

# Non-Native Writers' Errors – a Challenge to a Spell-Checker

1st Nordic workshop on evaluation of spellchecking and proofing tools

Björn Hammarberg and Gintarė Grigonytė

[ham@ling.su.se](mailto:ham@ling.su.se), [gintare@ling.su.se](mailto:gintare@ling.su.se)

Department of Linguistics, Stockholm University

# Motivation

Granska har granskat texten.

Jag har gjordade alla läxor .

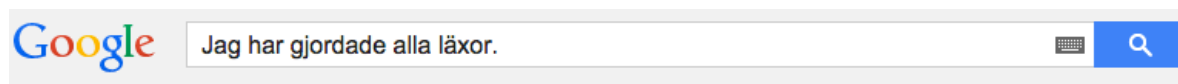
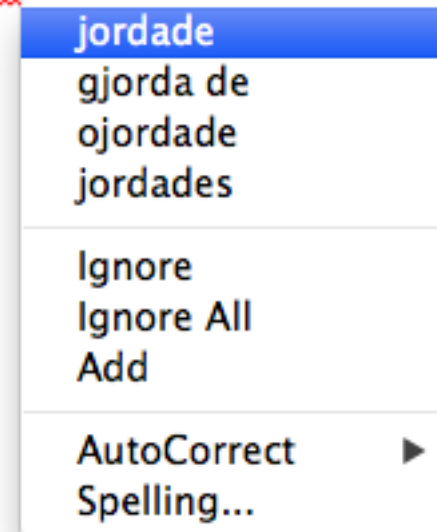
**gjordade** Misstänkt stavfel (stav1@stavning)

Förslag 1. gjordande

2. jordade

3. ojordade

Jag har gjordade alla läxor.



[Webben](#) [Bilder](#) [Videor](#) [Kartor](#) [Nyheter](#) [Fler ▾](#) [Sökverktyg](#)

10 resultat (0,50 sekunder)

Menade du: Jag har **gjort** alla läxor.

