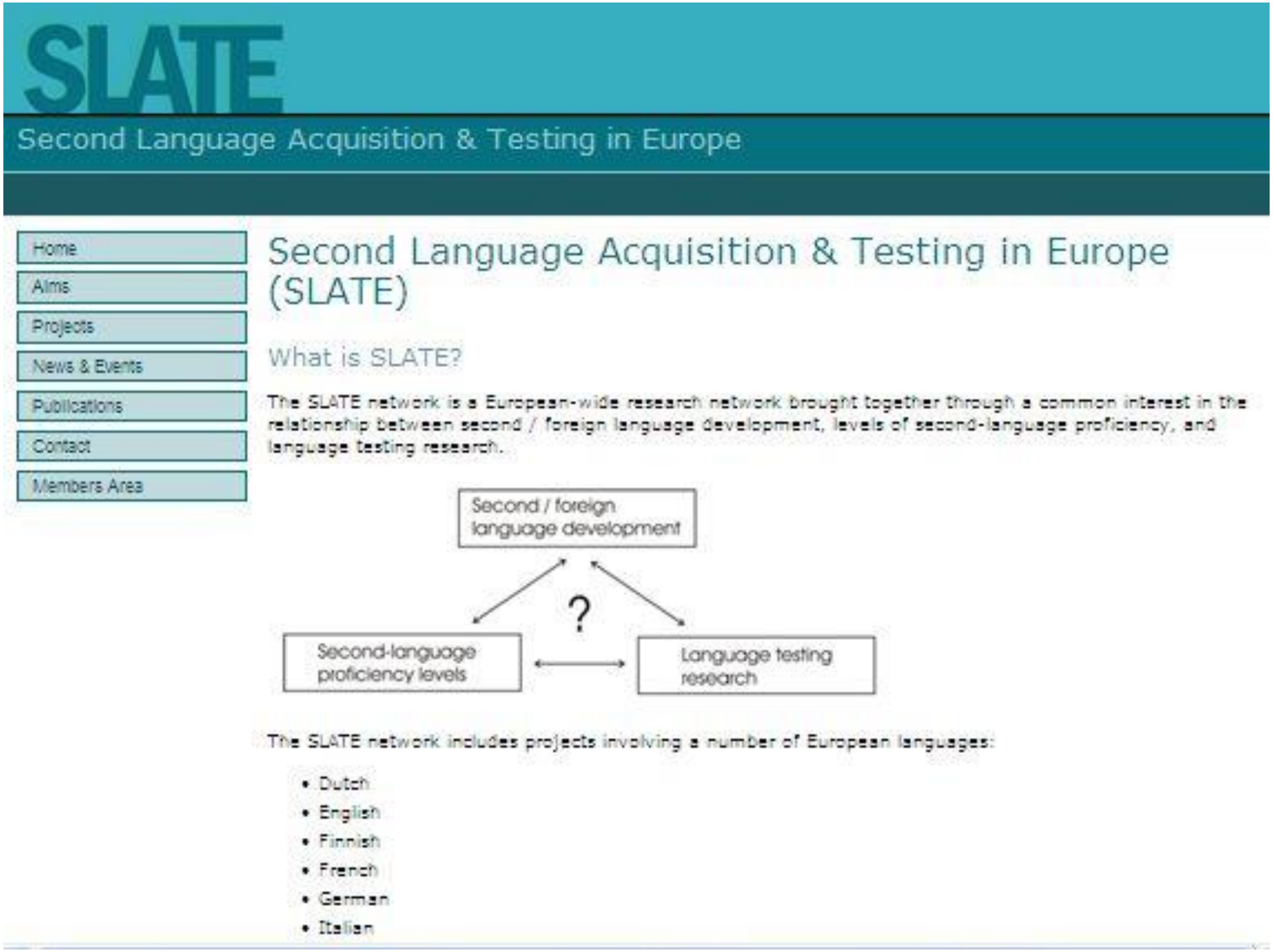
Karvi (Finnish Education Evaluation Centre) corpora

- **National educational evaluation** study 2011-13
- 9th grade (15-year-olds)
- Finnish and Swedish-speaking schools
- English (3500 learners), French (2300), German (2700), Russian (1000), Swedish (1700)
- **writing** (all wrote 2 tasks)
- **speaking** (50-80% completed this)
- we are copying part of the writing and speaking performances for research purposes

FIN-CLARIN & SLATE co-operation

- U. of Jyväskylä is a member of FIN-CLARIN
- SLATE (Second Language Acquisition and Testing in Europe)
  - network of researchers that combine language testing and SLA perspectives
  - Interested in investigating the **linguistic basis of the CEFR levels** (www.slate.eu.org & EuroSLA Monograph series #1; Bartning et al. 2010)
  - CEFLING, TOPLING & DIALUKI relate to the SLATE network

www.slate.eu.org


SLATE
Second Language Acquisition & Testing in Europe
Home
Aims
Projects
News & Events
Publications
Contact
Members Area

Second Language Acquisition & Testing in Europe (SLATE)

What is SLATE?

