The ISLE Corpus: Italian and German Spoken Learners' English

Eric Atwell, Peter Howarth and Clive Souter
University of Leeds

Background: ISLE project aims

Project ISLE (Interactive Spoken Language Education) aimed to exploit available speech recognition technology to improve the performance of computer-based English language learning systems, specifically for adult German and Italian learners of English. The English language teaching industry is showing increasing interest in and awareness of the relevance and potential of speech and language technology (Atwell 1999). The project conducted a detailed survey and analysis of prospective user requirements (Atwell et al. 2000): we sought expert advice and opinions from a range of prospective end-users (learners of English as a second language), as well as "meta-level experts" or professionals and practitioners in English language teaching (ELT teachers and researchers) and industry experts in the ELT market (publishers of ELT resources, textbooks and multimedia). The ISLE project partners included representative users, English language learners at all six sites in the ISLE project consortium: Dida*el S.r.l. (Milan, Italy), Entropic Cambridge Research Laboratory Ltd. (Cambridge, UK), Ernst Klett Verlag (Stuttgart, Germany), University of Hamburg (Germany), University of Leeds (UK), University of Milan Bicocca (Italy). Leeds University is a centre for English language teaching and research; Leeds University, Hamburg University and Entropic Cambridge had ready access to overseas students from Germany and Italy; Klett is a major German publisher of ELT resources and textbooks; and Dida*el is a major Italian publisher of multimedia educational systems. We developed a demonstrator English pronunciation tutor system, including an error diagnosis module to pinpoint and flag mispronounced words in a learner's spoken input (Herron et al. 1999).

Why collect a corpus?

The ISLE project also collected a corpus of audio recordings of German and Italian learners of English reading aloud selected samples of English text and

dialogue (Menzel et al. 2000). Note that this was **not** the main aim of the project; although corpus collection and annotation were a significant part of the original project proposal, when the budget was later slashed, these plans had to be cut back. Furthermore, we did not set out with the altruistic goal of building a corpus as a generic resource for the wider research community: the corpus was a necessary means to the project's own ends, and we did not have time to consider additional genres, annotations, etc. to make the resource more re-usable by others. Many corpus linguists advocate building more generic resources as tools for theoretical research into corpus-based methodologies for comparing and assessing learner pronunciations (e.g. Weisser 2001, Ramirez Verdugo 2002), but at least some corpus linguists agree with the principle of building specialised corpus datasets for specific problems (e.g. Thomas 2001, Pravec 2002).

The non-native speech corpus was used to optimise the ISLE system recognition and adaptation parameters for non-native speech and low-perplexity recognition tasks, and to evaluate the ISLE system's diagnosis of mispronunciations expected from intermediate learners of English. The corpus therefore contains a representative sample of the target non-native accents and exercise types to be found in the final ISLE system. In addition, the corpus provides empirical evidence of German and Italian English learners' pronunciation errors, which can be compared with expert perceptions in the ELT literature.

Corpus collection

Speech recordings were collected from non-native, adult, intermediate learners of English: 23 German and 23 Italian learners. In addition, data from two native English speakers (Atwell and Howarth) was collected for test calibration purposes. We also recorded data from some speakers with other L1, but did not add annotations (see below) as the ISLE system was to be targeted specifically at German and Italian L1; so the core ISLE corpus distributed via ELRA (see below) does not include these recordings. Two main sets of data were collected from each speaker:

i) The adaptation data was to be used to produce speaker-adapted non-native speech models for use in recognition experiments on the test data. The text prompts for the adaptation data recordings also serve as the enrolment texts in the ISLE demonstrator. This adaptation data allows us to evaluate how much enrolment data should be collected from each new ISLE user in order to give adequate non-native recognition performance. It also allows the adaptation parameters to be optimised for the system. We chose material from a non-fictional, autobiographical text describing the ascent of Mount Everest (Hunt 1996). The copyright for this material is owned by Klett-Verlag, one of the

project partners. It was also selected so that speakers/readers would not have to deal with reported speech or foreign words, which may cause them to alter their pronunciation. Speakers were asked to read aloud a passage of 82 sentences from the text, approximately 1,300 words of the text. This quantity was considered by Entropic to be sufficient for a representative range of phone co-occurrences to be included.

