GÖTEBORGS UNIVERSITET

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

SWELL – FORSKNINGSINFRASTRUKTUR FÖR SVENSKA SOM ANDRASPRÅK

SVENSKANS BESKRIVNING 36, UPPSALA UNIVERSITET, 25-27 OKTOBER 2017

JULIA PRENTICE & ELENA VOLODINA, INSTITUTIONEN FÖR SVENSKA SPRÅKET,
GÖTEBORGS UNIVERSITET

Andraspråksforskning – interdisciplinärt från början och än idag

- Kopplingar till
  - Lingvistik
  - Förstaspråksinlärning
  - Språkdidaktik
  - kognitiv psykologi

  - Idag även data- och korpuslingvistik

Learner corpus research

- Kärnkomponenter
  - Korpuslingvistik
  - Lingvistisk teori
  - Språkdidaktik
  - Andraspråksforskning

- Kopplingarna mellan LCR och andraspråksforskning inte så starka och tydliga som man skulle kunna tro
  - många forskare inom LCR är snarare korpuslingvister och språklärare än andraspråksforskare (Granger 2009)
- Men LCR och SLA integreras allt mer

En elektronisk forskningsinfrastruktur

- (fri tillgänglig) data i elektronisk format
- teknisk plattform för att utforska data, inkl. verktyg och algoritmer för analysen och visualisering av datan
- en verktygslåda med tekniska lösningar för insamling och bearbetning av ny data (inkl bl.a. automatisk annotering)
- ett nätverk med experter inom relevanta områden, inkl. bl.a. juridiska och etiska frågor

SweLL – forskningsinfrastruktur för svenska som andraspråk (Riksbankens jubileumsfond 2017-2019)

- Projektets syfte:
  - Att utveckla en elektronisk forskningsinfrastruktur för svensk andraspråksforskning
  - bearbeta andraspråksmaterial
  - utveckla språkteknologiska metoder som kan hantera avvikelser från standardsvenska
  - Större, sökbar inlärarkorpus för svenska (Volodina et al. 2016)
- Det finns idag stora inlärarkorpusar för andra språk men inte för svenska
  - Några exempel
    - Norska: ASK (Tenfjord et al 2004)
    - Tyska, tjeckiska, italienska : MERLIN (Boyd et al. 2014)
    - Engelska: International Corpus of Learner English, ICLE (Granger et al. 2002)

Utmaningar i samband med korpusinfrastruktur för L2-data (Volodina et al. 2016)

- Tillgång till och insamling av inlärartexter
  - måste samlas in direkt ifrån och med tillstånd från textförfattare som skriver på sitt L2
  - tillstånd av vårdnadshavare behövs för yngre barn
  - känslig information i texter och metadata kan förekomma, anonymisering och juridisk rådgivning behövs

- För lite dialog och samarbete mellan olika fält (jfr Granger 2009)
  - Olika projekt inom olika fält med olika utgångspunkter, syften och metoder
  - Brist på jämförbarhet av data från olika projekt

- Annoteringsverktyg är utvecklade för standardsvenska
  - svårt att hantera avvikelser
  - Annotering av sådant material är komplicerat och tar tid
  - texter behöver bl.a. ”normaliseras”

Mål för SweLL

- Bygga en korpus
  - ca 600 annoterade L2-texter
    - CLARIN-Priv-licens
  - Sökverktyg

- Analysverktyg
  - Normalisering
  - Annotering
  - Statistik
- Portal
  - databas för uppladdning av texter

Korpusdesign

- ”It is important to bear in mind, however, that for the SLA [second language acquisition] specialist big is not nessecarily beautiful” (Granger 2009:4)
  - Data- och korpuslingvister jobbar i huvudsak kvantitativt – poängen är tillgång till stora mängder data
  - Andraspråksforskare behöver kunna kontrollera för många bakgrundsfaktorer som kan påverka inlärares språkliga produktion
    - Språklig bakgrund – vilka L1 ska representeras i korpusen?
    - Andra viktiga faktorer som startålder för inlärning, övriga språkkunskaper, utbildningsbakgrund, tid i landet, tillgång till undervisning, tillgång till annan L2-input och majoritetsspråkskontext…
  - Det ställer höga krav på t.ex. korpusdesign och tillgång till metadata (ibid.)

