

Prosodic transcription of Glasgow English: an evaluation study of GlaToBI

Catherine J. Mayo

A thesis submitted in fulfilment of the requirements
for the degree of MSc in Speech and Language Processing
to the
University of Edinburgh
1996

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

September 1996

Acknowledgements

My thanks go to Bob and Ellen, supervisors extraordinaire—for the encouragement, and for setting the whole project going in the first place. Thanks should also go to Mike, who acted as pinch hitter when things looked messy.

Many thanks to Irene, Norman, Cedric, Morag and Eddie, who made sure that the computing end of things didn't fall apart too drastically. Thanks also to Ethel, for making sure that the rest of the department didn't fall apart either.

An electronic thank you to Doug Bors, who very kindly took the time to hunt down and learn a new statistic, and then explain it via e-mail to this mathematical duffer.

Thank you to Jim, Rob, Shane, Ellen, Matthew, Jacqueline and Paul: the victims of this EOUS (Experiment Of Unusual Size)—my apologies to you all. And finally, special thanks go to Rob for his unflagging patience with my programming questions, for not giving up on me completely when nothing would work and I insisted that “I haven't changed anything!”, and for just generally being there when things got rough.

Contents

Acknowledgements	ii
1 Introduction	1
1.1 Intonation	2
1.1.1 What is intonation?	2
1.1.2 Description of intonation	3
1.2 Intonational transcription systems—ToBI	5
1.2.1 Motivation for the system and system goals	5
1.2.2 System description	6
1.3 Dialectal variation in intonation	10
1.4 Adaptation of ToBI for other languages and dialects	12
1.5 GlaToBI—ToBI adapted for Glasgow English	13
1.6 The Map Task corpus	17
2 Experiment	19
2.1 Subjects	19
2.2 Data	20
2.3 GlaToBI	21
2.3.1 Presentation	21
2.3.2 Software support	24
2.4 Training	24
2.5 Experiment proper	25
2.6 Methods of data collection and analysis	26
2.6.1 Pairwise analysis	26
2.6.2 Kappa	26
2.6.3 Collection of data	27

3	Results, Analysis and Discussion	30
3.1	Pairwise analysis	30
3.1.1	Results	30
3.1.2	Discussion	30
3.1.3	Results of pairwise analysis compared with ToBI and GToBI . .	32
3.2	Kappa analysis	34
3.2.1	Results	34
3.2.2	Discussion	34
3.3	Evaluation of GlaToBI innovations	36
3.4	Evaluation of the study	39
3.4.1	The relationship between the training and the test	39
3.4.2	The current ToBI set-up	41
4	Conclusions	42
4.1	Recommendations for the evaluation study and suggestions for future analysis	42
4.2	Evaluation of the GlaToBI system	43

List of Figures

2.1	Blank GlaToBI transcription files	22
2.2	Completed GlaToBI transcription file.	23

List of Tables

1.1	Stylised representations of pitch final contours in ToBI and GlaToBI . . .	15
1.2	ToBI and GlaToBI tone menus.	16
2.1	Linguistic background of subjects	20
2.2	Dialectal background of subjects.	20
3.1	Table of results for pairwise analysis.	31
3.2	Comparison of pairwise results.	33
3.3	Table of results for kappa analysis.	35

Chapter 1

Introduction

There are efforts underway to prosodically transcribe the University of Edinburgh’s digitally encoded Map Task Corpus using a ToBI-style transcription system. The predominant dialect of the speakers in this corpus is Glasgow English, a dialect which has been found to have an intonation pattern which differs significantly from those of southern British or general American dialects (Cruttenden 1986).

The original ToBI system was introduced by its developers as “a standard for labelling English prosody” (Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert & Hirschberg 1992). With the current interest in speech synthesis, speech recognition and general computational models of speech, ToBI’s creators saw an increasing need for *consistency* in the prosodic transcription of speech. ToBI was therefore developed as a system for prosodic analysis which would be “analogous to IPA for phonetic segments” (Silverman et al. 1992). This, in addition to the fact that ToBI was specifically designed to be used with digital speech databases, make it an appropriate tool for the transcription of the Map Task Corpus. However, the original version of ToBI was developed for “the three most widely used varieties of spoken English—namely, general American, standard Australian, and southern British”, and in fact specifically states that it is inadequate for Glasgow English (Beckman & Ayers 1993). In light of this fact, a version of ToBI that *is* compatible with Glasgow English intonation has been developed by the university’s Linguistics Department and Centre for Speech Technology Research. Before GlaToBI (Glasgow English ToBI) is fully implemented, it has been recommended that an evaluation study be carried out to assess the new version’s reliability—this study, along with investigations of possible theoretical questions it introduces, will form the basis of this paper.

1.1 Intonation

1.1.1 What is intonation?

It is clear to anyone who listens to speech that intonation, or prosody, plays a role in conveying discourse information. Consider, for example, the difference in meaning between the following two sentences, said aloud: *Martha even did her HOMEWORK*, and *MARTHA even did her homework*. Syntactically the utterances are the same, but the shift in emphasis indicates a corresponding shift in meaning (in the first sentence Martha got carried away with diligence and even did her homework, in the second sentence, even Martha, who is generally the last in the class to pick up a pencil, did her homework). However, what exactly is it about the way these two sentences are uttered that indicates they mean different things—in other words, what is intonation and how is it different from other phonetic and phonological aspects of an utterance?

Ladd (1996*a*) separates intonation out from the other aspects of speech on three different levels. The first division he makes is one between segmental and suprasegmental features, the segmental features being all those which have to do with the manner of articulation of elements of an utterance. The lack of consensus in the speech sciences as to the exact nature of intonation has historically led to some confusion over the names of the suprasegmental components of speech. However, in general these features have been “traditionally defined as pitch, stress, and quantity (or f0...intensity, and duration)” (Ladd 1996*b*).

Ladd makes a second division, between the effects of pitch, stress and duration at a lexical level, and their effect at the post-lexical level of an utterance. Suprasegmental features can work at both levels, however their function, and whether or not they are required, differs across the two levels. Lexical level stress, for example, is an obligatory part of the composition of a word in an utterance: the word *complete* is said *com-plete* rather than *COM-plete*; the word *comfort* is said *COM-fort*, not *com-FORT*. Post-lexical prosody, on the other hand, is the prosody that is generally considered to be involved in the conveyance of overall sentence meaning. It is this level of interaction between the utterance and pitch, stress, and duration that makes up intonation—post-lexical, or sentence level stress, for instance, is what was at work on Martha and her homework in the sentences above.

The intonational use of suprasegmental features is separated from their use in other aspects of speech by a third division made by Ladd: that between linguistic and paralinguistic. The suprasegmental features that fall under the heading linguistic are those “intonational features [that] are organised in terms of categorically distinct entities (e.g. low tone or boundary rise) and relations (e.g. stronger than/weaker than)”

(Ladd 1996*a*). Paralinguistic features are those “in which continuously variable physical parameters (e.g. tempo or loudness) directly signal continuously variable states of the speaker” (Ladd 1996*a*). Paralinguistic features, while relevant emotionally as it were, do not carry any linguistic meaning—the utterance *Are you kidding?* will intonationally remain an incredulous question whether shouted or whispered.

1.1.2 Description of intonation

There are many different ways to go about describing intonation, limited only by the equally great number of intonational theories which have engendered these descriptions. As stated by Ladd (1996*a*) “Research on intonation has long been characterised by a number of unresolved basic issues and fundamental differences of approach. Until recently, these have precluded the emergence of any widely accepted framework for the description of intonational phenomena, or even any general agreement on what the interesting phenomena are”. The problem arises when a description has to be made, as this requires taking a stand on one side or the other of a still heated and contentious debate. The purpose of this study is the evaluation of a ToBI-style intonational transcription system. The description that follows, therefore, will be taken for the most part from the theory of intonational description by Pierrehumbert (1980), on which the original ToBI system was partially based.