ii) The test data was a series of short utterances which can be recognised using low-perplexity speech-recognition language models. This allows the recognition and diagnosis modules to be evaluated with tasks equivalent to those used in the ISLE demonstrator system. The second kind of data to be collected was intended to capture typical pronunciation errors made by non-native speakers of English. The constraints on this kind of data come firstly from the exercise types which the initial user survey revealed to be important (Atwell et al. 2000) and secondly from the tasks for which the Entropic speech recogniser would be likely to return very high accuracy. The exercises were chosen primarily to test speakers' competence in pronunciation of items within the context of a phrase or sentence. They consist of approximately 1,100 words contained in 164 phrases. We focussed on three known problems:

– Single phone pairs, e.g. "I said b**e**d not b**a**d", "I said g**o**t not g**oa**t"

– Phone clusters, e.g. *"I said **sn**ow not tomorrow", "I said **ch**eap not other"*

– Primary stress pairs, e.g. "*Children often **rebel** against their parents",*
"Singers learn how to **project** their voices"

The data was recorded using *Prompter*, a tool developed at Entropic for the purpose of recording waveforms from a list of text prompts. The tool is able to load any list of prompts, and gives the user functionality to start and stop recording, playback and view each utterance. The tool runs on both NT and Windows 95 and stores waveforms as WAV format files. A standard headset microphone (Knowles VR3565) was used in all recordings (at all six sites). 16 bit waveforms were sampled at 16kHz, the sampling rate used by the Entropic speech recogniser. Speakers took between 20 minutes and one hour each to complete the recording of the 2,400 words, depending on their proficiency and attention to detail. Some completed the whole exercise quite quickly, without bothering to re-record sentences where they knew they had not spoken all the words in the written prompt. Others carefully re-recorded if they realised they had misread the sentence. The total data collected per speaker averaged around 40 megabytes of WAV files.

