Automatic profiling of upper-level ESL learner texts

(Based on data from BAWE and MELD)
Background

- Granger and Rayson (1998) compare part-of-speech (POS) frequency in native English speaker (Louvain Corpus of Native English Essays (LOCNESS) database) and learner corpora (International Corpus of Learner English (ICLE) database).
Background

- Granger and Rayson (1998) compare part-of-speech (POS) frequency in native English speaker (Louvain Corpus of Native English Essays (LOCNESS) database) and learner corpora (International Corpus of Learner English (ICLE) database).
- Their major finding is that EFL learners’ (French mother tongue) written language reveals many features characteristic of spoken language and practically none of those of academic, written English, e.g.:
Granger and Rayson (1998) compare part-of-speech (POS) frequency in native English speaker (Louvain Corpus of Native English Essays (LOCNESS) database) and learner corpora (International Corpus of Learner English (ICLE) database).

- Their major finding is that EFL learners’ (French mother tongue) written language reveals many features characteristic of spoken language and practically none of those of academic, written English, e.g.:
  - Formal-informal doublets – learners opt for informal variants: *in spite of* and *till* are overused, *despite* and *until* are underused.
Background

- Granger and Rayson (1998) compare part-of-speech (POS) frequency in native English speaker (Louvain Corpus of Native English Essays (LOCNESS) database) and learner corpora (International Corpus of Learner English (ICLE) database).
- Their major finding is that EFL learners’ (French mother tongue) written language reveals many features characteristic of spoken language and practically none of those of academic, written English, e.g.:
  - Formal-informal doublets – learners opt for informal variants: in spite of and till are overused, despite and until are underused.
  - Academic vs. general vocabulary: nouns associated with academic writing (belief, reasoning, argument) are underused; general and/or vague nouns (people, thing, problem, difficulty) are overused.
Study 1: Research question

- Is the speech-like nature of learners’ written interlanguage (Granger & Rayson 1998) also characteristic of ESL/EFL students with different L1 backgrounds?
Study 1: Research question

- Is the speech-like nature of learners’ written interlanguage (Granger & Rayson 1998) also characteristic of ESL/EFL students with different L1 backgrounds?

- If yes, then the speech-like features previously found in the non-native speakers’ (NNS) corpus are not due to L1 transfer but constitute a developmental stage in SLA.
Study 1: Research question

- Is the speech-like nature of learners’ written interlanguage (Granger & Rayson 1998) also characteristic of ESL/EFL students with different L1 backgrounds?

- If yes, then the speech-like features previously found in the non-native speakers’ (NNS) corpus are not due to L1 transfer.

- Description and analysis of NNS written interlanguage have important implications for SLA theory and pedagogy.
Study 1: Materials

- Non-native speaker (NNS) corpus: Montclair Electronic Language Database (MELD).
  - essays of 34 MSU ESL undergraduate students;
  - 75,042 words;
  - dated 1999-2001;
  - Spanish (18), Chinese / Taiwanese (3), Hindi (2), Malayalam (1), Gujarati (4), Haitian Creole (2), Arabic (1), Bangla (1), Vietnamese (1), and Polish (1).
Study 1: Materials

- **Non-native speaker (NNS) corpus: Montclair Electronic Language Database (MELD).**
  - essays of 34 MSU ESL undergraduate students;
  - 75,042 words;
  - dated 1999-2001;
  - Spanish (18), Chinese / Taiwanese (3), Hindi (2), Malayalam (1), Gujarati (4), Haitian Creole (2), Arabic (1), Bangla (1), Vietnamese (1), and Polish (1).
- **Native speaker (NS) corpus: British Academic Written English (BAWE) corpus.**
  - NS students’ written assignments of a good standard;
  - from first year undergraduate to Master’s degree;
  - in many disciplines;
  - 4,687 255 words.
Study 1: Procedures

- MELD and BAWE were POS-tagged and lemmatized with the use of Tree-tagger.
MELD and BAWE were POS-tagged and lemmatized with the use of Tree-tagger (see next slide).

The tag set used in the research was taken from the Penn Treebank.
Study 1: Procedures

- MELD and BAWE were POS-tagged and lemmatized with the use of Tree-tagger (see next slide).
- The tag set used in the research was taken from the Penn Treebank.
- Frequency lists of major and subcategorized POS were extracted with Unix tools.
Study 1: Procedures

- MELD and BAWE were POS-tagged and lemmatized with the use of Tree-tagger (see next slide).
- The tag set used in the research was taken from the Penn Treebank.
- Frequency lists of major and subcategorized POS were extracted with Unix tools.
- POS profiles of the NNS and the NS corpora were then compared to each other.
Technology is playing an important role in people's life today.