Pierrehumbert (1990) states that intonational patterns are the result of the interaction between four distinguishable elements: stress, tune, phrasing and pitch range. The first of these, stress, acting on some syllables and not on others, creates a “pattern of relative prominence” (Pierrehumbert & Hirschberg 1990)—across an utterance there will be some syllables which will be perceived as more prominent or important than others. As mentioned above, stress can act at both a lexical and at a sentence level. Sentence level stress in general must follow all the requirements of the lexical stress rules (i.e. in the utterance *I REFUSE to do my homework!*, the main emphasis is on *REFUSE*, but the lexical stress dictates that it falls primarily on *-FUSE* rather than *RE-*).

More importantly, however, sentence stress “is affected by considerations of information structure” (Pierrehumbert & Hirschberg 1990): the main element of the utterance, or the speaker’s intended *focus* will dictate the placement of sentence level stress. If we consider the following exchange:

A—Martha did her HOMEWORK.

B—Whose homework did she do?

A—Martha did HER homework.

it is clear that the main stress can shift according to the intended meaning of the utterance. This main phrasal stress is generally referred to as the *nuclear* stress of the utterance (Pierrehumbert & Hirschberg 1990, among others), as it plays a central role in the assignment of stress throughout the rest of the utterance—in A’s second repetition above, the placement of the nuclear stress on *her*, forces *homework* to be de-stressed.

The following exchange features two utterances which differ in their fundamental frequency (f_0), or pitch pattern:

A—*Mary’s getting married.*

B—*Mary’s getting married?!*

This second element of intonation is called the tune of the utterance in Pierrehumbert’s analysis. The study of the composition of intonational tunes is one in which the lack of agreement in the field of intonation is very obvious. In general, pitch patterns are either described as whole tunes which are superimposed, blanket-style, over an utterance, or, as in the case of Pierrehumbert’s work, they are seen to consist of “sequences of discrete intonational events” (Ladd 1996*a*). For Pierrehumbert these events are described in terms of high or low tones; the course that the speaker’s pitch takes between these tones is seen as a type of non-message-bearing connective pitch. Those tones that occur at stressed or prominent syllables are called *pitch accents*, while those that occur at the ends of phonological phrases Pierrehumbert calls *phrasal tones*. It is important to note, as pointed out by Ladd (1996*a*) that in Pierrhumbert’s analysis pitch accents are not direct features of stress or prominence. Instead, “general conditions of prosodic well-formedness stipulate that pitch accents must occur *with* [my emphasis] prominent stressed syllables, and the occurrence of a pitch accent therefore serves as a *cue* to the location of prominence”. If the speaker’s overall intended meaning changes, therefore, the pitch accents will also change, as will the phrasal tones if the division of the utterance into phrases is changed (Pierrehumbert & Hirschberg 1990).

Pierrehumbert identifies two types of phrasal tone, corresponding to the two levels of intonational phrasing that she has postulated for English. *Phrase accents* mark the end of the smaller level of phrasing: the intermediate phrase, which is made up of one or more pitch accents and the phrase accent. The larger level of phrasing, the intonational phrase, is made up of one or more intermediate phrases, and is marked at the end by the phrase accent of the final intermediate phrase, plus a *boundary tone* to mark the end of the intonational phrase. (Pierrehumbert & Hirschberg 1990)

The final element of Pierrehumbert’s analysis, pitch range, is another issue in the field of intonation that has not been resolved. Other than those shifts in pitch due to paralinguistic considerations, it has also been observed that “ f_0 tends to decline over

the course of phrases and utterances” (Ladd 1996*a*), a phenomenon that is known as declination. Prosodic analysts tend to follow whichever method of description they have adopted for examining tunes, therefore those who see pitch in terms of whole tunes will tend to view this movement as the result of an overall downward sloping of the pitch. Pierrehumbert’s analysis, on the other hand, posits that the movement is the result of a local stepping down of pitch—an accented high tone following another accented high tone will have a slightly lower actual pitch relative to the first tone (Ladd 1996*a*).

This is only a brief overview of the intricacies of intonation. Specific details of Pierrehumbert’s autosegmental theory of intonational description, as well as a number of other prevailing theories, will be discussed in further detail in relation to the ToBI transcription system (see section 1.2.2).

1.2 Intonational transcription systems—ToBI

1.2.1 Motivation for the system and system goals

There were number of developments in the speech sciences which led to the creation of the original ToBI transcription system. The growing field of speech technology requires large databases of labelled speech (on which to train recognition and synthesis systems, for example). In addition, it has become clear that “prosody is central to the interface between speech and natural language processing technologies” (Silverman et al. 1992): in other words, prosody seems to carry a lot of the information in an utterance, and may play a large role in the disambiguation of speech. Therefore, not only does speech technology need large corpora of speech, it also needs access to the prosodic content of that speech. New techniques have been successfully developed to elicit large quantities of spontaneous speech, the most appropriate type of material for training machines which attempt to model “real” speech. Unfortunately, as stated by the ToBI designers “large corpora are of little use unless they are annotated in some way that permits retrieval and analysis of similar phenomena” (Silverman et al. 1992). Phonetic (and phonological) information can be consistently accessed from speech by means of the International Phonetic Alphabet (IPA)—an agreed upon standard for phonetic transcription. However, because of the traditional splits along theoretical and methodological lines in the study of prosody, until recently no such agreed industry standard existed for the transcription of intonation. ToBI, according to its designers, was an attempt to fill that need.

The name ToBI stands for *Tones* and *Break Indices*, and gives a good indication

of the main components of the system. The tonal aspect of a ToBI transcription targets the actual tune or melody of an utterance. Based primarily on Pierrehumbert's work, intonation in this part of the system is stripped down to a sequence of the most relevant events in the pitch contour. The part of the system that is concerned with break indices is based on work on phrasing in speech by Price et al. (1991). This group was interested in the relationship between prosody and syntax, and in particular in the way that speakers use prosody to disambiguate potentially misleading utterances. In the course of their investigation, the group developed an impressionistic coding system for rating the strength of boundaries or *breaks* between words in an utterance, the central idea of which was adopted for ToBI.

The goals of the original ToBI system designers were clear—to design a system of intonational transcription with the following features:

1. reliability: agreement between different transcribers must be at least 80%.
2. coverage: [the system must be] sufficiently comprehensive to capture the most important prosodic phenomena in spontaneous speech.
3. learnability in a relatively short time, in order to be used in multi-site data collections.
4. capability of being related to current approaches to speech recognition, to parser outputs, and to formal representations of semantics and pragmatics.

(Silverman et al. 1992)

1.2.2 System description

There are three elements necessary for a ToBI transcription: the speech, in the form of a recording of the utterance (it may also additionally be represented visually by a speech waveform), a representation of the pitch contour, in the form of a fundamental frequency (f0) track, and a set of labels associated with both of the above. The labels consist of four tiers, each representing one element of the actual transcription of the utterance.

THE ORTHOGRAPHIC TIER. This is generally the first tier to be transcribed (for a strictly prosodic transcription, this may even be pre-transcribed before presentation to the user). This tier of the labels holds the orthographic transcription of the utterance—i.e. a written version of the spoken utterance. As this part of the system technically only exists as an aid to the prosodic transcription, it is transcribed using the standard roman alphabet, rather than the IPA. However, orthographic transcribers have some

freedom in terms of how they represent parts of the utterance which are not strictly words: disfluencies, pauses (filled, i.e. “um”, “er”, and unfilled), and contractions (i.e. “gonna”) can be represented in a number of ways. The designers of the system point out that there are various existing conventions for the transcription of these and other phenomena of spontaneous speech which can be adopted for this part of the system (Beckman & Ayers 1993).

THE TONE TIER. This tier and the following (the break index tier) make up the core of the prosodic transcription. As mentioned above, the transcription in this part of the system is autosegmental in nature—this tier therefore holds a symbolic representation of pitch movement in terms of discrete pitch prominences (*pitch accents*) and intonational boundary pitch (*phrasal tones*). Following Pierrehumbert’s system, ToBI makes use of two main tones: high (**H**) and low (**L**), which can be combined in various ways to mark all the tonal events of an utterance.