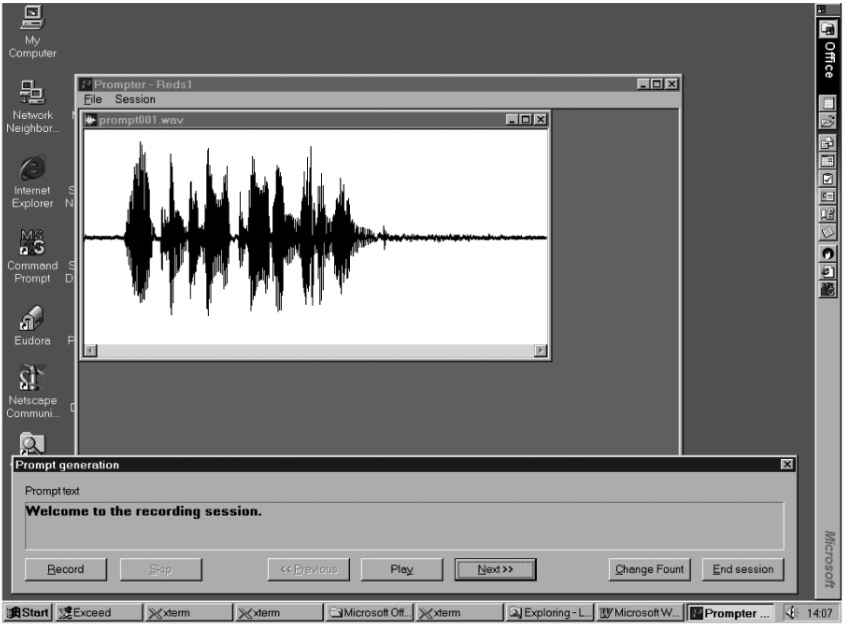


Figure 1: Entropic Prompter Recording Tool

Corpus annotation

The error localization and mispronunciation diagnosis modules of the ISLE system need to pinpoint errors at the phone level. In order to evaluate the performance of these modules, each utterance in the test data set has been annotated at the phone level (adaptation data only needed to be verified at the word level and was not annotated at the phone level). The annotation contains a transcription of how the utterance was spoken by the speaker in relation to a reference transcription containing a canonical native pronunciation. The phone-level reference transcription for each utterance was produced automatically using the Entropic UK English speech-recogniser (Young et al. 1999) running in a forced-alignment mode: the recogniser "knew" the target transcription (i.e. what was being said); it merely had to find the best alignment to the audio signal. Although phoneticians might prefer International Phonetic Alphabet (IPA) labeling as an international standard agreed by academics, the Entropic speech recogniser uses an ASCII-based label set, Entropic's UK English phone set (Power et al. 1996); see Figure 2. This was simpler for us to adopt as it did not require special fonts

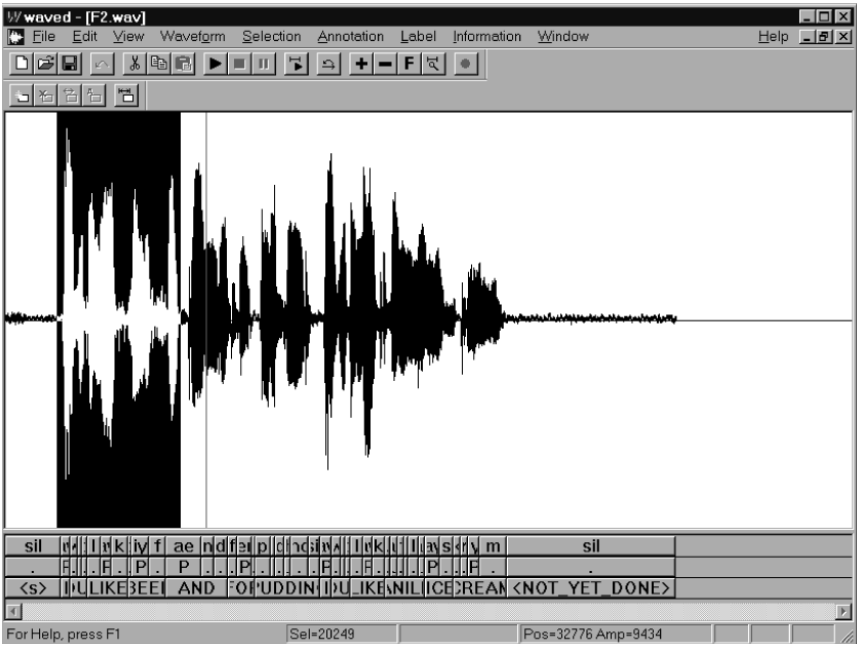
for display and printing. Note however that a mapping exists from IPA to the UK phone set if needed.

A team of five annotators led by Howarth at Leeds University added manual annotations, using the Entropic *WavEd* speech editor and annotator tool. Annotators marked deviations from the reference transcription at the phone level; they also added word-stress markup, and an overall proficiency rating for each speaker. WavEd displays the time-amplitude speech waveform, and aligned reference transcriptions at each of the three levels (word, phone and stressed syllable), as shown in Figure 3. The waveform can alternatively be displayed as a spectrogram. The audio file for a whole utterance can be played back, or the annotator can highlight a section to listen to. When viewing a whole utterance, it is usually not possible to display all the phone and stress labels, so a zoom facility exists so they can be seen more easily (Figure 4). After practice, annotators were able to complete work on each speaker in 5–6 hours (though not as a continuous block of work); the total time taken for all annotation was approximately 300 hours.