<table>
<thead>
<tr>
<th>Word</th>
<th>Tag</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>NP</td>
<td>Technology</td>
</tr>
<tr>
<td>is</td>
<td>VBZ</td>
<td>be</td>
</tr>
<tr>
<td>playing</td>
<td>VVG</td>
<td>play</td>
</tr>
<tr>
<td>important</td>
<td>JJ</td>
<td>important</td>
</tr>
<tr>
<td>role</td>
<td>NN</td>
<td>role</td>
</tr>
<tr>
<td>in</td>
<td>IN</td>
<td>in</td>
</tr>
<tr>
<td>people</td>
<td>NNS</td>
<td>people</td>
</tr>
<tr>
<td>'s</td>
<td>POS</td>
<td>'s</td>
</tr>
<tr>
<td>life</td>
<td>NN</td>
<td>life</td>
</tr>
<tr>
<td>today</td>
<td>NN</td>
<td>today</td>
</tr>
<tr>
<td>.</td>
<td>SENT</td>
<td>.</td>
</tr>
</tbody>
</table>
Study 1: Results

Major parts of speech distribution

Nouns | Articles | Finite lexical verbs | Auxiliaries | Prepositions & SC | Adjectives | Adverbs | Coordinating Conjunctions

NS | NNS

0 | 5 | 10 | 15 | 20 | 25 | 30

16
Study 1: Results (continued)

**Major & subcategorized POS’s underuse & overuse in MELD**

**Overuse:**
1 & 2 personal pronouns
Coordinating conjunction *but*
Infinitives
Auxiliaries
Indefinite determiners & pronouns
Particles
Adverbs
Present participles

**Underuse:**
Nouns
Articles
Lexical finite verbs
Past participles
Prepositions & subordinating conjunctions
Adjectives
Since the NS corpus is much bigger (N = 4,687,255) than the NNS (MELD) corpus (n = 75,042), one-sample t-test and not Chi-square test was used.

The **One-Sample t-test** compares the mean score of a **sample** to a known value (N).

- **H₀**: observed differences in POS’s frequencies in the two corpora were due to chance.
  Under **H₀**, the difference between POS’s frequencies in the NNS (MELD) and in the NS corpora should be small and t-statistics not too large.

- **Results**: All the compared frequencies in the two corpora (BAWE & MELD) showed t > 3 (p value < 1%). Thus, **H₀** is rejected.
Study 1: Statistical analysis (continued)

\[ t = \sqrt{n} \frac{f_{NNS} - f_{NS}}{\sqrt{f_{NS}(1 - f_{NS})}} \]
Study 1: Discussion

Learners’ POS profile:

- **Nouns**: overall underuse. Specifically, among underused: abstract nouns *(belief, argument)*; among overused: general nouns *(people, difficulty)*.

- **1 & 2 personal pronouns**: overuse.

- **Indefinite determiners & pronouns**: overuse.

English academic texts (based on BNC and LOB):

- nominal style, predominance of abstract nouns (Johansson 1985).

- impersonal style (Pool & Field 1976).

Study 1: Discussion (continued)

Learners’ POS profile:

- **Coordinators**: overuse of *but*.

- **Prepositions & SC**: underuse.

- **Adverbs**: short adverbs of native origin (especially of time and place) are overused, *-ly* adverbs are underused.

English academic texts:

High frequency of *but* has been found to be a feature of spoken, not formal language (Chafe 1982).

A high proportion of prepositions is associated with written English (Rayson et al. forthcoming).

English academic texts favor *-ly* adverbs of Romance origin and disfavor short adverbs of native origin (especially of time and place) (Johansson 1985).
Study 1: Discussion (continued 2)

Learners’ POS profile

- **Verbs:** overuse of auxiliaries and infinitives.
- underuse of past participles.

English academic texts

- overuse of auxiliaries and infinitives is a characteristic of spoken English (O’Donnell 1974).
- underuse of lexical verb forms, both finite and non-finite (past and present participles) is a characteristic of spoken English (Chafe & Danielewicz 1987).
Study 1: A closer look at the learner data

- Overused POS’s in the learners’ corpus are mainly represented by **closed (functional)** classes of parts of speech, i.e.: 1 & 2 personal pronouns, coordinating conjunction *but*, auxiliaries, indefinite determiners & pronouns, and particles.

- Underused POS’s in the learners’ corpus are mainly represented by **open (content)** classes, i.e.: nouns, lexical finite verbs, past participles, and adjectives.
Study 1: A closer look at the learner data

- Overused POS’s in the learners’ corpus are mainly represented by **closed (functional)** classes of parts of speech, i.e.: 1 & 2 personal pronouns, coordinating conjunction _but_, auxiliaries, indefinite determiners & pronouns, and particles.

- Underused POS’s in the learners’ corpus are mainly represented by **open (content)** classes, i.e.: nouns, lexical finite verbs, past participles, and adjectives.