- Utveckling av forskningsinfrastruktur för, och med tydlig teoretisk förankring i, andraspråksforskning

Annotering och val av taxonomi

- Toxonomier för andra korpusar:
  - ASK : 23 feltyper

    Lexikon (8), morfologi (3), syntax (7), interpunktion (4), oidentifierbart(1)

  - MERLIN: 64 feltyper

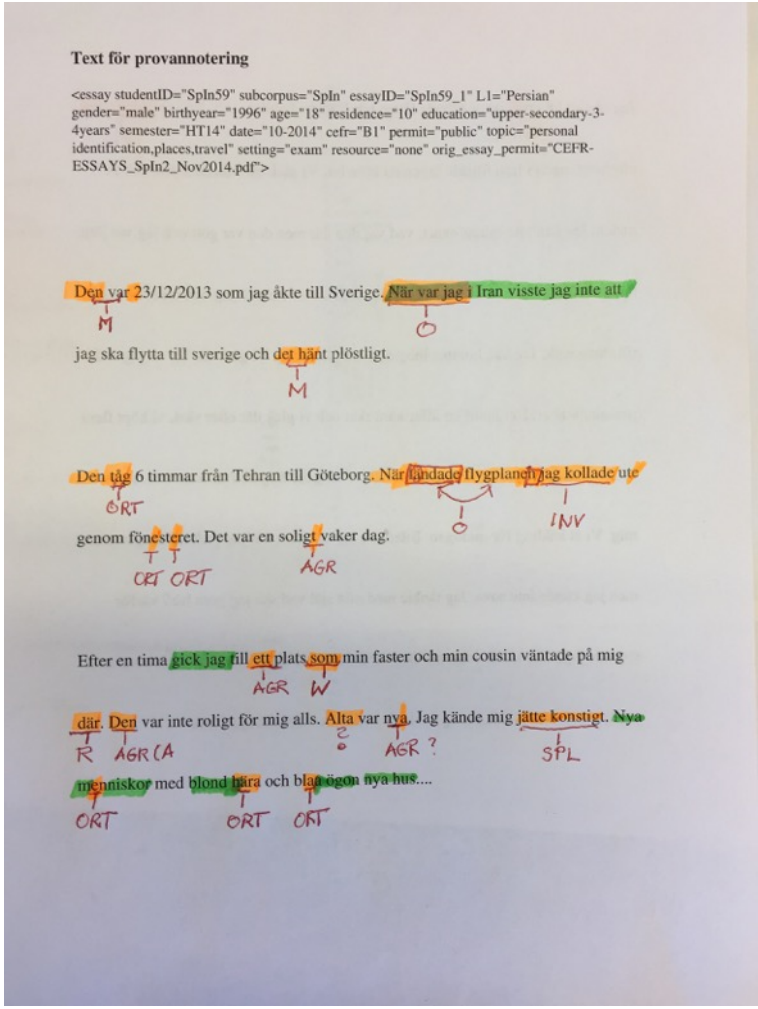
    grammatik (21), ortografi (8), begriplighet (8), vokabulär (10), koherens (4), sociolingvistisk lämplighet, (10), pragmatik (3)

- Hur detaljerad behöver taxonomin vara?
- Hur stor roll spelar de enstaka språken här
  - likhet mellan norska och svenska en faktor
  - jämförbarhet mellan ASK och SweLL önskvärd

- ASK + (språkspecifika vanliga feltyper i svenska som andraspråk)

Annotering och val av taxonomi

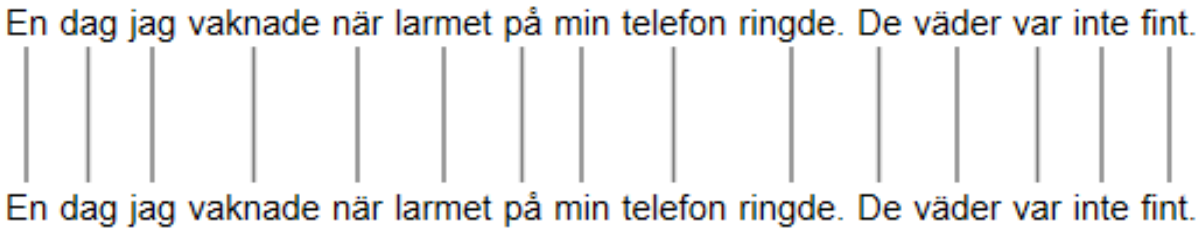
- Annoteringsexperiment
  - manuell annotering av inlärartext
  - Inter- annotator agreement?
- Vad är “målhypotesen”?
- Exempel: ’Orthografic error’ eller ’wrong word’?
- jag bara dricker te med två <u>broad [>bröd/skivor bröd/mackor?]</u>
  - minimala ändringar vid normalisering och annotering?