Symbol	**Example**	**Symbol**	**Example**
Vowels		**Plosives**	
Aa	B**al**m	B	**b**et
Aa	b**ar**n	D	**d**ebt
Ae	b**a**t	G	**g**et
Ah	b**u**t	K	**c**at
Ao	b**ou**ght	P	**p**et
Aw	b**ou**t	T	**t**at
Ax	**a**bout	**Fricatives**	
Ay	B**i**te	Dh	**th**at
Eh	b**e**t	Th	**th**in
Er	b**ir**d	F	**f**an
Ey	B**ai**t	V	**v**an
Ih	b**i**t	S	**s**ue
Iy	B**ee**t	Sh	**sh**oe
Oh	B**o**x	Z	**z**oo
Ow	B**oa**t	Zh	Mea**s**ure

Oy	B**oy**	**Affricates**	
Uh	B**oo**k	Ch	**ch**eap
Uw	B**oo**t	Jh	**j**eep
Semi-Vowels		**Nasals**	
L	**l**ed	M	**m**et
R	**R**ed	N	**n**et
W	**W**ed	Ng	thi**ng**
Y	**y**et	**Silence**	
Hh	**h**at	Sil	silence
		Sp	short pause

Figure 2: Entropic Graphvite UK Phone Set



"I would like beef and for pudding I would like vanilla ice cream"

Figure 3: Entropic WavEd Speech Editor and Annotator

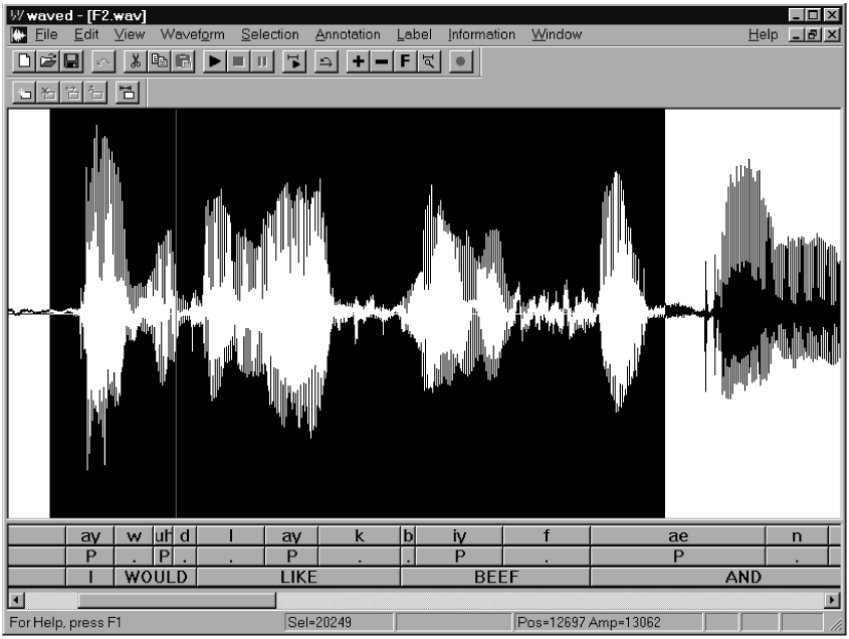


Figure 4: Zooming in on the annotation

Target words in the word-stress subset of the test data were annotated with their expected stress pattern. The stress patterns are defined as sequences of primary and secondary stress. The stress level was annotated in the reference transcription alongside the vowels of the target word.

The phonetic annotations were marked for three kinds of pronunciation errors at the phone level: substitutions, insertions and deletions, plus stress substitution errors. The error annotations take the form ***E-O*** where ***E*** is the expected form seen in the reference transcription and ***O*** is the observed form.

If time/funding had allowed, we wanted to collect "goodness of pronunciation" scores from the human annotators at the utterance and word level and possibly at the phone level. This would have given some finer indication of how well the subject is speaking and could have been used to calibrate and compare the localization and diagnosis components. In practice, however, we decided that goodness of pronunciation was a subjective metric likely to fall foul of inter-annotator variation, and in any case it tended to vary little between utter-

ances of a single speaker, so we only annotated an overall proficiency rating to each speaker-dataset.

