GÖTEBORGS UNIVERSITET

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

DEVIATION (ERROR) TAXONOMY
AND OTHER CONSIDERATIONS IN THE SWELL PROJECT

CLARIN L2 workshop
Gothenburg, December 7 2017

JULIA PRENTICE, ELENA VOLODINA, DAN ROSÉN, BEÁTA MEGYESI, MATS WIRÉN, GUNLÖG SUNDBERG, LENA GRANSTEDT, MONICA REICHENBERG

SweLL – Research infrastructure for Swedish as a second language

- Larger learner corpora for other languages exist – but not for Swedish
  - Examples:
    - English: International Corpus of Learner English, ICLE (Granger et al. 2002)
    - Norwegian: ASK (Tenfjord et al 2004)
    - German, Czech, Italian : MERLIN (Boyd et al. 2014)

Aims for SweLL

- Build a corpus
  - ca 600 annotated, searchable L2-texts
    - CLARIN-Priv-licens
- Tools for
  - Normalization
  - Annotation
  - Statistical analysis
- Portal
  - data base for uploading texts

Annotation - choosing a taxonomy

- Other taxonomies :
  - ASK : 23 Error types

    Lexical (8), morphological (3), syntactical (7), punctuation (4), unidentified (1)

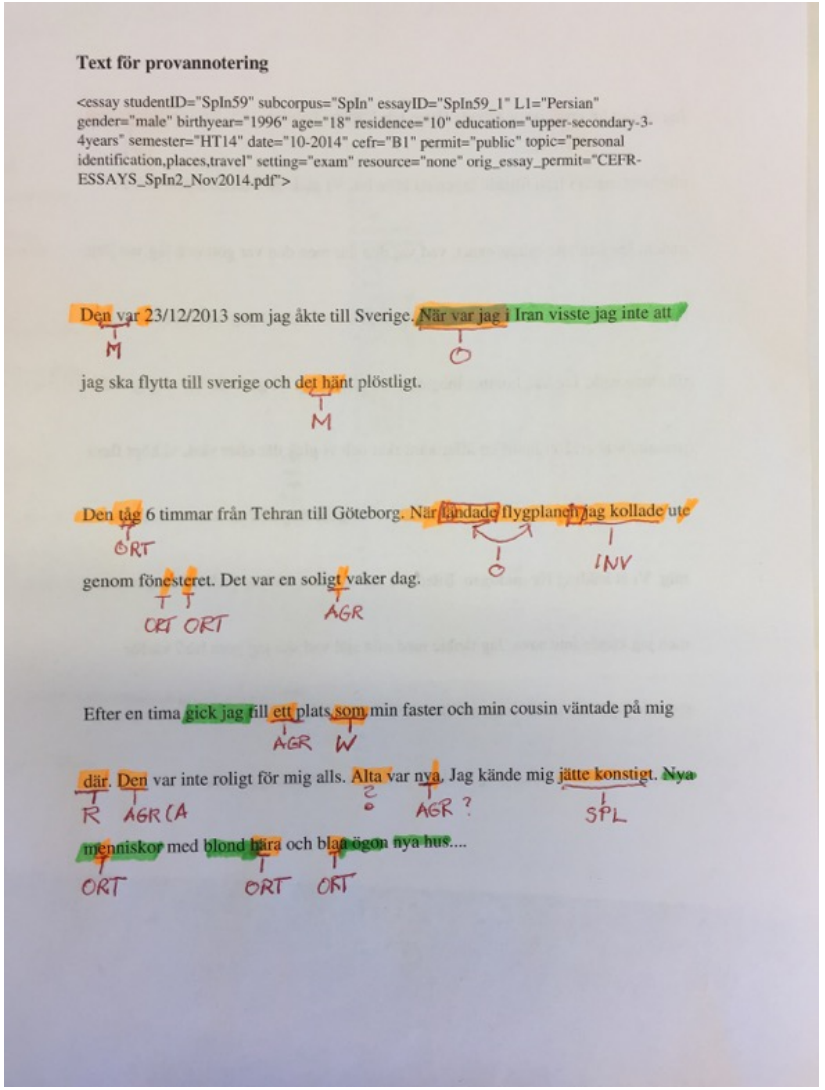
  - MERLIN: 64 error types

    grammar (21), orthografic (8), intelligibility (8), vocabulary (10), coherence (4), sociolinguistic (10), pragmatics (3)

- How detailed should the taxonomy be?
- How important is the target language?
  - similarity between Norwegian and Swedish
  - Comparability between ASK and SweLL wanted

Annotation – target hypothesis and level of normalization

- Pre-pilot annotation experiment
  - manual annotation of a learner text
  - Inter-annotator agreement?
- What is the target hypothesis?
- Example: ’Orthografic error’ or ’wrong word’?
  - jag bara dricker te med två broad
  - tee with two bread
  - [ORT>bröd]
  - [W > mackor]
    - minimal changes in normalization and annotation?