Pitch accents: All syllables in an utterance that are perceived as accented (stressed) by the transcriber are marked by one of the following pitch accents:

H*, **L***, **L+H***, **L*+H**

or their downstepped counterparts:

!H*, **L+!H***, **L*+!H**, **H+!H***

A simple H* or L* on a syllable corresponds to a tone target on a local peak or valley in the speaker’s pitch range. The two compound accents L+H* and L*+H both correspond to pitch movement from a point low in the speaker’s range to higher in that range. However, the part of the tone to which the stressed syllable is aligned in time (indicated by the * diacritic) is different for each tone. The L*+H accent is “a *low* tone target on the accented syllable which is immediately *followed* by [a] relatively sharp rise”, while the L+H* accent is “a *high* peak target on the accented syllable which is immediately *preceded* by a relatively sharp rise” (my emphases) (Beckman & Hirschberg 1994).

The downstepped tones are used to mark the overall decline of the speaker’s fundamental frequency across an utterance. As mentioned above, in the autosegmental approach this declination is the result of “the stepwise lowering of pitch...at specific pitch accents” (Ladd 1996a). In other words, the gradual lowering of pitch across an utterance comes about because, while all the H and L tones remain high or low relative to each other, the actual pitch of each H tone may be stepped down from the pitch of the preceding H tone. It should be noted that the use of the diacritic ! to mark downstepped tones is a departure from Pierrehumbert’s work. Pierrehumbert builds the phenomenon of downstep directly into the tones themselves, where this system, based on suggestions from Ladd (1983) makes the choice of downstep independent from the

choice of the actual tone. Because of this, the inventory of pitch accents used in the ToBI system does not correspond exactly to Pierrehumbert’s tonal inventory.

Phrasal tones: Again, following Pierrehumbert, there are two types of phrasal or boundary tone which are marked in the ToBI system—intermediate phrase tones, and intonational or full phrase tones. According to Pierrehumbert & Hirschberg (1990), “a well-formed intermediate phrase consists of one or more pitch accents, plus a simple high or low tone...which marks the end of the phrase”. This simple high or low tone is called a *phrase accent* and is represented in the ToBI system by the following:

H-, L-.

These phrase accents, along with indicating the end of an intermediate phrase, also mark the level of the pitch from the last pitch accent to the boundary. In addition, the H- phrase accent has an *up-step* cuing property, raising any boundary tone which it precedes.

An intonational phrase is “composed of one or more intermediate phrases” (Pierrehumbert & Hirschberg 1990), the end of which is marked with a *boundary tone*:

H%, L%.

In addition, every boundary tones is, by rule, preceded by a phrase accent, “since the end of every intonational phrase is also the end of an intermediate phrase” (Pierrehumbert & Hirschberg 1990). Unfortunately, however, this gives rise to a theoretical problem regarding the actual placement of the final phrase accent. In Pierrehumbert’s analysis phrase accents do double duty, acting as the marker both for the end of an intermediate phrase and for the relative position of the pitch from the nuclear accent to the end of the phrase. The problems begin when an utterance has a nuclear H* pitch accent (for example) which occurs early in the utterance, and which is followed by a drop in the pitch and a low pitch plateau stretching to the end of the utterance. If the phrase accent is to mark the fall in pitch after the nuclear pitch accent, it should be placed fairly close to the nuclear accent itself, and therefore some distance from the end of the utterance. However, the phrase accent must also mark the end of the intermediate phrase—which in this case is also the end of the utterance—meaning that it should be placed near the final boundary tone. The designers of the ToBI system were forced to make a choice between these two placements by another complication: the fact that a phrase accent must be accompanied by a break index of 3. If the phrase accent is placed immediately after the nuclear pitch accent, then a level 3 break index may have to be marked on the utterance long before the end of the intermediate phrase. As a result ToBI convention chooses to align all final phrase accents with end-of-utterance boundary tones, listing them as four compound tones (Ladd 1996*a*):

H-H%, H-L%, L-L%, L-H%.

Both the boundary tones in the two compound tones H-H% and H-L% are affected by the up-step cuing H- phrase accent. This causes the L% tone to be realised in the middle of the speaker's range, and the H% to be realised at a point higher than any previous pitch. The resulting tones described by the two compounds are a mid to high plateau (H-L%), and a high followed by a rise to an even higher pitch (H-H%). The L-L% compound also needs some explanation—this tone describes a low pitch which falls even lower in the speaker's range at the edge of the intonational phrase.

THE BREAK INDEX TIER. Essentially, this tier is used to mark the strength of boundaries or breaks between words. The labels represent “a rating for the degree of juncture perceived between each pair of words and between the final word and the silence at the end of the utterance” (Beckman & Hirschberg 1994). *Words* here include word fragments as well as filled pauses (“er”, “um”). There are five basic values of break indices in a ToBI transcription, varying in strength from 0 to 4:

0, 1: These two break indices are used for junctures between words, and mark the weakest boundaries. 0 indicates no break at all (as in the boundary between *got* and *you* in “gotcha”), while 1 is the standard break between words within the same intermediate phrase.

3, 4: Break indices 3 and 4 are used to represent breaks at phrase endings—the strongest boundaries. The end of an intermediate phrase is marked with break index 3, and the end of a full intonational phrase is marked with a 4.

2: Break index 2 is also used at phrase boundaries, but in two very specific situations. Because of the co-ordination between the break index tier and the tone tier that is required by the ToBI system, there are occasions when a strict transcription could lead to a contradiction between the two tiers. In some cases where a phrase tone is obligatory (i.e. to mark the state of the pitch between one pitch accent and the next, or a pitch accent and a boundary) there may be no obvious audible break in the speech, making the use of break index 3 questionable. The opposite may also occur—there may be an audible break in the speech, requiring a break index of 3 or 4, without a corresponding break in the movement of the pitch (and therefore without any possible tonal representation). In both these cases break index 2 is used to indicate a less precise match between the tiers (Ladd 1996*a*, Beckman & Hirschberg 1994).

THE MISCELLANEOUS TIER. This final tier is used for all those elements of an utterance which cannot be orthographically transcribed, but which could play a role in the interpretation of the intonation. Marked in this tier, therefore, are general disfluencies, silence, speaker's breathing or laughter and any environmental noise which has affected the pitch track (interruption by another speaker, speaker movement).

DISFLUENCIES, UNCERTAINTY. In addition to the miscellaneous tier the ToBI system provides labels for disfluencies in both the tone tier and the break index tier, as well as labels to be used when the user is unsure as to the placement or choice of break index or tone.

(Beckman & Hirschberg 1994)

1.3 Dialectal variation in intonation

The study of intonation in English has for the most part concentrated on the intonational patterns of “standard” or more widely occurring dialects of English, specifically Received Pronunciation (RP) and Standard American English (SAE), with very little investigation into local variants. In fact, as a whole there has been very little investigation into the variation in intonation across dialects of English—as stated by Cruttenden (1986), “all that is available...is a number of sketchy articles, and paragraphs in books and articles which are only suggestive”. What these books and articles do suggest is that dialectal variation occurs; what they don’t go into in much detail are the actual forms that these variations take. As a result, the specifics of intonation in dialects of English which are not RP or SAE are negligible to nonexistent. One group of dialects which *has* received some attention is the group which Cruttenden (1994) (also previously 1986) calls urban north British (UNB). According to Cruttenden the most salient intonational difference between RP and UNB is “the more extensive use of rising tones in many northern [British] cities” (1986). Cruttenden’s descriptions of the intonational patterns of these dialects have therefore been more or less restricted to this one feature.