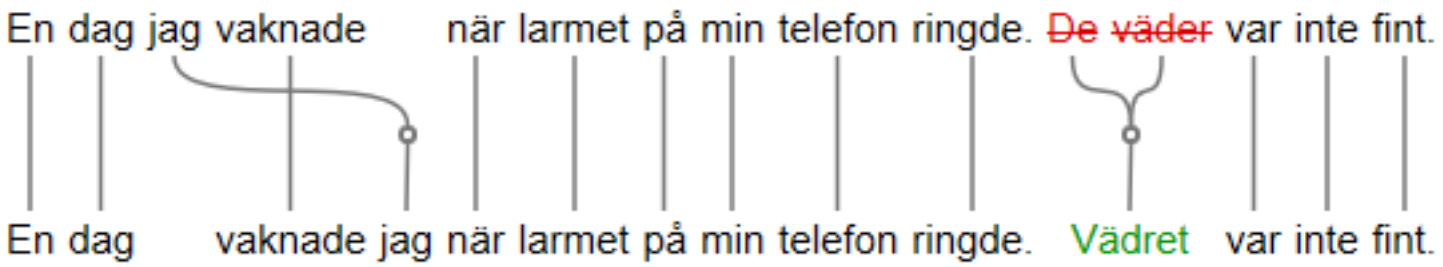
Annoteringsverktyg
(Dan Rosén, Mats Wirén)

- Steg 1: ”Normalisering” (målhypoteser)

Alignment of source text and normalised text



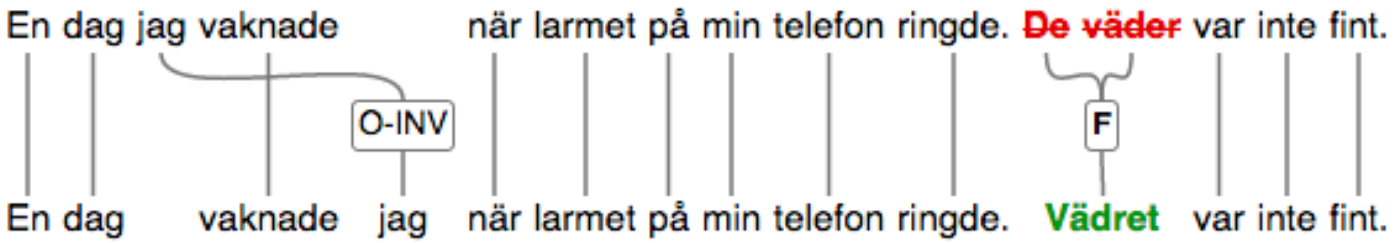
Alignment of source text and normalised text



Annoteringsverktyg 2

- Steg 2: Avvikelseannotering

Alignment of source text and normalised text



I	
W	wrong word
ORT	orthographic error
PART	overcompounding
SPL	oversplitting
DER	deviant derivational affix used
FL	Non-Norwegian word
F	deviant selection of morphosyntactic category
CAP	deviant letter case (upper/lower)
PUNC	wrong selection of punctuation mark

Demo

http://demo.spraakdata.gu.se/dan/swell-editor/

Interdisciplinär kommunikation – felet med felannotering

- Inom LCR talar man om feltaxonomier och felannotering (*error taxonomies*, *error annotation*)
  - problematiskt inom andraspråksforskningen
- Vad ska man säga istället?
  - Normavvikelseannotering…?
  - Inlärarfenomen?
  - Interimspråksfenomen?
    - både avvikande drag och sådana som visar på en utveckling mot målspråksnormen
- Vad är problemet?
  - Andraspråksforskningens förflutna
    - Kontrastiv analys: jämförelse av inlärares L1 och L2 för att förutsäga svårigheter och fel
  - Idag : Interimspråket (Selinker 1972) ska studeras i sin egen rätt – ett system under utveckling – inte en bristfällig version av målspråket (jfr Granger 2009)

Can-do taxonomy?