SweLL-taxonomy = ASK+

- First draft:
  - 8 error codes added to the ones taken from ASK
    - Some language specific error types common in Swedish L2-production
    - Partially more fine grained error categories
  - 2 reductions
    - wrong punctuation mark [punc] and missing punctuation mark [PUNCM] are included (as tokens) in categories wrong word [W] and missing word [M] in SweLL

https://spraakbanken.gu.se/eng/swell/swell_codebook

SweLL- additions

Div. type [code]	Example	Explanation	ASK-category
[W-ref]	Min morbror var där. *Hon (She)* -->[W-REF] --> **Han (He)** är snart 60 år	Diviation in reference	[W] Wrong word
idiomaticity [ID]	Jag gick till *sängen* --> [ID] --> **sängs** direkt när jag kom hem (alt. jag gick och la mig)	Idiomaticity Can stretch over groups of words. Nothing is grammatically wrong, but "you don't say so " in Swedish. Covers deviations with respect to idiomatic expressions.	-
definiteness [F-DEF]		Deviation in definite/indefinite forms, may apply to groups of words	[F] Deviant selection of morphosyntactic category
tense [F-TENSE]	Det *var* --> [F-TENSE] --> **är** det bästa jag vet.	Covers all deviations with verbs and verb groups, incl aspect	[F]

SweLL- additions 2

Div. type	Example	Explanation	ASK-category
[F-NUM]	De kan bli *stressad* --> [F-NUM] --> **stressade**	Deviation in number agreement	[F]
[F-AGR]	*ingen problem* --> [F-AGR] --> **inga problem** *en god liv* --> [F-AGR] --> **ett gott liv**	Diviation in agreement,e.g. between adjective and noun; pronoun and noun, etc.	[F]
[M-SUBJ]	Varje tisdag har *[...]* --> [M-SUBJ] -- > **vi** studiebesök.	Subject missing	[M] word or phrase missing
[R-PREP]	*I* --> [R-PREP] --> **[...]** varje fredag har vi idrott	Preposition redundant	[R] word or phrase redundant

Pilot for annotation tool and taxonomy

- 9 texts, different proficiency levels
- 8 annotators from our research group, 3 texts each
  - (1 mandatory, 2 optional/L-2 researchers 2 texts at different levels mandatory, 1 optional)

- 1 week for annotation and a short evaluation report with comments on
  - the time it takes to become proficient in using the tool and error codes
  - the taxonomy
  - the code book including error codes, examples and explanations
  - guidelines for normalization and deviation (error) annotation
  - need fot multiple taget hypotheses
  - use of, and manual for the annotation tool

Some tentative insights from the pilot

- Time needed for training annotators
  - ca 2 days (answers 1,5 – 3 or 10 texts)
    - easy to use the tool but it takes time to get used to and remember the codes and to actually be able to apply them to different kinds of texts
- Taxonomy
  - more work to make categories as mutually exclusive as possible needed
    - e.g. [ID] not easy to delimite in relation to other categories like [W]
  - categories should as much as possible pertain to the same level of abstraction/specificity
    - Common (frequent) deviation types in Swedish L2-production should have their own category though
  - Some additions might not be necessary

  - Need for multiple TH?
    - Difficult to say at this point – but in some cases it can be important to show that there are alternative ways of interpretation/normalization

Ongoing interdisciplinary discussion – whats wrong with ”error” annotation?

- *Error taxonomies*, *error annotation*
  - Established terms within LCR
  - problematic terms within SLA
- What is the problem?
  - Beginning of SLA:
    - Contrastive analysis: comparison of learners' L1 and L2 to predict difficulties and errors
  - SLA today :
    - Learner language should be studied as a developing system in its own right (cf Selinker 1972) - not a deficient version of the target language (jfr Granger 2009)
    - Bilingual turn: The (monolingual) native speaker as the norm is being questioned (Ortega 2009)
- What to say instead?
  - Norm deviation taxonomy…?
  - Learner phenomenon taxonomy?
  - Interlanguage taxonomy?
    - covers both features that deviate from the TL-norm and those that agree with it – interlanguage development

Can-do taxonomy?

- A can-do taxonomy for Swedish learner language?
  - Annotating ”success” in L2-production
- Being able to study developmental features in L2-texts
  - Criteria/can-do-categories based on Processability theory? See Flyman Mattsson & Håkansson (2010)
    - Model based on stages in grammatical development

    - How can developmental features that do not fully agree with the target language norm be handled within such a taxonomy?
    - What else, aside from grammatical features, do we want to/can we include?
      - e.g. lexicon, constructions?

Towards interlanguage annotation?

- Díaz-Negrillo, Meurers , Valera & Wunsch (2009:2)
  - ”In sum, SLA research essentially observes correlations of linguistic properties, whether erroneous or not. In consequence, learner corpora should ideally provide annotation of linguistic properties, including but not limited to errors.” (ibid., cf. t.ex. Pienemann, 1998)

- Interlanguage annotation (ILA), Díaz-Negrillo & Lozano (2013:65):

  ” We build on common annotation practices in learner corpora. But we argue for a type of annotation that can disclose a wider picture of specific features of learner’s interlanguage, that is, tagging that (i) is purpose-oriented, (ii) is fine-grained and (iii) describes not just learners’ subtle errors but also their correct uses.”