Cruttenden states that the dialects which comprise the UNB dialect group are strictly those found in large cities, in particular Belfast, Derry, Newcastle, Liverpool, Birmingham and, most importantly to this study, Glasgow. (It is interesting to note that the dialect of Edinburgh is *not* considered to be one of the UNB dialects.) Of these dialects, the intonation of Glasgow English seems to be the least well documented. Cruttenden (1994) cites a report by Currie which states that “each unit [in Glasgow English] ends with a rise on the final stressed word of the unit”, and that “the final stressed element [i.e. the nuclear tone] is low and not high”. Cruttenden surmises from this that a rising tone at the end of the intonational phrase is the common tonal structure in Glasgow English, however there is little given in the way of description of this structure. Cruttenden does point out, though, that “impressionistically the intonation of Glasgow is very similar to that of Belfast” (Cruttenden 1994); details from the description he makes of the rises in Belfast English could therefore help to build up a better picture of the standard intonation patterns of Glasgow English.

According to Cruttenden there are three main tones found in Belfast English: a **low rise**, a **high rise**, and a **rise fall**. The low rise is characterised by a low pitch on the nuclear or main accent in the utterance, which rises to mid-range and remains there for the rest of the intonational phrase. Cruttenden claims that for Belfast English this configuration acts as the unmarked or standard tone for declarative utterances, for tag interrogatives (i.e. “It’s a nice day, *isn’t it?*”, or “It’s a nice day, *eh?*”), and interrogative utterances beginning with a ‘wh-’ word. RP, in contrast, marks these with a simple falling tone. The high rise tone is similar to the low rise, in that the pitch begins in the speaker’s low to mid range on the nucleus, however in the high rise tone the point to which the pitch rises and remains is high in the speaker’s range. This is the unmarked tone for interrogatives which expect a yes or no answer; it may also be used for what Cruttenden calls “involved declaratives [or] wh-interrogatives” (Cruttenden 1994) (he does not make clear what the manner of involvement is likely to be). The final rise, the rise fall, is used for making a contrast, or to show the speaker’s reservation. The version of the rise fall found in Belfast differs from the rise fall found in RP, which falls to the bottom of the speaker’s range abruptly. In Belfast English the rise fall tone rises from a low point on the nucleus of the utterance, to the top of the speaker’s pitch range, and then falls gradually through the rest of the utterance.

Cruttenden does puts forward one particular and significant difference between Belfast and Glasgow English. In general across all the dialects that make up UNB, there is an increase in the use of four particular tones, all of which can be seen to be predominantly rising in nature—a **rise**: a glide up on the nuclear syllable or a jump-up between the nuclear and the following syllable, a **rise-plateau**: a rise with the high level maintained, a **rise-plateau-slump**: a rise with the high level maintained until the last one or two syllables when the pitch declines, and a **rise-fall**: a rise followed by a fall with no plateau in between. Cruttenden states that while all dialects of UNB use these rises more often than RP or SAE, each individual dialect has a tendency to use one or two more predominantly than the others. According to Cruttenden, Belfast English prefers the rise-plateau (encompassing the low rise and the high rise, both of which maintain a higher pitch after the rise), and the rise-plateau-slump (which would be equivalent to the rise fall, as the fall in both cases is gradual, remaining high for most of the tone). By contrast, the inventory of rises in Glasgow English seems to include (and prefer) something equivalent to the simple rise without a plateau or fall, but with the addition of a lengthening of the nuclear syllable. (Cruttenden 1986).

It should be noted as well that there have been suggestions by Samuels (as cited by Cruttenden (1994)) that Western Scottish English does not differentiate intonationally between statements and questions (both presumably using a rising final tone). This

would definitely seem to indicate that the use of rising tones in Glasgow English is significantly more common than in RP or SAE.

1.4 Adaptation of ToBI for other languages and dialects

As seen in the previous section, intonational patterns can vary widely, even across dialects of one language. While ToBI was developed as a standard for prosodic transcription, it was not designed to encompass the intonational patterns of all languages. In point of fact, ToBI's design was based primarily on Standard American English, and can really only be said to cover this dialect of English, plus southern British English and Standard Australian English with any accuracy (Beckman & Ayers 1993). That said, the underlying principles of the ToBI system are not restricted to the description of Standard American English, and the ToBI system has been adopted, with changes, for a number of languages and dialects other than SAE and RP. The advantage that the ToBI system has over whole tone analyses is that its tonal inventory is built up from only two fundamental building blocks, high tones and low tones, which can be combined and manipulated to suit the needs of languages and dialects other than Standard American English. Two good examples of the flexibility that the ToBI system affords to adaptations are GToBI, a ToBI-style system for German, and J_ToBI, a Japanese ToBI system.

The tonal inventory of J_ToBI is much smaller than that of the original ToBI system—in particular the system has only one pitch accent, H*+L, which can align early or late relative to the accented mora (Japanese syllable). This relationship is marked by a diacritic (“<” or “>”) in a similar manner to the marking of downstep in the original ToBI. Downstep itself, in fact, is not marked in J_ToBI, as in Japanese it is a feature of lexical rather than sentence-level intonation. The end of an accentual phrase (which corresponds roughly to a ToBI intermediate phrase) is marked by the system's single phrasal tone, H-, and by either an L% boundary tone (called a “strong” low) or a wL% (a “weak” low). This use of boundary tones at an intermediate level is a departure from their use in the original system. In addition, the pitch at the beginning of utterances or phrases is also marked—either with a %L or a %wL (again, “strong” and “weak” low tones). The end of a full intonational phrase is marked, like in ToBI, by a final boundary tone—in J_ToBI either an H% or HL%. The compound tones formed by J_ToBI's intermediate boundary tones and the full boundary tones are therefore L%H% and L%HL%.

ToBI's original break index system has also been changed for use with Japanese—an adjustment not made by any other adaptations of the system for languages other

than Standard American English. Again, the inventory of break indices is smaller than in the original ToBI, ranging from a break index of 0 (to mark “junctures...common in fast speech processes” (Vendetti 1995)) to a break index of 3 (to mark a “strong degree of disjuncture between words” (Ibid.)). The reason for this is that, as stated by Vendetti (1995) “the question of what is a “word” in Japanese is a difficult one”, and one that J_ToBI does not try to answer, but simply to describe using a system that is as simple as possible.

Two significant changes were made to the original ToBI system to make it more appropriate for the transcription of German—the first tonal, the second theoretical. The complete set of tones used with the GToBI system is identical to the original set, with the addition of a new compound pitch accent: **H+L***. This accent corresponds to movement from high in the speaker’s pitch range to quite low in the range, with the accented syllable aligned with the low point of the tone (the “starred” tone). The movement from high to low in this tone is described as a large jump down, as opposed to a fall or a glide—there does not seem to be any intermediate connecting pitch between the high and low tones (Benzmüller & Grice 1996).

The theoretical alteration made to the system involves the description of intermediate phrases. As mentioned above, in both Pierrehumbert’s intonational description of English, and in the original ToBI system, an intermediate phrase consists of at least one pitch accent and a phrase accent to mark the end of the phrase. In the GToBI system this fundamental description is altered to allow for intermediate phrases without *any* pitch accents. GToBI intermediate phrases “when accentless...are subordinate to a preceding or following accented intermediate phrase within the same intonational phrase” (Grice, Reyelt, Benzmüller, Mayer & Batliner 1996).

There are no available results from reliability testing of J_ToBI, but GToBI achieved comparable results to the original ToBI system (Grice et al. 1996). This would seem to indicate that despite the fact that “ToBI is specifically a transcription system for English, and not...a kind of high tech IPA alphabet for intonation generally” (Ladd 1996*a*), the fundamentals of the system can be applied with success to other languages and dialects.

