Developing and exploiting spoken learner corpora: Challenges and opportunities

Gaëtanelle Gilquin
University of Louvain – FNRS
gaetanelle.gilquin@uclouvain.be

1 Learner language and SLA
• Interlanguage = a learner’s approximate system, which shares features with the learner’s first language and with the target language (cf. Selinker 1972)
• ‘Introduction to second language acquisition research’ (Ellis 2008: 9-17): study of five Mexican children (Wong Fillmore 1976, 1979), study of Alberto (Schumann 1978), study of Wes (Schmidt 1983), study of three classroom learners (Ellis 1984, 1992) and study of Patty (Lardiere 2007)

2 Learner corpus: definition
• Learner corpora are ‘systematic computerized collections of texts produced by language learners’ (…) ‘the selection is representative and balanced’ (Nesselhauf 2004: 125, 127)

3 The advent of learner corpora
• Early 1990s: Longman Learners’ Corpus and International Corpus of Learner English
• First release of ICLE in 2002
• Today: c. 150 learner corpora in ‘Learner Corpora around the World’ repertoire
• BUT more have been collected

4 Typology of learner corpora
• General vs. ESP
• Written vs. spoken vs. multimedia
  • Mute spoken corpora
  • Speech corpora
• Synchronic vs. longitudinal
• Monolingual vs. multilingual
• Mono-L1 vs. multi-L1

5 Louvain International Database of Spoken English Interlanguage (Gilquin et al. 2010)
• International project launched in 1995
• Spoken counterpart of ICLE
• Informal interviews with advanced learners of English from different L1s (11 in LINDSEIv1, 20 in LINDSEIv2)
• Each interviews comes with (socio-)linguistic information about the interviewee (learner) and the interviewer
DEVELOPING A SPOKEN LEARNER CORPUS: CHALLENGES

Corpus design

- All corpora need be assembled according to explicit design criteria (Atkins et al. 1992) but even more important for learner corpora given the inherent heterogeneity of learner output
- "‘mixed bag’ collections of L2 data present little interest' (Granger 2012: 3235)
- ‘there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one may want to control’ (Granger 2004: 126)

Learner recruitment

- Ideally: probability sampling
- Most of the time: convenience sampling
- Volunteering ➔ self-selection ➔ predominance of certain learner profiles
- ‘Corpus-building is of necessity a marriage of perfection and pragmatism’ (McEnery et al. 2006: 73)

Proficiency level

- Proficiency level - A fuzzy variable in computer learner corpora (Carlsen 2012)
- Atkins et al.’s (1992: 5) distinction between
  - Internal criteria = linguistic criteria
  - External criteria = non-linguistic or extra-linguistic criteria
- Common criterion: age or years of instruction (Granger 1998, Tono et al. 2001)
- Unreliability of external criteria for proficiency-level assignment (Thomas 1994)

Degree of naturalness

- ‘genuine communications of people going about their normal business’ (Sinclair 1996)
- ‘(near-) natural foreign or second language learner texts’ (Granger 2008: 338)
- Limited number of functions fulfilled by second and, especially, foreign languages
- The tasks carried out in the classroom may lack authenticity and involve some stress (cf. exams, Luzón et al. 2007: 17)
Transcription

- Requires high-quality recordings
- From basic orthographic transcription to detailed phonological and phonetic transcription
- Costly: one word (‘simple’ orthographic transcription) = c. 1 € (Ballier & Martin 2013: 33)
- Time-consuming: one minute of learner speech = 20-30 minutes of transcription (including post-transcription checks)

Transcription

- Some elements can go unnoticed, cf. Lindsay & O’Connell (1995): discourse markers (e.g. well, now) and hesitation phenomena (filled pauses, false starts, repetitions, etc.) likely to be missing
- ‘auditory hallucinations’ (Stubbs 1983: 228), cf. Randolph (1917: 323): stenographer transcribed ‘they wuz’ as there was but believed that ‘her reports were absolutely faithful’ ⇔ correction of errors

Consent

- Written consent to use the data for research purposes should be obtained
- If the learner is under-age: consent from parent/guardian
- If the learner is of age: he/she should fully understand the nature of the research before giving consent
- If data collection takes place in a school/university: consent from teacher, headmaster or higher-level authority