- Utveckling av en Can-do taxonomi för svensk inlärarspråk?
  - Annotera ”succé” i inlärares produktion

- Synliggöra språklig utveckling i inlärartexter
  - Kriterier/can-do-kategorier utifrån processbarhetsteorins utvecklingsstadier?
    - Se Flyman Mattsson & Håkansson (2010)
      - Analysmodell baserad på grammatiska utvecklingsstadier
    - Hur kan utvecklingsdrag som inte stämmer överens med målspråksnormen hanteras inom en sådan taxonomi?
    - Vad behövs utöver grammatiska drag?
      - lexikon, konstruktioner?

Mot interimsspråksannotering

- Díaz-Negrillo , Meurers , Valera & Wunsch (2009:2)
  - Andraspråksforskningen är bl.a. intresserad av olika utvecklingsstadier, som inte nödvändigtvis manifesteras genom målspråksenligt användning av de strukturer som ses som indikatorer för utveckling till nästa stadium (jfr. t.ex. Pienemann, 1998)

”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.)

- 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.”
  - Tar avstamp i de s.k. engelska morfemstudierna – MOS (jfr Dulay, Burt & Krashen 1982, DeKKeyser 2001)
- För SweLL
  - normavvikelse (ASK+) + can-do-annotering

Tack!

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):

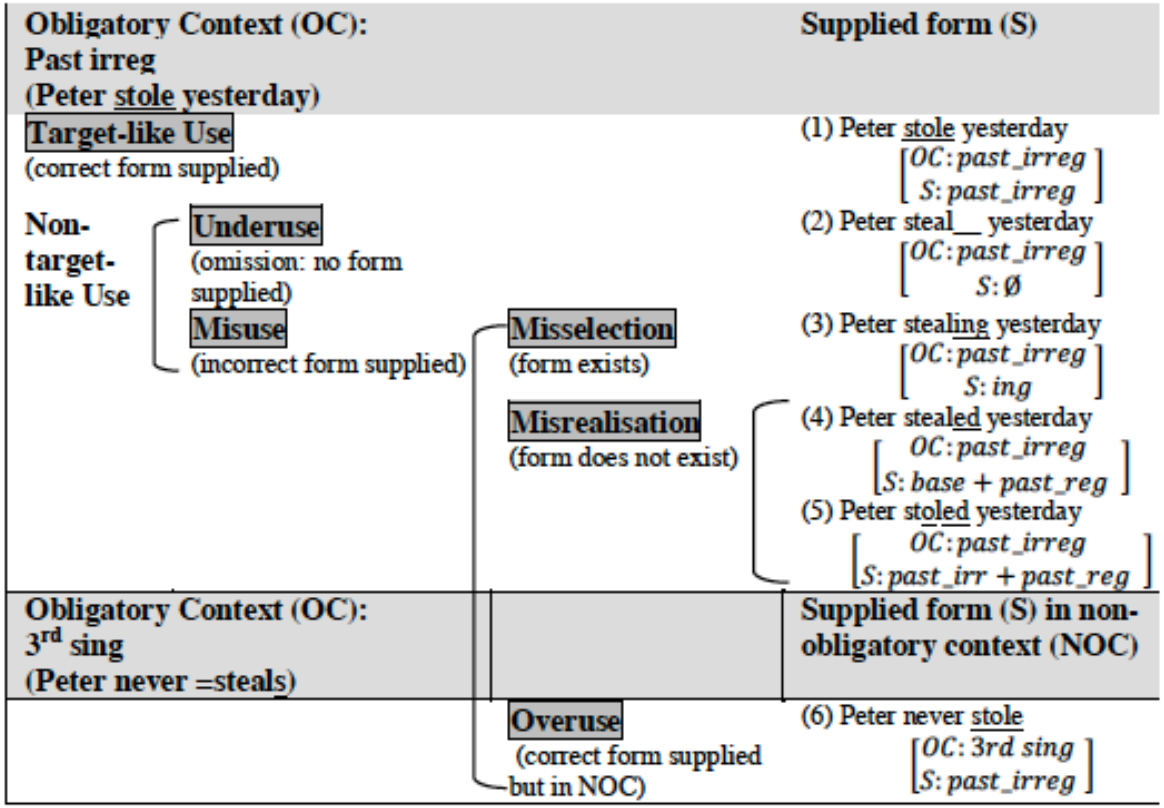


Figure 1: Tagset for irregular past