1.5 GlaToBI—ToBI adapted for Glasgow English

As discussed above, Glasgow English, along with the other urban north British dialects, differs from Standard American English and RP English in its more frequent use of nuclear rises. The original ToBI system has two compound pitch accents which are used to mark movement from a low pitch to a higher one: **L+H*** and **L*+H**, either of

which could be adopted for Glasgow rises. As can be seen from the * diacritic, these tones differ in their alignment with the accented syllable. The L+H* tone corresponds to a movement from low to high just before the accented syllable, with the accent itself aligned with the high. The L*+H tone, on the other hand, corresponds to a low target tone aligned with the accented syllable, followed closely by a movement from that low point to a higher pitch. However, as pointed out by Cruttenden (see section 1.3), the predominant type of rise in Glasgow English is not a step up from an unaccented syllable to the following accented syllable, nor a step up from the accented syllable to the following unaccented one. In fact, the typical rise in Glasgow English is a glide up *on* the accented syllable itself, a movement which is not captured adequately by either of the original ToBI rising tones. To solve this problem the designers of the GlaToBI system removed both original ToBI rises from the tonal inventory and replaced them with just one compound pitch accent: **L*H**. The placement of the * diacritic in between the L and H tones indicates that the stressed syllable does not align with either one or the other, but with the movement *from* one *to* the other. In point of fact, it may be possible that Glasgow English does have tones that rise before or after the accented syllable—as pointed out by Cruttenden (1986) the simple rise preferred by Glasgow English could, in addition to being a glide up, also be a step up occurring from the accented syllable to the next. The system developers did not find any evidence of other types of rise in their investigations; nonetheless this is an element of the system which will be closely examined in this evaluation study.

The second alteration made to the original ToBI system was, like the GToBI adaptation, a theoretical change. While the simple rise found in Glasgow English does not seem to be realised in any way other than that described above, there does seem to be evidence of the rise–fall tone that Cruttenden states is present in Belfast (and therefore possibly Glasgow) English. Cruttenden’s alternate name for this tone, the rise–plateau–slump, is in fact quite an accurate description of the pitch movement of this tone in Glasgow English: the pitch rises at the accented syllable and then remains high until very near the edge of the phrase, where it falls again. The first part of the tone, the rise, is clearly the simple Glasgow English rise, and can be represented by the L*H tone. Unfortunately, the second part of the tone, the high plateau followed by a fall, is not so easily represented. The original ToBI inventory of phrasal tones includes an L phrase accent followed by an L boundary tone (L–L%) which is used to represent final lowering of the pitch, but this combination does not convey the fact that the pitch in Glasgow English remains high before falling. The original ToBI also has an H phrase accent followed by an L boundary (H–L%), which would seem more appropriate, however because of the up–step rule this tone is realised as a level tone

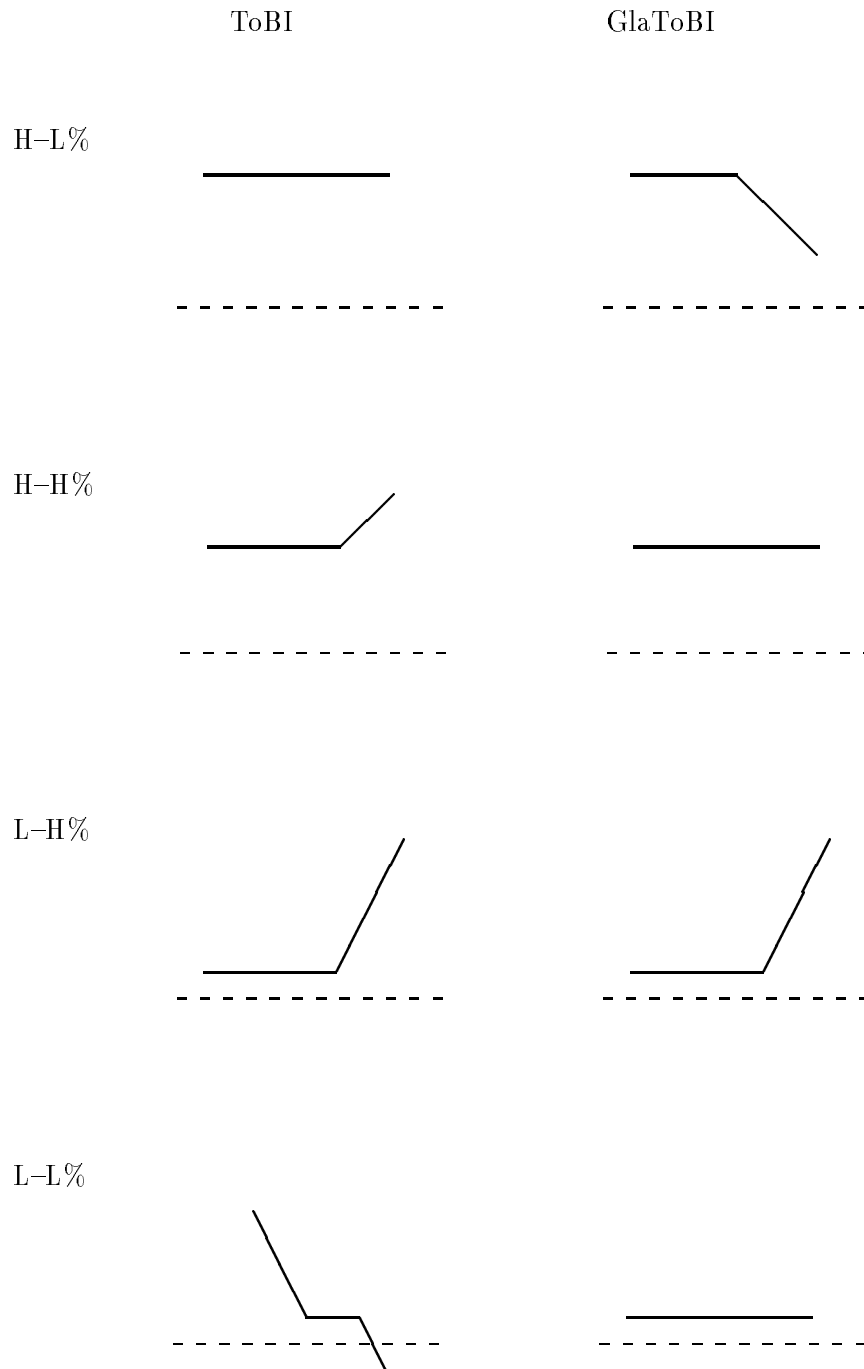


Table 1.1. Stylised representations of pitch final contours in ToBI and GlaToBI

in the mid to high range of the speaker’s pitch. The solution adopted by the designers of the GlaToBI system was to eliminate the intrinsic up–step cuing property of the H phrase accent. This then allowed the compound H–L% phrasal tone to be used in a slightly more intuitive manner, to represent a high–low sequence.

Unfortunately, however, this theoretical departure from the original ToBI required some other, compensatory adjustments to be made to the system. The adoption of H–L% to describe a falling pitch leaves the system without a phrase accent–boundary tone sequence to describe a level tone. Fortunately, the fact that up–step is no longer taken into consideration means that the compound H–H% can now be used simply to represent a high pitch accent followed by an equally high boundary (i.e. a mid to high plateau), rather than a high pitch accent followed by a higher boundary. Following suit with the low tones, the sequence L–L% is used in the GlaToBI system to represent a low pitch accent immediately preceding an equally low boundary tone—a low plateau. Table 1.1 shows stylised pitch contours for the GlaToBI realisations of the four possible combinations of pitch accent and boundary tone as compared to the original ToBI f0 realisations of the same compound tones.

Table 1.2 compares the final GlaToBI tone menu to the original ToBI tone menu.

	ToBI	GlaToBI
Pitch accents	H*	H*
	L*	L*
	L+H*	L*H
	L*+H	L*H
	!H*	!H*
	L+!H*	L*!H
	L*+!H	L*!H
Phrasal tones	H–	H–
	L–	L–
	H–L%	H–L%
	H–H%	H–H%
	L–H%	L–H%
	L–L%	L–L%

Table 1.2. ToBI and GlaToBI tone menus.