In addition to the blocks of individual speaker data, we created five pseudo-speaker blocks of data by selecting some utterances covering all speakers, in order to be able to check inter and intra-annotator consistency. All annotators marked up pseudo-speaker 1 first, then annotated some of the individual speakers, with pseudos 2–5 interspersed in the remaining work (for a detailed analysis, see Menzel et al. 2000). Overall, agreement rates were low: at best, annotators agreed in only 55 per cent of cases when deciding where and what an error is. Even localisation of the error alone, deciding where the error is but not what the correction should be, shows at best a 70 per cent agreement between annotators. In some cases this was because annotators flagged errors in the same word but not the same exact location (phoneme). Furthermore, similar results on the consistency of phone-level annotations have been obtained elsewhere (e.g. Eisen et al. 1992).

Analysis: what does the Corpus tell us about learners' pronunciation errors?
Statistics extracted from the error-annotated corpus allowed us to see which were the most common sources of English pronunciation errors for native speakers of Italian and German:

Italian Native Speakers: Most difficult phones:
/UH/ (51% wrong, often /UW/)
/ER/ (45% wrong, often /EH/+/R/)
/AH/ (42% wrong, often /AX/)
/AX/ (41% wrong, often /OH/)
/NG/ (39% wrong, often /NG/+/G/)
/IH/ (38% wrong, often /IY/)

Italian Native Speakers: Phones that account for the most errors:
/AX/ (13% of errors)
/IH/ (12% of errors)
/T/ (8% of errors; due to schwa insertion)
/AH/ (7% of errors)
/ER/ (6% of errors)
/EH/ (5% of errors)
(schwa insertion accounts for ~15% of errors)

Italian Native Speakers: Words that account for the most errors:
"a" 8% of errors, wrong 42% of the time
"the" 6% of errors, wrong 60% of the time
"to" 4% of errors, wrong 58% of the time
"said" 4% of errors, wrong 49% of the time
"I" 2% of errors, wrong 18% of the time
"and" 2% of errors, wrong 55% of the time
"of" 2% of errors, wrong 34% of the time

German Native Speakers: Most difficult phones:
/Z/ (21% wrong, often /S/)
/AX/ (20% wrong, often /UH/)
/AH/ (20% wrong, often /AX/)
/V/ (17% wrong, often /F/)
/W/ (10% wrong, often /V/)
/UW/ (10% wrong, often /UH/)

German Native Speakers: Phones that account for the most errors:
/AX/ (24% of errors)
/AH/ (9% of errors)
/Z/ (8% of errors)
/T/ (8% of errors; deletion)
/IH/ (7% of errors)
/T/ (8% of errors)

German Native Speakers: Words that account for the most errors:
"to" 9% of errors, wrong 44% of the time
"the" 8% of errors, wrong 31% of the time
"a" 6% of errors, wrong 14% of the time
"of" 3% of errors, wrong 27% of the time
"and" 2% of errors, wrong 31% of the time
"with" 1% of errors, wrong 41% of the time
"potatoes" 1% of errors, wrong 49% of the time

The Italian speakers made an average of 0.54 phone errors per word with a standard deviation of 0.75, while the Germans made an average of 0.16 phone errors per word with a standard deviation of 0.42. This difference may be partly due to the greater phonological similarities between German and English than between Italian and English. Examples of pronunciation errors at each level, subdivided

between German and Italian native speakers are given below, with an indication of whether these are expected (owing to L1 interference and attested in the EFL literature) or unpredictable/idiosyncratic. Annotators reported some difficulty in deciding which errors to mark at word level and which to mark as phone level; for example in the case of a spurious *s* being appended onto a noun or verb, it is difficult to decide whether the speaker is performing a systematic pronunciation error, or intending to pronounce a different word from the one in the prompt.