The SLATE network is a European-wide research network brought together through a common interest in the relationship between second / foreign language development, levels of second-language proficiency, and language testing research.

Second / foreign language development
?
Second-language proficiency levels
Language testing research

The SLATE network includes projects involving a number of European languages:

- Dutch
- English
- Finnish
- French
- German
- Italian

Rating of performances against CEFR (and other) scales

- set of CEFR writing scales used in the CEFLING, TOPLING and DIALUKI projects (Alanen et al. 2010; Alanen et al. 2012; Huhta et al. 2014)
- One of purposes of these projects was to study the linguistic (lexical, grammatical & textual) features that characterise and/or differentiate CEFR levels
- This required that learners' writing performances could be placed at the CEFR levels → comparison of A1, A2, B1 etc in terms of their linguistic features
- Placement at the CEFR levels was done by rating the written scripts

	OVERALL WRITTEN PRODUCTION	WRITTEN INTER-ACTION	CORRESPONDENCE & NOTES, MESSAGES, FORMS	CREATIVE WRITING & THEMATIC DEVELOPMENT & COHERENCE AND COHESION
A1	Can write simple isolated phrases and sentences.	Can ask for or pass on personal details in written form.	Can write a short simple postcard. Can write numbers and dates, own name, nationality, address, age, date of birth or arrival in the country, etc. such as on a hotel registration form.	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do. Can link words or groups of words with very basic linear connectors like 'and' or 'then'.
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.	Can write short, simple formulaic notes relating to matters in areas of immediate need.	Can write very simple personal letters expressing thanks and apology. Can take a short, simple message provided he/she can ask for repetition and reformulation. Can write short, simple notes and messages relating to matters in areas of immediate need.	Can write about everyday aspects of his/her environment, e.g. people, places, a job or study experience in linked sentences. Can write very short, basic descriptions of events, past activities and personal experiences. Can write a series of simple phrases and sentences about their family, living conditions, educational background, present or most recent job. Can write short, simple imaginary biographies and simple poems about people. Can tell a story or describe something in a simple list of points. Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points. Can link groups of words with simple connectors like 'and', 'but' and 'because'.