1.6 The Map Task corpus

As mentioned above, the creation of the original ToBI system was driven in part by the development of new techniques to elicit and record large corpora of spontaneous speech. One such corpus, the HCRC Map Task corpus was the main motivation behind the development of the GlaToBI system. This corpus is a large digital database of spontaneous, task-oriented dialogues. Previously much of the work in the speech sciences has been done on read or scripted speech, despite the fact that there are many aspects of spontaneous speech which are not present in read speech. This bias arises due to the problems involved in controlling the content and the context of spontaneous speech—it is difficult to elicit a specific utterance in a specific manner without directly asking the speaker for it, at which point, whether the utterance is read or repeated, the speech ceases to be spontaneous. The HCRC corpus was recorded in a context which encourages unscripted/spontaneous speech but in which the content and context of the speech can be manipulated to the experimenter’s specific design. Each of the 128 conversations in the corpus revolves around the speakers’ performance of the Map Task, an experiment which involves two participants, each with their own copy of a map. One of the participants is designated the instruction giver—on this person’s map there is a route marked—while the other participant is the follower. Neither participant can see the other’s map. The task required of the participants is to reproduce the route that is marked on the instruction giver’s map onto the follower’s blank map—by conversation only. More involved and elaborate conversation is encouraged by the fact that there are deliberate mismatches between landmarks on the two maps: some are present on one map but not on the other, or present once on one map and twice on the other, or the same landmark might have different names on each map. The participants are made aware that their two maps might be different in some respects, but are not given details of the mismatches before they begin the experiment. They are therefore forced into more complicated conversation, both to try and understand the situation, and to try and solve the problem. The Map Task lends itself to fairly careful control by experimenters, without forcing them to interfere directly in the conversation. Naturally as the maps are designed by the experimenter, the characteristics (phonetic, phonological etc.) and placement of the landmark names can be altered to suit the study. Additionally, in the HCRC corpus there was further manipulation of the environment by varying the familiarity of the two speakers to each other, and by controlling their ability to make eye contact by placing or removing a barrier. In this fairly straightforward manner both the context—the specific task, the familiarity of speakers, their ability to make eye contact—and the content—the names and placement of landmarks—were

successfully controlled by the HCRC experimenters, while maintaining the spontaneity of the participants' speech. (Anderson et al. 1991).

The fact that the Map Task corpus is made up of spontaneous speech is important to note—both for the future study of intonation, and, more specifically, for this evaluation study. There is still some debate as to whether intonation shows the marked differences in read and spontaneous speech that are seen phonetically or syntactically, however it should nevertheless be kept in mind that all of the data for this study comes from a corpus of unscripted speech, as this could play a role in the outcome of the evaluation.

Chapter 2

Experiment

The following evaluation study was designed primarily to test the new GlaToBI system in terms of two of the original ToBI system goals—reliability and learnability. In addition, this experiment should also give some indication as to how GlaToBI performs in terms of a third ToBI goal—that of the coverage or comprehensiveness of the system.

2.1 Subjects

The subjects for this experiment were three of the four designers of the GlaToBI system and five novice transcribers who had not been previously exposed to the system. Four of the five novices met the pre-test criterion (described in section 2.4) and were allowed to go on to perform the experiment proper, giving a total of 7 transcribers. The subjects were staff and students from the University of Edinburgh, Queen Margaret College (Edinburgh), and Glasgow University. All were native speakers of English, with one of the four final non-expert transcribers a native speaker of Glasgow English.

Both expert and novice subjects were used in this experiment because, in assessing the reliability of a ToBI-style system, the performance of both types of transcriber must be taken into account. One of the primary purposes of the experiment was to ascertain the level of agreement achieved by users of the system. If this had been the only aspect of the system to be examined, the transcription of those users who were already familiar with the system (i.e. the developers) would have been sufficient. However, like the original ToBI system, GlaToBI was developed to meet other, additional criteria as well—in particular “learnability in a relatively short time” (Silverman et al. 1992). For the purpose of testing this part of the system, volunteers with varied backgrounds in linguistic studies but with no previous experience with GlaToBI were asked to perform the same tasks as the experts.

Table 2.1 below gives the background in linguistics of both the experts and the novices. Table 2.2 shows all the subjects' dialectal backgrounds.

	Not experi- enced with prosodic transcription	Experienced with prosodic transcription	Experienced with ToBI	GlaToBI system designer
JK				•
PT				•
MA				•
RC			•	
JS		•		
SM		•		
EB	•			

Table 2.1. Linguistic background of subjects

	Standard American	Southern British	Irish	Glasgow English
JK	•			
PT			•	
MA		•		
RC		•		
JS				•
SM			•	
EB	•			

Table 2.2. Dialectal background of subjects.

2.2 Data

The data used to evaluate the GlaToBI system was taken from the HCRC Map Task corpus. Each utterance in the corpus has been digitally recorded and orthographically transcribed, and word endings have been labelled. While there are some instances of missed or mis-transcribed words, and some lack of consistency as to the transcription of contractions and compound words, for the most part the orthographic transcriptions are quite accurate. In addition to files containing the speech and orthographic transcription, each utterance also consists of an f0 file, containing the utterance's digital

pitch contour, and a file containing the speech waveform. In addition, for the purpose of this experiment, three blank files were also created for each of the utterances to be used in the study: a tone file, a break index file, and a miscellaneous file.

Each dialogue in the corpus has been divided up into utterances—essentially one turn from one participant in a Map Task dialogue—which can vary in length from one to twenty words or more, and which are coded according to the conversation, speaker and placement in the dialogue. Because of the variability not only of the length of the utterances but also of the quality, the 50 utterances used for this evaluation study were taken from a larger group of 150 randomly chosen utterances, allowing unsuitable utterances to be eliminated. Those utterances considered to be unsuitable were i) any which consisted of only one word (i.e. “umhum” or “okay”), as the intonational movement in these utterances was generally negligible, ii) any in which the orthographic transcription did not completely correspond to the speech, and iii) any in which the speech was extremely unclear. This last criterion eliminated all utterances in which the speakers themselves were inaudible, and any in which the recording equipment had in some way interfered with the speech (there were a number of utterances in which the recording equipment had picked up environmental noise which obliterated the speech, and two utterances in which there was mains feedback which prevented the f0 tracker from finding a pitch curve). Of the 50 utterances chosen, only 41 were transcribed completely by all the transcribers (due to time constraints). A further 11 were not analysed because they were completed with a high number of grammatical errors—while theoretically the transcriber can use any combination of tones and break indices to describe an utterance, there are some transcriptions which are grammatically not allowed: word or phrase boundaries without correctly corresponding break indices, missing phrasal tones, illegal combinations of tones etc. In total, therefore, the subjects transcribed 30 utterances, comprised of 273 words. The length of each utterance ranged from 2 words to 19 words, averaging approximately 9 words per utterance.

2.3 GlaToBI

2.3.1 Presentation

The GlaToBI presentation as it appeared to the transcribers is as shown in figures 2.1 and 2.2. (It should be noted that this is an Entropics *waves+* presentation of a ToBI-style system—this type of system is also compatible with other computer and non-computer platforms.) There are three main sections to this presentation: the speech waveform (a), the fundamental frequency contour (b), and the label files (c),