Word level (not systematic or easily predictable):

Italian	German
photographic → photography	not be → be not
than/then → that	the → a
deserted → desert (phone error?)	month → week
like to → to like	of → about
+ the	+ more
- to	- in

Stress level (largely as predicted):

Italian	German
′photographic	′report
′convict / con′vict	′television
′components	contrast / contrast

Phone level (as predicted + idiosyncratic):

Italian vowels	German vowels
said: eh → ey	′produce: oh → ow
bed: eh → ae	cupboard: ax → ao
planning: ae → ey	pneumatic: uw → oy
ticket / singer / visit: ih → iy	outside: aw → ow
biological:	staff: aa → ae
ay → iy, oh → ow, ih → iy, ax → ae	dessert: ih → iy

Italian consonants

sheep: + ax
honest: + hh
thin: th → t
sleep: s → z
ginger: jh → g (x2)
singer: ng + g
bait: - t

German consonants

pneumatic: + p
said: s → z
visa: v → w
weekend: w → v
the: dh → d
biscuit: + w
thumb: + b
finger: - g
dessert: - t

Conclusions

The goal of project ISLE was to exploit available speech recognition technology to improve the performance of computer-based English language learning systems. The ISLE project also collected a corpus of audio recordings of German and Italian learners of English reading aloud selected samples of English text and dialogue, to train the speech recognition and pronunciation error-detection modules. Speech recordings were collected from non-native, adult, intermediate learners of English: 23 German and 23 Italian learners. In addition, data from two native English speakers was collected for test calibration purposes. The corpus contains 11,484 utterances, 1.92 gigabytes of WAV files, and 17 hours, 54 minutes, and 44 seconds of speech data. The corpus is based on 250 utterances selected from typical second language learning exercises. It has been annotated at the word and the phone level, to highlight pronunciation errors such as phone realisation problems and misplaced word stress assignments.

We aimed to balance the speaker set for gender, age and accent variation as much as possible, but ended up with more male than female volunteers (32:14). However, this might be excused on at least two grounds: (1) given we only have 46 speakers it would be unwise to attempt to draw conclusions about gender-based language variation from our sample, even if the genders were evenly split 23:23; and (2) the target market for the ISLE system, home and business PC users, is predominantly male.

In addition to the blocks of individual speaker data, we created five pseudo-speaker blocks of data by selecting some utterances covering all speakers, in order to be able to check inter and intra-annotator consistency. Overall, agreement rates were low: at best, annotators agreed in only 55 per cent of cases when deciding where and what an error is. Even localisation of the error alone, deciding where the error is but not what the correction should be, shows at best a 70

per cent agreement between annotators. In some cases this was because annotators flagged errors in the same word but not the same exact location (phoneme). Given the poor inter-annotator agreement on the exact location and nature of errors, the target one might reasonably set for diagnosis programs should be limited to only those errors which annotators agree on; this applies not only to the ISLE system but also to other pronunciation correction systems.

Statistics extracted from the error-annotated corpus allow us to see which are the most common sources of English pronunciation errors for native speakers of Italian and German. For both Italian and German native speakers, we have empirical evidence on which are the most difficult phones and which phones account for most errors (equivalent to the type/token distinction in corpus frequency counts), and which words account for the most errors. The Italian speakers made an average of 0.54 phone errors per word with a standard deviation of 0.75, while the Germans made an average of 0.16 phone errors per word with a standard deviation of 0.42. This difference may be partly due to the greater phonological similarities between German and English than between Italian and English. Examples of pronunciation errors at each level have been evidenced, with an indication of whether these are expected (owing to L1 interference and attested in the EFL literature) or unpredictable/idiosyncratic.

We welcome corpus re-use by other researchers, who can acquire a copy (on four CDs) from ELDA. The data has been used to develop and evaluate automatic diagnostic components, which can be used to produce corrective feedback of unprecedented detail to a language learner. At the end of the project, development of the ISLE pronunciation tutor system stopped at the Demonstrator stage, and future prospects for migration to a commercial ELT package are uncertain. However, we hope that the ISLE Corpus may be a useful achievement of the project.