  → Upper-level ESL learners’ vocabulary is **limited**. Instead of content parts of speech, function words are used (e.g., learners use pronouns instead of nouns).

- Type-token ratio in MELD: 0.045.
Study 1: Conclusion

- The data re-confirm the previous finding by Granger & Rayson (1998): ESL upper-level learners’ written language reveals many features of spoken English and practically none of written English.
Study 1: Conclusion

- The data re-confirm the previous finding by Granger & Rayson (1998): ESL upper-level learners’ written language reveals many features of spoken English and practically none of written English.
- So, we can assume that this language behavior is not due to L1 transfer.
Study 2: Materials

- Native speaker (NS) corpus: British Academic Written English (BAWE) corpus.
- NS students’ written assignments of a good standard;
- from first year undergraduate to Master’s degree;
- in many disciplines;
- 4,687 255 words.
Study 2: Materials

- **Native speaker (NS) corpus:** British Academic Written English (BAWE) corpus.
  - NS students’ written assignments of a good standard;
  - from first year undergraduate to Master’s degree;
  - in many disciplines;
  - 4,687 255 words.

- **Non-native speaker (NNS) corpus:** British Academic Written English (BAWE) corpus.
  - NNS students’ essays of a good standard;
  - from first year undergraduate to Master’s degree, in many disciplines;
  - Dutch, Hindi, Chinese (Mandarin), Korean, French & some other L1’s;
  - 2,010 556 words.
Study 2: Research question

How different and/or similar is the performance in academic English of more proficient upper-level NNS (college students in UK) compared to less proficient ESL undergraduate students (study 1)?
Study 2: Results

POS distribution in BAWE NS and BAWE NNS

- Nouns
- Articles
- 1 & 2 personal pronouns
- Finite lexical verbs
- Auxiliaries
- Prepositions & SC
- Adjectives
- Adverbs
- Coordinating Conjunctions

Legend:
- NS
- NNS
Study 1 and Study 2: Results

POS distribution in MELD, BAWE (NS) & BAWE (NNS)
Study 2: Discussion

- Similar POS’s frequencies across NNS (BAWE) and NS (BAWE) corpora (see slide 31).
Study 2: Discussion

- Similar POS’s frequencies across NNS (BAWE) and NS (BAWE) corpora (see slide 32).

- A possible explanation of the observed discrepancy (study 1 vs. study 2; see slide 33):

  - BAWE NNS essays are samples of writing in academic subjects (not in ESL/EFL classes), therefore a higher level of writing was expected.
  - BAWE NNS essays were written by both undergraduate & graduate students (longer exposure to academic English register).
Conclusions & Suggestions for further research

- The study re-confirms the speech-like nature of learner writing of ESL students (Granger & Rayson 1998) of different linguistic backgrounds (study 1).
Conclusions & Suggestions for further research

- The study re-confirms the speech-like nature of learner writing of ESL students (Granger & Rayson 1998) of different linguistic backgrounds (study 1). *Could it be a general pattern of SLA?*

- Possible explanation: communicative ELT approach places more emphasis on speech; lack of exposure to good quality academic writing (Granger & Rayson 1998).
Conclusions & Suggestions for further research

- The study re-confirms the speech-like nature of learner writing of ESL students (Granger & Rayson 1998) of different linguistic backgrounds (study 1). *Could it be a general pattern of SLA?*

- Possible explanation: communicative ELT approach places more emphasis on speech; lack of exposure to good quality academic writing (Granger & Rayson 1998) but with more extended exposure ESL learners’ skills may match those of NSs.

- Further research: do NNS’s writing skills approach NS’s writing skills (study 2; POS’s frequency profiles)? Other features in the NNS data need examination, e.g. use of pre-fabricated expressions, specific syntactic constructions (passive vs. active), etc.
Limitations

- NNS corpus (MELD) was compiled in ESL settings (USA). Two other corpora (BAWE NNS & NS samples) are representative of British academic English. Possible difference: American vs. British English.
NNS corpus (MELD) was compiled in ESL settings (USA). Two other corpora (BAWE NNS & NS samples) are representative of British academic English. Possible difference: *American vs. British English*.

No distinction was made between ‘word choice’ and mistakes. Learner corpora used for the study were not tagged for errors. However, high proficiency of the ESL students allows high validity of the study. 100 random tags (from MELD) were reviewed, 95 out of them were correct.

CI (confidence interval) = (0.937; 1), n=100.

\[
\text{lower} = p - 1.96*\text{sqrt}((p*(1-p))/n) \\
\text{upper} = p + 1.96*\text{sqrt}((p*(1-p))/n)
\]
References

- Fitzpatrick, E. Montclair Electronic Language Database.
References (continued)


*The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).*