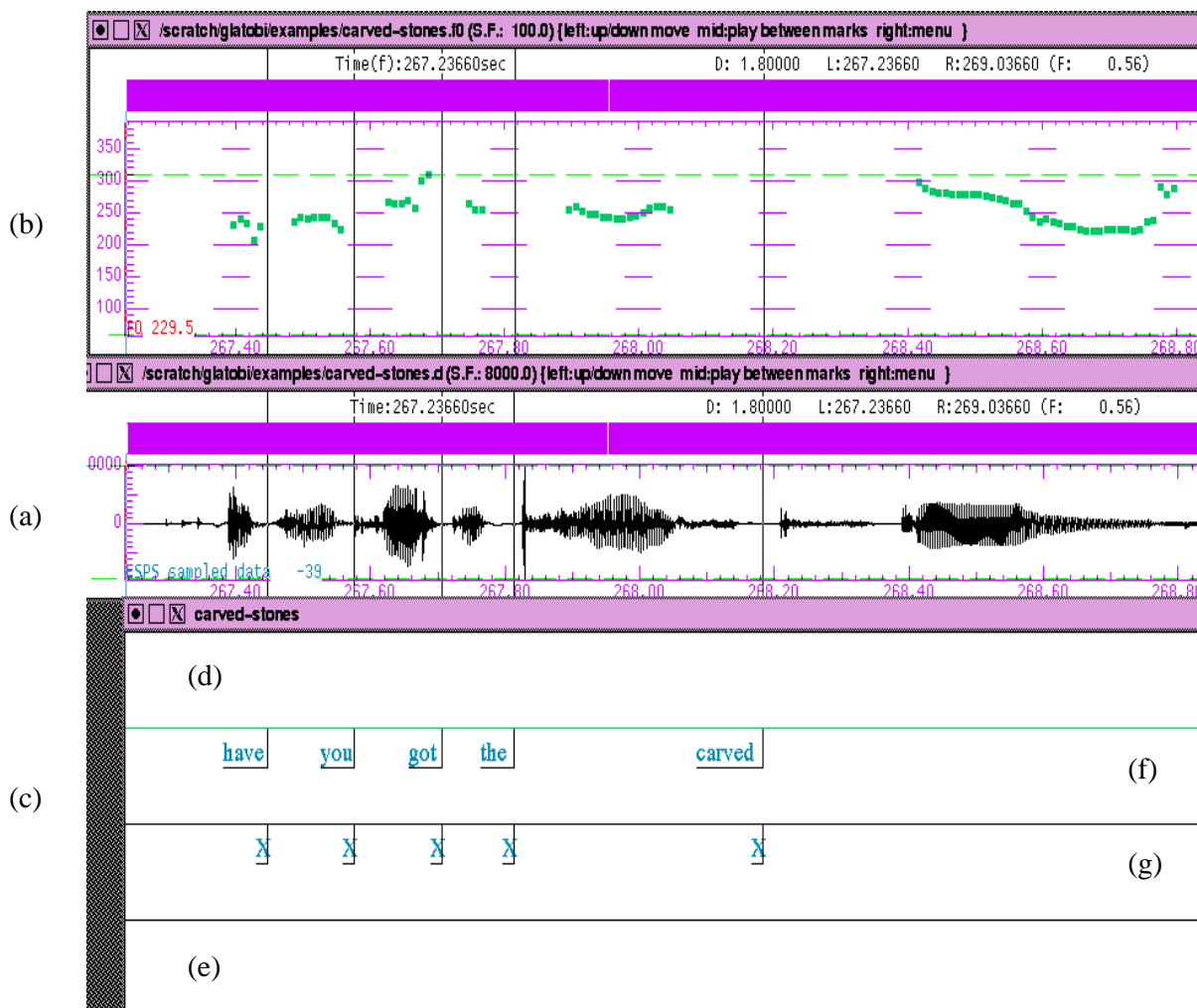


Figure 2.1. Blank GlaToBI transcription files

all of which align themselves temporally to the speech waveform. Two of the tiers in the label file are presented completely blank to the transcribers—these are the tone tier (d) and the miscellaneous tier (e). The two remaining tiers in the label file are the orthographic tier (f) and the break index tier (g). In the orthographic tier the end of each word (the backwards “L”) is lined up in time with the corresponding end-of-word boundary labelled in the speech waveform. Because there is no option as to the placement of break indices—there must be one inserted at the end of every completed and incompleted word—the break index tier is presented with dummy break indices (“X”), again lined up with the corresponding end-of-word boundary on the waveform. When the user does the transcription, they simply have to replace each “X” with the

appropriate break index (i) and insert the tones in the appropriate place in the tone tier (h) (figure 2.2).

Each tier in the label file except for the orthographic tier has a pull-down menu to allow ease of transcription. This limits the user to the symbols which are presented in the menus, which should prevent “illegal” symbols (i.e. those not appropriate to the system) from being used. The miscellaneous tier also allows the user to insert comments, in addition to using the available markings for laughter, audible breaths, coughs etc. This facility is useful for those who may find the system slightly restrictive,

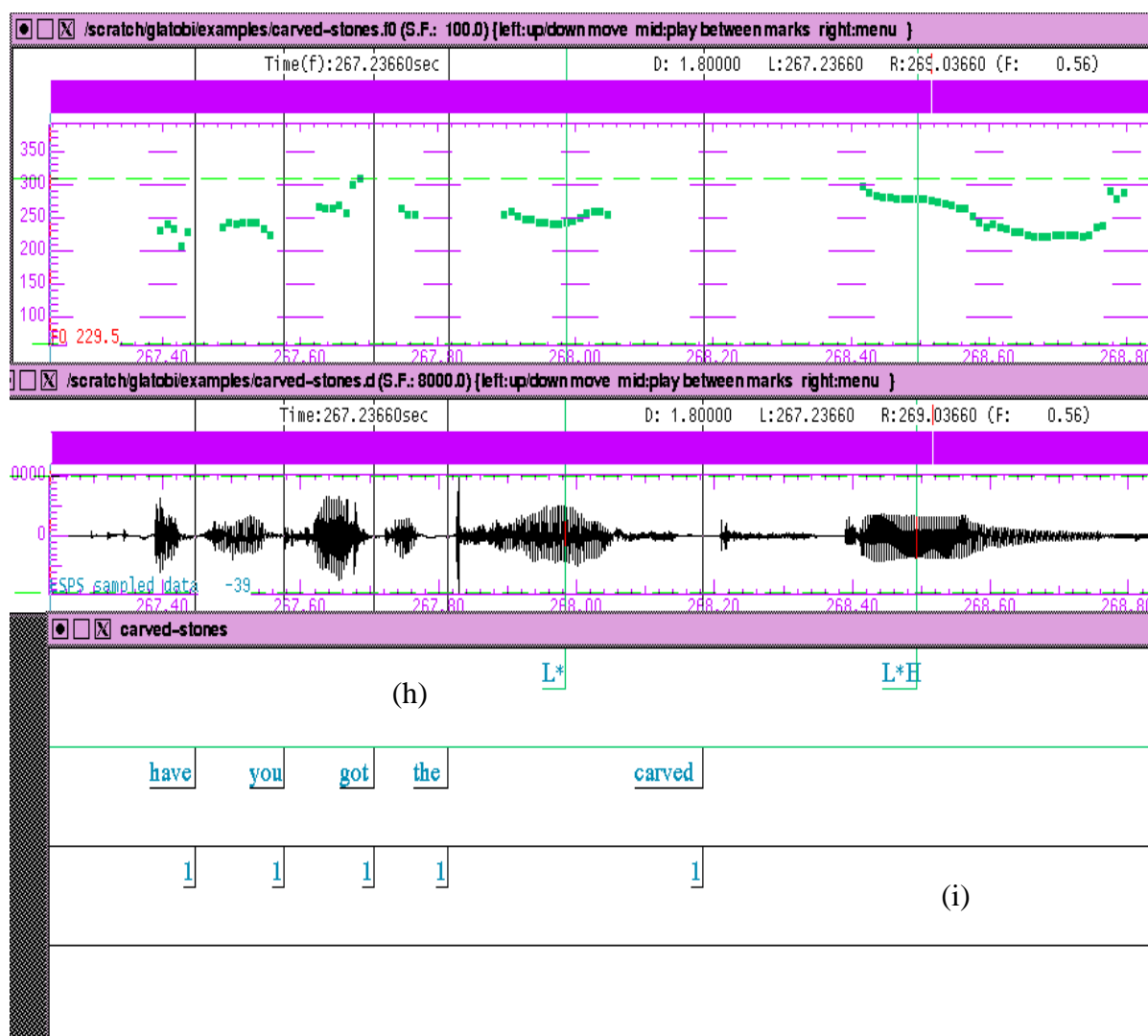


Figure 2.2. Completed GlaToBI transcription file.

as it allows them to explain what may seem like inappropriate transcriptions, or point out problem areas to other transcribers. Each menu has three modes: **INSERT**, **REPLACE**, and **DELETE**, which allow the transcribers to place, move and correct tones and break indices with relative freedom.