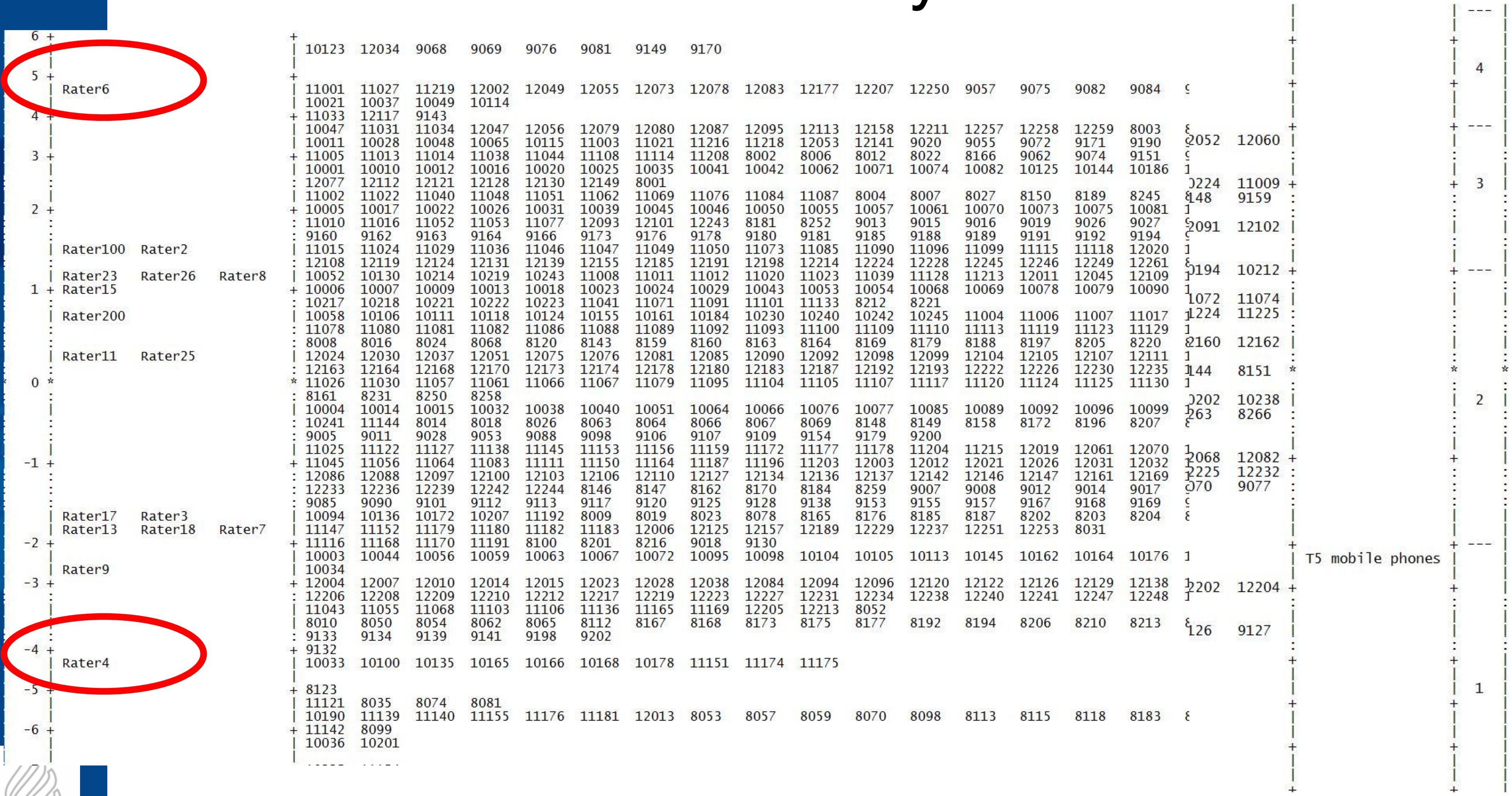
After ratings: analysis of the quality of the ratings

- To investigate how consistently and similarly raters work (intra-rater & inter-rater consistency / reliability) & to increase the dependability of the placements of learners' scripts at the CEFR levels
  - double or multiple ratings required in most cases
- Many approaches to such analyses
  - proportions of agreement, correlations (several different), Cronbach's alpha, Kendal's tau (see e.g. Kaftandjieva & Takala 2000)
  - Multifaceted Rasch analysis (with Facets) (see Huhta et al. 2014)

Facets analyses

- For analysing multiple ratings of learners' performances on one or multiple tasks based on one or two rating scales
- To enable a more reliable / accurate placement of the texts on the CEFR (or other) levels
- We can spot and remove (if we so decide):
  - (too) **inconsistent** raters
  - (too) unexpected / inconsistent ratings (invidual data points)
  - (too) **different** raters in terms of severity / leniency