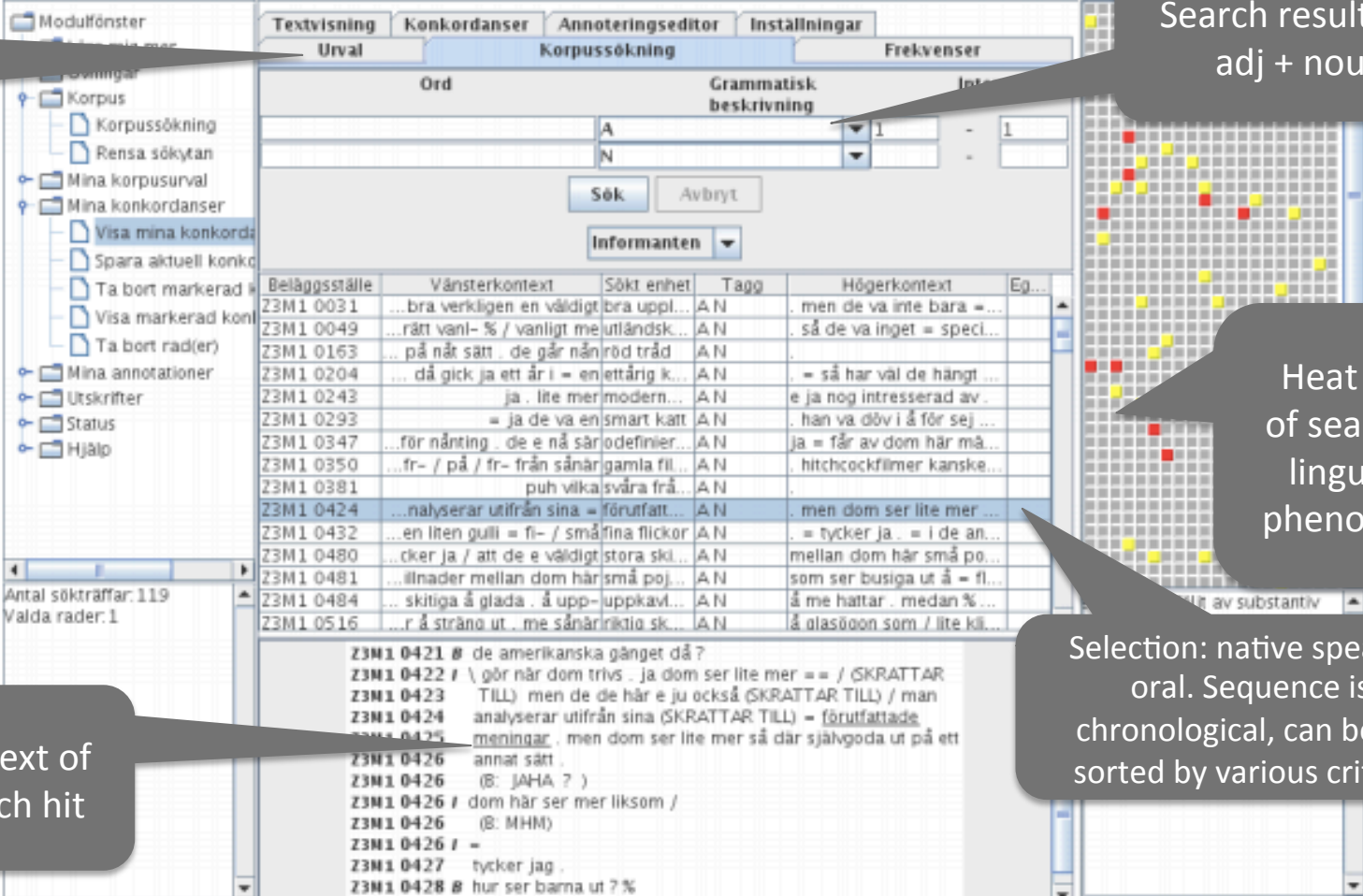
## The ASU corpus

|                  | <i>Learner</i>  | <i>Native</i>   | <i>Total Lnr+Nat</i>      |
|------------------|---|---|---------------------------|
| <b>Oral</b>      | 10 persons x 10 sessions<br>= 100 text units,<br>ca 269,000 / 147,000 word<br>tokens <sup>1</sup> | 7 persons x 5 sessions<br>= 35 text units,<br>ca 149,000 / 98,000 word<br>tokens <sup>1</sup> | ca 418,000 word<br>tokens |
| <b>Written</b>   | 10 persons x 11 sessions<br>x 2 texts = 220 text units,<br>ca 50,000 word tokens                  | 7 persons x 5 sessions<br>x 2 texts = 70 text units,<br>ca 25,000 word tokens                 | ca 75,000 word tokens     |
| <b>Total ASU</b> |   |   | ca 493,000 word<br>tokens |

Hammarberg (2010a, 2010b)

The ASU corpus is available through the ITG interface at  
<http://spraakbanken.gu.se/swe/itg>

# The ITG interface on Språkbanken



The search mode

Search results for adj + noun

Heat map of searched linguistic phenomena

The context of the search hit

Selection: native speaker, oral. Sequence is chronological, can be re-sorted by various criteria

| Beläggsställe | Vänsterkontext                | Sökt enhet    | Tagg | Högerkontext                  | Eg... |
|---------------|-------------------------------|---------------|------|-------------------------------|-------|
| Z3M1 0031     | ...bra verkligen en väldigt   | bra uppl...   | AN   | . men de va inte bara = ...   |       |
| Z3M1 0049     | ...rätt vani- % / vanligt me  | utländsk...   | AN   | . så de va inget = speci...   |       |
| Z3M1 0163     | ... på nåt sätt . de går nån  | röd tråd      | AN   | .                             |       |
| Z3M1 0204     | ... då gick ja ett år i = en  | lettårig k... | AN   | . = så har väl de hängt ...   |       |
| Z3M1 0243     | ja . lite mer                 | modern...     | AN   | e ja nog intresserad av ...   |       |
| Z3M1 0293     | = ja de va en                 | smart katt    | AN   | . han va dövd i å för sej ... |       |
| Z3M1 0347     | ...för nånting . de e nå sär  | odefinier...  | AN   | ja = får av dom här må...     |       |
| Z3M1 0350     | ...fr- / på / fr- från sänär  | gamla fil...  | AN   | . hitchcockfilmer kanske...   |       |
| Z3M1 0381     | puh vilka                     | svåra frå...  | AN   | .                             |       |
| Z3M1 0424     | ...analyserar utifrån sina =  | förutfatt...  | AN   | . men dom ser lite mer ...    |       |
| Z3M1 0432     | ...en liten gulli = fi- / små | fina flickor  | AN   | . = tycker ja . = i de an...  |       |
| Z3M1 0480     | ...cker ja / att de e väldigt | stora ski...  | AN   | mellan dom här små po...      |       |
| Z3M1 0481     | ...ilnader mellan dom här     | små poj...    | AN   | som ser busiga ut å = fi...   |       |
| Z3M1 0484     | ...skitiga å glada . å upp-   | uppkavl...    | AN   | å me hattar . medan % ...     |       |
| Z3M1 0516     | ...r å stråno ut . me sänär   | riktio sk...  | AN   | å glasöon som / lite kll...   |       |