Acknowledgements

This paper reports on a collaborative research project; we gratefully acknowledge contributions of a number of collaborators, principally: Wolfgang Menzel, Dan Herron, and Patrizia Bonaventura, University of Hamburg (Germany); Steve Young and Rachel Morton, Entropic Cambridge Research Laboratory Ltd. (Cambridge, UK); Jurgen Schmidt, Ernst Klett Verlag (Stuttgart, Germany); Paulo Baldo, Dida*el S.r.l .(Milan, Italy); Roberto Bisiani and Dan Pezzotta, University of Milan Bicocca (Italy). We are particularly endebted to Wolfgang Menzel for setting up and leading the ISLE project and to Uwe Jost (Canon Research Europe) for proposing Leeds University as a contributor to the project.

This research was supported by the European Commission under the 4th framework of the Telematics Application Programme (Language Engineering Project LE4-8353). The corpus is distributed for non-commercial purposes through the European Language Resources Distribution Agency (ELDA).

References

Atwell, Eric. 1999. *The language machine*, British Council, London.

Atwell, Eric, Peter Howarth, Clive Souter, Paulo Baldo, Roberto Bisiani, Dan Pezzotta, Patricia Bonaventura, Wolfgang Menzel, Dan Herron, Rachel Morton, and Juergen Schmidt. 2000. User-guided system development in interactive spoken language education. *Natural Language Engineering Journal*, vol. 6 no. 3–4, Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, 229–241.

Eisen, Barbara, Hans Tillmann, and Christoph Draxler. 1992. Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases, *Proceedings of ICSLP'92: International Conference on Spoken Language Processing*, 871–874.

Herron, Dan, Wolfgang Menzel, Eric Atwell, Roberto Bisiani, Fabio Daneluzzi, Rachel Morton, and Juergen Schmidt. 1999. Automatic localization and diagnosis of pronunciation errors for second language learners of English. *Proceedings of EUROSPEECH'99: 6th European Conference on Speech Communication and Technology*, vol. 2, 855–858. Budapest, Hungary.

Hunt, John. 1996. *The ascent of Everest.* Stuttgart: Ernst Klett Verlag, English Readers Series.

Menzel, Wolfgang, Eric Atwell, Patricia Bonaventura, Dan Herron, Peter Howarth, Rachel Morton, and Clive Souter. 2000. The ISLE Corpus of non-native spoken English. In M. Gavrilidou, G. Carayannis, S. Markantionatou, S. Piperidis and G. Stainhaouer (eds.). *Proceedings of LREC2000: Second International Conference on Language Resources and Evaluation*, vol. 2, 957–964. Athens, Greece. Published and distributed by ELRA – European Language Resources Association.

Power, Kevin, Rachel Morton, Colin Matheson, and Dave Ollason. 1996. *The Graphvite book 1.1*, Entropic, Cambridge.

Pravec, Norma. 2002. Survey of learner corpora. *ICAME Journal* 26: 81–114.

Ramirez Verdugo, Dolores. 2002. Non-native interlanguage intonation systems: a study based on a computerized corpus of Spanish learners of English. *ICAME Journal* 26: 115–132.

Thomas, Jenny. 2001. Negotiating meaning: a pragmatic analysis of indirectness in political interviews. Invited plenary paper, Corpus Linguistics 2001 Conference, Lancaster University, UK.

Weisser, Martin. 2001. A corpus-based methodology for comparing and evaluating different accents. In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.). *Proceedings of the Corpus Linguistics 2001 Conference*, 607–613. UCREL: University Centre for Computer Corpus Research on Language, Lancaster University, UK.

Young, Steve, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Philip Woodland. 1999. *The HTK book 2.2*, Entropic, Cambridge.