Facets results – an example of stage 1 of an analysis



```
 6 +                                 +
   |                                 | 10123  12034  9068   9069   9076   9081   9149   9170
   |                                 |
 5 +                                 +
   | Rater6                          | 11001  11027  11219  12002  12049  12055  12073  12078  12083  12177  12207  12250  9057   9075   9082   9084
   |                                 | 10021  10037  10049  10114
 4 +                                 + 11033  12117  9143
   |                                 | 10047  11031  11034  12047  12056  12079  12080  12087  12095  12113  12158  12211  12257  12258  12259  8003
   |                                 | 10011  10028  10048  10065  10115  11003  11021  11216  11218  12053  12141  9020   9055   9072   9171   9190
 3 +                                 + 11005  11013  11014  11038  11044  11108  11114  11208  8002   8006   8012   8022   8166   9062   9074   9151
   |                                 | 10001  10010  10012  10016  10020  10025  10035  10041  10042  10062  10071  10074  10082  10125  10144  10186
   :                                 : 12077  12112  12121  12128  12130  12149  8001
   |                                 | 11002  11022  11040  11048  11051  11062  11069  11076  11084  11087  8004   8007   8027   8150   8189   8245
 2 +                                 + 10005  10017  10022  10026  10031  10039  10045  10046  10050  10055  10057  10061  10070  10073  10075  10081
   :                                 : 11010  11016  11052  11053  11077  12093  12101  12243  8181   8252   9013   9015   9016   9019   9026   9027
   :                                 : 9160   9162   9163   9164   9166   9173   9176   9178   9180   9181   9185   9188   9189   9191   9192   9194
   | Rater100  Rater2                | 11015  11024  11029  11036  11046  11047  11049  11050  11073  11085  11090  11096  11099  11115  11118  12020
   :                                 : 12108  12119  12124  12131  12139  12155  12185  12191  12198  12214  12224  12228  12245  12246  12249  12261
   | Rater23   Rater26   Rater8      | 10052  10130  10214  10219  10243  11008  11011  11012  11020  11023  11039  11128  11213  12011  12045  12109
 1 + Rater15                         + 10006  10007  10009  10013  10018  10023  10024  10029  10043  10053  10054  10068  10069  10078  10079  10090
   :                                 : 10217  10218  10221  10222  10223  11041  11071  11091  11101  11133  8212   8221
   | Rater200                        | 10058  10106  10111  10118  10124  10155  10161  10184  10230  10240  10242  10245  11004  11006  11007  11017
   :                                 : 11078  11080  11081  11082  11086  11088  11089  11092  11093  11100  11109  11110  11113  11119  11123  11129
   :                                 : 8008   8016   8024   8068   8120   8143   8159   8160   8163   8164   8169   8179   8188   8197   8205   8220
   | Rater11   Rater25               | 12024  12030  12037  12051  12075  12076  12081  12085  12090  12092  12098  12099  12104  12105  12107  12111
   :                                 : 12163  12164  12168  12170  12173  12174  12178  12180  12183  12187  12192  12193  12222  12226  12230  12235
 0 *                                 * 11026  11030  11057  11061  11066  11067  11079  11095  11104  11105  11107  11117  11120  11124  11125  11130
   :                                 : 8161   8231   8250   8258
   |                                 | 10004  10014  10015  10032  10038  10040  10051  10064  10066  10076  10077  10085  10089  10092  10096  10099
   :                                 : 10241  11144  8014   8018   8026   8063   8064   8066   8067   8069   8148   8149   8158   8172   8196   8207
   :                                 : 9005   9011   9028   9053   9088   9098   9106   9107   9109   9154   9179   9200
   |                                 | 11025  11122  11127  11138  11145  11153  11156  11159  11172  11177  11178  11204  11215  12019  12061  12070
-1 +                                 + 11045  11056  11064  11083  11111  11150  11164  11187  11196  11203  12003  12012  12021  12026  12031  12032
   :                                 : 12086  12088  12097  12100  12103  12106  12110  12127  12134  12136  12137  12142  12146  12147  12161  12169
   :                                 : 12233  12236  12239  12242  12244  8146   8147   8162   8170   8184   8259   9007   9008   9012   9014   9017
   :                                 : 9085   9090   9101   9112   9113   9117   9120   9125   9128   9138   9153   9155   9157   9167   9168   9169
   | Rater17   Rater3                | 10094  10136  10172  10207  11192  8009   8019   8023   8078   8165   8176   8185   8187   8202   8203   8204
   | Rater13   Rater18   Rater7      | 11147  11152  11179  11180  11182  11183  12006  12125  12157  12189  12229  12237  12251  12253  8031
-2 +                                 + 11116  11168  11170  11191  8100   8201   8216   9018   9130
   |                                 | 10003  10044  10056  10059  10063  10067  10072  10095  10098  10104  10105  10113  10145  10162  10164  10176
   | Rater9                          | 10034
-3 +                                 + 12004  12007  12010  12014  12015  12023  12028  12038  12084  12094  12096  12120  12122  12126  12129  12138
   :                                 : 12206  12208  12209  12210  12212  12217  12219  12223  12227  12231  12234  12238  12240  12241  12247  12248
   |                                 | 11043  11055  11068  11103  11106  11136  11165  11169  12205  12213  8052
   |                                 | 8010   8050   8054   8062   8065   8112   8167   8168   8173   8175   8177   8192   8194   8206   8210   8213
   :                                 : 9133   9134   9139   9141   9198   9202
-4 +                                 + 9132
   | Rater4                          | 10033  10100  10135  10165  10166  10168  10178  11151  11174  11175
   |                                 |
-5 +                                 + 8123
   |                                 | 11121  8035   8074   8081
   |                                 | 10190  11139  11140  11155  11176  11181  12013  8053   8057   8059   8070   8098   8113   8115   8118   8183
-6 +                                 + 11142  8099
   |                                 | 10036  10201
   |                                 |
```

```
             |                      |  ---  |
             |                      |       |
             +                      +       |
             |                      |   4   |
             +                      +       |
             |                      |       |
             +                      +  ---  |
2052  12060  |                      |       |
             :                      :       :
0224  11009  +                      +   3   |
148   9159   :                      :       :
2091  12102  |                      |       |
0194  10212  +                      +  ---  |
1072  11074  |                      |       |
1224  11225  :                      :       :
2160  12162  |                      |       |
144   8151   *                      *       *
0202  10238  |                      |   2   |
263   8266   :                      :       :
2068  12082  +                      +       |
2225  12232  :                      :       :
070   9077   :                      :       :
             +                      +  ---  |
             | T5 mobile phones     |       |
2202  12204  +                      +       |
126   9127   |                      |       |
             +                      +       |
             |                      |   1   |
             +                      +       |
             +                      +       |
             +                      +       |
```