Antal sökträffar: 119  
Valda rader: 1

Z3M1 0421 # de amerikanska gänget då ?  
Z3M1 0422 / \ gör när dom trivs . ja dom ser lite mer == / (SKRATTAR  
Z3M1 0423 TILL) men de de här e ju också (SKRATTAR TILL) / man  
Z3M1 0424 analyserar utifrån sina (SKRATTAR TILL) = förutfattade  
Z3M1 0425 meningar . men dom ser lite mer så där självgoda ut på ett  
Z3M1 0426 annat sätt .  
Z3M1 0426 (B: JAHA ? )  
Z3M1 0426 / dom här ser mer liksom /  
Z3M1 0426 (B: MHM)  
Z3M1 0426 / =  
Z3M1 0427 tycker jag .  
Z3M1 0428 # hur ser barna ut ? %

## Spell checker prototype

- Grigonyte and Hammarberg (2014)
- Uses:
  - SALDO, Borin et al.(2008)
  - SLME, Östling (2012)

| Original word                                  | Intended word                       | Probability in given context |
|--|-------------------------------------|------------------------------|
| försöka att förstå hela <b>menningen</b>       | meningen<br>( <i>En. sentence</i> ) | 5.06e-05                     |
| ( <i>En. try to understand all senntence</i> ) | mynningnen<br>( <i>En. mouth</i> )  | 4.71e-09                     |

## Type 1. Lexical shape of words

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| motsidig      | ömsesidig     | motstridig               |
| oilen         | oljan         | ollen                    |

Hypothetical word for an intended concept with incorrect result

- 1) lexical form made up hypothetically on the basis of the German form *gegenseitig*
- 2) word formed on the basis of English *oil* and the Swedish definite suffix *-en*

## Type 2. Morphology

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| sittade       | satt          | siktade                  |
| feodaliskt    | feodalt       | feodalism                |

Inflection errors and word formation errors

- 1) wrong conjugation leading to the wrong past tense form
- 2) adjective forming suffix *-isk* added to the adjective *feodal*

## Type 3. Phonology-phonetics

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| flygan        | flugan        | flyga                    |
| hella         | hela          | hylla                    |

Difficulty of discriminating phonetically between the sounds affecting phonemic distinctions

- 1) Vowels y and u
- 2) Short and long consonant



# IPA table vowels

| Swedish orthography | Phonemes (Swed-ortho notation) | Phonemes (IPA notation) | Major realizations (IPA) | Common international spelling of corresponding sounds | Comments                     |
|---------------------|--------------------------------|-------------------------|--------------------------|---|------------------------------|
| o                   | /o:/                           | /u:/                    | [u:]                     | u   |                              |
|                     | /å/                            | /o/                     | [ɔ]                      | o   |                              |
|                     | (/å:/                          | /o:/                    | [o:]                     | o   | <i>Limited distribution)</i> |
|                     | (/o/                           | /u/                     | [ɔ]                      | u   | <i>Limited distribution)</i> |
| u                   | /u:/                           | /ɯ:/                    | [ɯ:]                     | - (y, ü)  | [ɯ:] typolog. unusual        |
|                     | /u/                            | /ɯ/                     | [ə]                      | -   | [ə] typolog. unusual         |
| y                   | /y:/                           | /y:/                    | [y:]                     | y, ü  |                              |
|                     | /y/                            | /y/                     | [ɻ]                      | y, ü  |                              |
| å                   | /å:/                           | /o:/                    | [o:]                     | o   |                              |
|                     | /å/                            | /o/                     | [ɔ]                      | o   |                              |
| ä                   | /ä:/                           | /ɛ:/                    | [ɛ:], [æ]                | e   | Low before /r/               |
|                     | /ä/                            | /ɛ/                     | [ɛ], [æ]                 | e   | Low before /r/               |