Anonymisation

- Cf. Guidelines for Research with Children and Young People published by the National Children’s Bureau (Shaw et al. 2011)
  - ‘As far as possible, CYP participating in research should be afforded the same degree of protection regarding confidentiality, anonymity and data protection as adult participants.’ (p. 33)
  - ‘If anonymity cannot be assured we make this clear to participants, and where possible check whether they are comfortable with inclusion of material that might identify them.’ (pp. 25-26)

Alignment + annotation

- Text-sound alignment: never fully automatic
- Most software programs used to annotate corpus data were designed to deal with standard written data
  - Exceptions:
    - SOUP (Gavaldà 2004): parser for spontaneous speech
    - COALA (Pienemann 1992) and COMOLA (Jagtman & Bongaerts 1994): parsers for interlanguage data
  - Several attempts at annotating written learner corpora (e.g. Van Rooy & Schäfer 2002), but few at annotating spoken learner corpora
Alignment + annotation

- Usually require a (relatively) large amount of manual work
- Degree of accuracy not always tested comprehensively – or quite low (e.g., automatic error-tagging of the NICT JLE Corpus: 35% recall, 48% precision; Izumi et al. 2004)
- Output not necessarily fully reliable as a basis for further analysis

DEVELOPING A SPOKEN LEARNER CORPUS: OPPORTUNITIES

Corpus design: metadata

- Rich source of information about the learner, his/her sociolinguistic profile, context of acquisition, etc.
- Could be the starting point of studies investigating the potential impact of each of these variables on learner language

Learner recruitment: local learner corpora

- Global learner corpora: data collected among learners who are subjects providing data for inclusion in the corpus
- Local learner corpora: data collected among one’s students, who are both contributors to and users of the corpus
  - Identify one’s own learners’ specific needs
  - Provide tailor-made solutions to their problems

Proficiency level: from quasi-longitudinal to longitudinal corpora

- Corpora including data from different (identified) proficiency levels can be used as quasi-longitudinal corpora, cf. NICT JLE Corpus (Izumi et al. 2004)
- Corpora can be compiled directly with a longitudinal design (same learners at different proficiency levels)
- (Quasi-)longitudinal corpora make it possible to study learners’ progress (or lack thereof) over time

Degree of naturalness: task effect

- Having access to learner data representing different degrees of naturalness makes it possible to investigate the effect of the task
- More constrained data help answer more specific questions, esp. in terms of lexical knowledge
Transcription

- Having access to transcriptions makes it possible to use the usual text retrieval software to study learner speech (e.g. WordSmith Tools, AntConc)
- The more delicate the transcription, the more refined the types of analysis that are feasible on the basis of the transcripts

Alignment + annotation

- Alignment makes it possible to query text and sound simultaneously (e.g. disambiguation of an item or pronunciation of a word)
- Adding layers of annotation enhances the value of learner corpora
- The more layers of annotation are added, the more research questions can be examined

EXPLOITING A SPOKEN LEARNER CORPUS: CHALLENGES AND OPPORTUNITIES

Focus on speech

- The Written Language Bias in Linguistics (Linell 2005)
  - Writing = “the ‘grammatical’ or correct language” (Linell 2005: 24)
- Speech as “a stepchild of descriptive linguistics” (Tottie 1991: 255)
- Biber et al’s (1999) Longman Grammar of Spoken and Written English
  - Conversational English is “treated as equal in standing to written English” (Hirst 2001:132)
  - Whole chapter devoted to grammar of conversation

Focus on learner speech

- Morphology (e.g. Jantunen & Brunni 2013)
- Lexis (e.g. Hatzidaki 2006)
- Grammar (e.g. Kaneko 2004)
- Phraseology (e.g. De Cock 2004)
- Discourse (e.g. Pérez-Paredes et al. 2008)
- Performance phenomena (e.g. Gilquin 2008)
- Intonation (e.g. Ramírez & Romero 2005)
- Etc.
**Interlanguage phonology**
- Most spoken learner corpora are mute spoken corpora
  - Lack of studies on pronunciation and prosody in learner speech (Ballier & Martin fc.)
  - Most existing studies rely on more constrained data types (e.g., read speech)
  - Phonetic studies require fine-grained levels of speech annotation, often lacking due to 'the difficulty of automating annotation of the speech chain' (Ballier & Martin fc.)