Future plans

- Continue using automated analyses of English learner texts
- Annotate at least some of the ’corpora’: Explore automated annotation tools for English, Swedish & Finnish (possibly other languages, too)
  - Including error tagging
  - In compliance with standards created within CLARIN
- Continue collaboration with Karvi to make the writing and speaking performances collected by available for research

Questions, issues

- Length of performances / scripts
  - very short texts are a problem for investigating individual texts or learners
- Certain kinds of rating issues (severity/leniency)
- Quality / accuracy of automated analyses of learner language
  - Effect of errors / non-standard language on the results?
  - Correction of learner texts: what and how much? (McNamara et al. 2014, p.155-6)
  - Jarvis: correction of misspellings increases accuracy of lexical analyses
  - Identification of sentence boundaries seems to be important (missing sentence final punctuation)

Sample texts (DIALUKI, longitudinal, 6th grade)

TASK: A variant of the ’narrative’ task used in CEFLING (telling about a memorable event from the previous year, e.g., during the summer holiday)

E1016-102

Last year, when I was twelve years old, I visited my frend’s house. I went swim and we fried barbegue. I don’t know why I remembet this moment, it just was fun!

E1026-102

In 2010 I’m going to Solkila with my dad and sister. We are going to watch Solki Rally. There is very awesome cars and many people. In Solkila were very hot day, about +°30. I wear a t-shirt, shorts and sunglasses. My dad was happy, because he pääsi watching cars. My sister was verybored and he doesn’t want come to here.

References

Alanen, R. Huhta, A. & Tarnanen M.. 2010. Designing and assessing L2 writing tasks across CEFR proficiency levels. In Bartning, Inge, Martin, Maisa & Vedder Ineke (eds.) *Communicative proficiency and linguistic development: intersections between SLA and language testing research*. EUROSLA Monograph Series, 1. 21-56. http://eurosla.org/monographs/EM01/EM01home.html

Alanen, R., Huhta, A., Jarvis, S., Martin, M. & Tarnanen, M.. 2012 Issues and challenges in combining SLA research and language testing. In Tsagari, Dina & Csepes, Ildiko (eds.) *Collaboration in Language Testing and Assessment*. (pp.15-30). Language Testing and Evaluation Series, Grotjahn, R. &. G. Sigott (general eds). Frankfurt: Peter Lang. 15-30.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M. & Hirvelä, T. 2014. Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing 31*(3) 307–328.

Kaftandjieva, F. & Takala, S. 2002. Council of Europe scales of language proficiency: A validation study. In CEFR Case Studies, 106-129. https://www.coe.int/t/dg4/linguistic/Source/case_studies_CEF.doc

Khushik, G. & Huhta, A. (submitted). Syntactic complexity features across the Common European Reference Framework levels in the writing of Finnish and Pakistani learners of English.

McNamara, D., Graesser, A., McCarthy, P., & Cai, Z. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: CUP.

CEFLING writing tasks 2

Message to your teacher

You've been away from school for a week. Soon you'll have an English exam. Your teacher, Mary Brown, speaks only English. Send an email message to the teacher.

- Tell her why you've been away.
- Ask two things about the exam.
- Ask two things about the English lessons that were held during the week.

Remember to **begin** and **end** the message appropriately. Write **in English / Finnish** in clear characters in the space below.

CEFLING writing tasks 4

Opinion

Choose one of the topics and write about what you think about the matter. Give reasons for your opinion.

1. Boys and girls should go to different classes at school.
2. No mobile phones at school!

Write **in English / Finnish** in clear characters in the space below (continues on the reverse side).