# IPA table consonants

| Swedish orthography                   | Phonemes (Swed-ortho notation) | Phonemes (IPA notation) | Major realizations (IPA) | Common international spelling of corresponding sounds | Comments                           |
|---------------------------------------|--------------------------------|-------------------------|--------------------------|---|------------------------------------|
| c                                     | /s/                            | /s/                     | [s]                      | c, s  |                                    |
| g                                     | /g/ /j/                        | /g/ /j/                 | [g] [j]                  | g, j  |                                    |
| j                                     | /j/                            | /j/                     | [j]                      | j, y  |                                    |
| k                                     | /k/                            | /k/                     | [k]                      | k, c  |                                    |
| ng, n, g                              | /ng/                           | /ŋ/                     | [ŋ]                      | ng  | n before /g/, /k/;<br>g before /n/ |
| sj, sk, stj, skj, sch, ch <i>etc.</i> | /sj/                           | /ʃ/ (/f/)               | [ʃ], [s]                 | sh, sch, ch   | Great spelling variation for /ʃ/   |
| tj, kj, k, ch <i>etc.</i>             | /tj/                           | /tʃ/                    | [tʃ]                     | (ch)  | Great spelling variation for /tʃ/  |

## Type 4. Orthographical norms

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| lexion        | lektion       | legion                   |
| sommna        | somna         | somna                    |

Reasonable but not occurring spellings that do not deviate from Swedish spelling conventions, simply words are not written this way

- 1) *x* is not used in this word, unlike *reflexion/reflektion*
- 2) *o* is short in the morpheme *somn-*, *double m is not required, as already two consonants are present (m and n)*

## Type 5. Code switching

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| fashion       | fashion       | fusion                   |
| exciting      | exciting      | excitering               |

Word are not intended to be Swedish in the first place, but are temporary switches from Swedish into another language, in this case English

## Type 6. Random spelling or typing lapses

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| utblidningen  | utbildningen  | utbildningen             |
| envenemanger  | evenemang     | evenemangen              |

Spell-checking errors by definition

- 1) Character swap
- 2) Insertion of characters

## Type 7. Multiple errors

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| skylnat       | skillnad      | skyltat                  |
| forberedat    | förberett     | förbereda                |

Words can contain several errors

- 1) Phonological vowel distinction y-l + single-double consonant + phonological consonant distinction t-d
- 2) o-ö distinction + morphological error in verb conjugation

## Type 8. Strongly deviant words

| Produced form | Intended word | Spell-checker correction |
|---------------|---------------|--------------------------|
| verstao       | förstår       | värsta                   |
| showte        | duschade      | shorts                   |

Words that are difficult to recognize even for a human reader. A context may help.

- 1) ? Hypothetically: v=f, ver=för (in German phonology)
- 2) ? Hypothetically: shower (en.) root + past tense conjugation

## Discussion

- **Distinction between languages**, (Lui et al., 2014)
- **Morphological segmenters** (Grönroos et al., 2014) could cover the following issues:
  - verb conjugation (e.g. *sittade* can be discovered by recognizing the word-final morpheme *-ade*)
  - noun and adjective inflection
  - “double consonant” issue
- **Phonological errors** like failure of distinguishing separate phonemes or segment quantity i.e., long vs. short, and orthographic conventions i.e. like the doubling of the consonants after the stressed short vowel, can be solved to some extent by learning “rules”, for instance e.g.: u-y, o-u.



## References

- L. Borin, M. Forsberg and L. Linngren.** 2008. The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. *Studia Linguistica Upsaliensia*, p.21–32.
- G. Grigonytė and B. Hammarberg.** 2014. Pronunciation and Spelling: The Case of Misspellings in Swedish L2 Written Essays, In proceedings of the 6<sup>th</sup> Baltic HLT conference, IOS Press. p. 95-98.
- Grönroos, S.-A., Virpioja, S., Smit, P., and Kurimo, M.** 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In proceedings of the 25th International Conference on Computational Linguistics, p. 1177-1185.

## References

- B. Hammarberg.** 2010. The ASU corpus. <http://spraakbanken.gu.se/swe/itg>
- B. Hammarberg.** 2010. Introduction to the ASU Corpus, available from:  
[www.ling.su.se/polopoly\\_fs/1.13705.1320939577!/Introduction\\_to\\_the\\_ASU\\_Corpus.pdf](http://www.ling.su.se/polopoly_fs/1.13705.1320939577!/Introduction_to_the_ASU_Corpus.pdf)
- M. Lui, J.H. Lau, T. Baldwin.** 2014. Automatic Detection and Language Identification of Multilingual Documents. *TACL 2*: 27-40.
- R.Östling,** SLME tool inspired by the Stupid Backoff Model (Brants et al., 2007), (2012), <http://www.ling.su.se/english/nlp/tools/slme>