**Comparison of NS and NNS**
- Identify native and non-native features
  - **MISUSE**: the incorrect use of a particular phenomenon
  - **OVERUSE**: the use of more instances of a particular phenomenon than native speakers
  - **UNDERUSE**: the use of fewer instances of a particular phenomenon than native speakers

**Comparison of speech and writing**
- Degree of monitoring
  - E.g., use of the relative pronoun which with animate antecedents:
    - 15% in speech (LINDSEI-SP)
    - Less than 2% in writing (ICLE-SP)
  - Learners seem to monitor their use of relative pronouns to a greater extent in writing
- Register awareness
  - Spoken-like nature of learner writing (Gilquin & Paquot 2008) and written-like nature of learner speech (De Cock 2011)

**Metadata as factors of explanation**
- Any recorded variable can be examined for its potential effect on learner speech
- Comparison of learners displaying the variable and those not displaying it (statistically significant) difference?
- Main focus on L1 as a possible factor of explanation
- Other variables tend to be neglected, despite their obvious interest and potential impact

**Comprehensive overview?**
- Integrated Contrastive Model (Gilquin 2000/2001, based on Granger 1996)
Comprehensive overview?

- Investigation of a single variable to account for a certain behaviour
- BUT ‘the factors that can bring about variation in learner output are numerous, perhaps infinite’ (Ellis 1994: 49)
- Need for statistical methods that take several variables into account simultaneously

L1 TRANSFER

Granger 2004

APPLICATION

Learner corpus-based applications

- Many corpus-based applications for learners, but few learner corpus-based applications, and even fewer spoken learner corpus-based applications
- Potential of spoken learner corpora for applications (esp. in language teaching) with some rather timid attempts

Teaching materials

- Create materials that include more authentic spoken material (cf. Carter & McCarthy 1997, Exploring Spoken English)
- Include sections on ‘the grammar of speech’
- Help learners become ‘stylistically aware’
- Tailor-made materials, focusing on those aspects that appear to be problematic in learner speech

Teaching materials

- Not everything that is typical of speech should be taught: ‘features of learner language uncovered by LC research need not necessarily lead to targeted action in the classroom’ (Granger 2009: 22)
- Examples:
  - there’s + plural N
  - I wonder where the hell is he gone
  - The thing is is that
Data-driven learning

- Use of native spoken data as a model to imitate
- Use of learner spoken data with a corrective function (Gilquin & Granger 2010)
  - Error-tagged learner corpus (cf. Campillos Llanos 2012; interface to search error-tagged oral corpus of L2 Spanish produced by learners from 9 L1 backgrounds)
  - Comparison of NS/NNS corpora: ‘In the teaching of spoken skills there are still many things that may be observed and learnt by contrasting native speaker and learner corpora’ (Luzón et al. 2007: 16)

CEF descriptors

- Improvement / fleshing out of CEF ‘Can Do’ statements on the basis of analysis of CEF rated transcripts
- E.g. refinement of the fluency descriptors (Hasselgren 2002, Götz 2007, Osborne 2007, Dumont 2015)
  - B1: ‘pausing for grammatical and lexical planning and repair is very evident’
  - B2: ‘there are few noticeably long pauses’

NLP

- Automatic error detection (Leacock et al. fc.)
- Automated scoring (Higgins et al. fc.) → language testing/assessment
- Native language identification (Jarvis & Paquot fc.)
- Applications in CALL (e.g. selection of activities based on learner’s automatically identified L1/proficiency = customization), speech recognition, etc.

CONCLUSION

Summary

- Quite a lot of challenges, but many opportunities too, both in the development and the exploitation of spoken learner corpora
- Try to avoid the pitfalls of spoken learner corpus research to take advantage of its benefits

Wish list

- More spoken learner corpora
- More varied spoken learner corpora (languages, proficiency levels, tasks, etc.)
- More exploitation of spoken learner corpora (including concrete applications for the teaching of speech)
- Links with related types of corpora, e.g. ELF corpora (Mauranen 2011, Seidlhofer 2015), corpora of teacher talk (Nicaise 2015) – and with other types of data, esp. experimental data (Gilquin & Gries 2009)