- For SweLL
  - norm deviation + can-do-annotation (?)

Tack!

Thank you!

https://spraakbanken.gu.se/swe/swell_infra

Referenser

Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová and Chiara Vettori. *The MERLIN corpus: Learner Language and the CEFR* . Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, May 26-31, 2014.

Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. I: Aijmer, Karin (red.), *Corpora and Language Teaching.* Amterdam & Philadelphia: John Benjamins*, 13-32.*

Díaz-Negrillo, A. & Lozano, C. (2013). Using learner corpus tools in SLA research: the morpheme order studies revisited', Paper presented at Corpus Linguistics 2013, University of Lancaster (UK).

Díaz-Negrillo, A. ,D. Meurers, S.Valera & H. Wunsch (2009). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum, Vol. 36, No 1-2. 139-154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair, edited by María Moreno Jaén and Carmen Pérez Basanta. 2010.*

Språkbanken 2017. <https://spraakbanken.gu.se/swe>

Tenfjord, Kari, Meurers, P. & Hofland, K. 2004. The ASK corpus - a language learner corpus of Norwegian as a second language. Paper presented at the TALC 2004 conference, Granada - Spain, 6-9 July 2004.

Volodina, Elena & Lars Borin. 2012. Developing a freely available web-based exercise generator for Swedish. *EuroCALL 2012 Proceedings*, Gothenburg.

Volodina, Elena, Beata Megyesi, Mats Wirén, Lena Granstedt, Julia Prentice, Monica Reichenberg, Gunlög Sundberg. 2016. A Friend in Need? Research agenda for electronic Second Language infrastructure. *Proceedings of SLTC 2016*, Umeå, Sweden

Díaz-Negrillo & Lozano (2013:65):

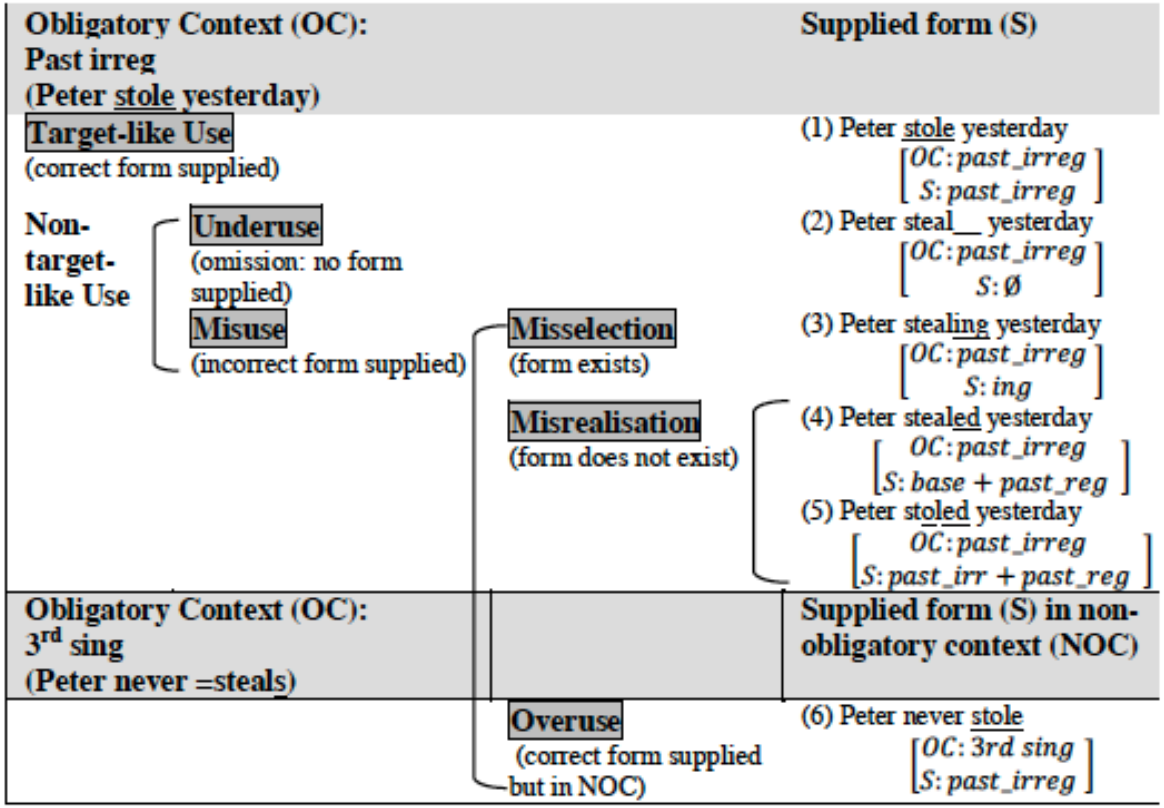


Figure 1: Tagset for irregular past