2.3.2 Software support

Two support tools were used in this evaluation study to further simplify the use of the system. The first tool invokes *x-waves* once and allows a list of utterance names to be sent to the system. The speech waveform, f0 contour and label files are loaded one utterance at a time into GlaToBI for transcription, while at the same time all three are copied to the user's local directory—this allows multiple users to have access to the same utterances, without having access to the other users' transcriptions. The second programme collects all of the utterances it finds in a specified directory and sends them as a list to the first programme, allowing the user access to GlaToBI and the all the utterances for transcription with only one command. This second simplification was especially useful for this evaluation study, as some of the subjects were unfamiliar with the *waves+* environment.¹

The original ToBI system was designed with a software tool which acted as a grammar checker for transcriptions. This checker allowed the users of the original ToBI to “correct some ‘slips of the mouse’ and misunderstandings of details of the system” (Pitrelli, Beckman & Hirschberg 1994). Unfortunately this tool was unavailable to the users of GlaToBI. The effect of this will be discussed further in section 3.4.2.

2.4 Training

The four novice transcribers were given an intensive one day training session before participating in the experiment proper. To avoid any particular bias in the novices' transcriptions compared to the experts', the session was run by the one GlaToBI designer who did not participate in the experiment proper. The session consisted of two lectures which alternated with practical lab sessions in which the trainees could become familiar with both a ToBI-style transcription, and with the new GlaToBI system. The material covered in the lectures included background information on intonation, the autosegmental metrical theory of intonation, phrasing and break indices, the development of the original ToBI system, and the system's adaptation for Glasgow English. In addition, potential problem areas were highlighted, specifically those which arise

¹Thank you to Norman Dryden and Matthew Aylett for providing the programmes.

when transcribing spontaneous speech. The example utterances were taken both from the training material provided with the original ToBI and from the Map Task corpus itself. The original ToBI training utterances were pre-transcribed; in the case of the Map Task material, 16 utterances, ranging in length from 4 to 16 words and chosen for their intelligibility, were transcribed by two of the GlaToBI designers. The trainees were paired up for the practice sessions to facilitate collaboration and a more thorough understanding of the system, and throughout the day they were encouraged to consult the system designers (two were present for all practice sessions) or the notes provided for them if anything was unclear. At the end of the training day a pre-test was given to make sure that the system had been fully understood. Again the utterances used for this test came from the Map Task corpus, and were pre-transcribed by two expert transcribers for use in comparison with the trainee transcriptions. The pre-test contained 8 utterances in all, with examples of most phenomena covered in the training lectures. The novices were asked to work separately on these transcriptions without consulting each other, although they could consult their notes. In order to be considered to have passed the pre-test the novice transcribers had to complete transcription of the 8 utterances with no more than two grammatical errors per utterance, and with a reasonable overall transcription compared to the expert pre-transcription (as judged by the designer running the training session).

2.5 Experiment proper

All subjects (novice and expert) were tested individually within the week following the novices' training session. During the test, the subjects were not permitted to consult with each other or with the experimenter, however the novice subjects were informed that they could use their training notes and any personal notes that they had made during the training session. (One expert transcriber was also provided with a copy of the training notes for reference.) The subjects were encouraged to take breaks whenever they felt tired or found their attention wandering. There was no time limit put on the test—most subjects finished transcription of 30–40 utterances in one day (8 hours approximately, including breaks), with one expert and one novice transcriber spreading the transcription out over two consecutive days.

2.6 Methods of data collection and analysis

2.6.1 Pairwise analysis

Two methods were used to assess the level of agreement reached by the subjects in this evaluation study. The first method repeats the pairwise analysis used in evaluating the original ToBI system (Pitrelli et al. 1994), and the version developed for German (Grice et al. 1996). This method, instead of comparing the labels of individual transcribers against the group, compares the labels of each transcriber against the labels of every other transcriber, for that particular aspect of the utterance. For instance, if four out of five transcribers label a word H*, and the fifth labels the same word L*H, the level of agreement is **not** considered to be 4 out of 5. Instead, comparing pairwise, the level of agreement is found to be 6 out of 10—6 pairs out of a potential 10 pairs of transcribers who agreed with each other as to the exact label to be placed on the word. In a description of this method of analysis, Pitrelli et al. (1994) state that “the basic unit for measuring agreement is the *transcriber-pair-word*”, or the set of two labels given to one word by a pair of transcribers, and that “the measure of inter-transcriber consistency is then the percentage of transcriber-pair-words exhibiting agreement on a particular element in the transcription.” For the example above, therefore, the percentage agreement would be 60%, or 6 out of 10. Pitrelli et al. (1994) make the claim that this method of analysis is stringent because it does not return the deceptively high 80% agreement which would be given by an analysis which stated simply that of 4 out of 5 transcribers agreed.

2.6.2 Kappa

The kappa statistic is the second method used to assess the reliability of transcriber agreement. Again, as above, this statistic measures agreement among the transcribers by examining their coding pairwise, but in addition, this statistic corrects for expected chance agreement (Carletta 1996). Carletta claims that while the pairwise agreement used by Pitrelli and Grice is relatively reliable “in that it produces one figure which sums reliably over all coder pairs”, it does not take into account the number of possible categories available to the transcriber at any one time. This means that while the results obtained may be good, there is no way of knowing if they are better than what would be achieved by chance over the same number of categories. As Carletta points out, “the amount of agreement we would expect coders to reach by chance depends on the number and relative proportions of the categories used by the coders”: if there are only two available categories, both of which having an equal chance of occurring,

then two coders using these categories would agree 50% of the time; if the number of categories is increased to four, the chance agreement would be 25%. Carletta suggests that the kappa statistic, which takes into account both the number and proportion of categories, *and* chance agreement would be more useful in judging the reliability of a system such as GlaToBI.

The kappa coefficient of agreement (K) is “the ratio of the proportion of times the raters [transcribers] agree (corrected for chance) to the maximum proportion of times that the raters [transcribers] could agree (corrected for chance)” (Siegel & Castellan Jr. 1988):

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times that the transcribers agree (i.e. the percentage agreement from the pairwise analysis above) and $P(E)$ is the number of times that the same number of transcribers would be expected to agree by chance. K ranges from 0, where there is no more agreement than what would be expected by chance, to 1, where there is complete agreement among the transcribers (Siegel & Castellan Jr. 1988). Both Carletta and Siegel state that in some cases it is useful to find out the significance of the value of K —i.e. whether, and by how much the value of K was greater than that which would occur by chance. However, Carletta states that “interpretation of the scale of agreement is possible” and more important. She suggests that values of $K > .8$ indicate good reliability, with values of $.67 < K < .8$ indicative of possible reliability.

The important aspect of this statistic to note in relation to this particular evaluation study, however, is the notion of independence. Siegel & Castellan Jr. (1988) put forward the kappa statistic as appropriate for use when objects are being rated by a number of judges, and the objects are not ranked or ordered, but simply put into categories. The example they give is of a group of psychologists diagnosing a number of patients—in this case each psychologist categorises each patient *independently* of all the other patients, and of the other raters. The effect that this has on the outcome of this particular study will be discussed further in section 3.2.2.

2.6.3 Collection of data

Of the 50 utterances presented to the transcribers, 30 were completely and grammatically transcribed by all 7 subjects. 273 words were transcribed by all the subjects, totalling 1911 individual transcriptions for accent tones, 1911 for phrasal tones, and 1911 for break indices. In a pairwise analysis, this totals 5733 transcriber–pair–words (273 words times 21 transcriber–pairs), for each of accent tones, phrasal tones and

break indices. Analyses (pairwise and kappa) were done on all utterances individually, and on all utterances together. In addition, the rate of agreement within each group of transcribers (expert and novice) was examined and compared with the rate of agreement over all transcribers. For all analyses the following elements were examined:

1. pitch accents:

- a) rate of agreement as to the presence or lack of accent
 - categories: yes (accent is marked), no (no accent marked)
- b) rate of agreement as to the accent marked
 - categories: H*, L*, L*H, none (no accent).

2. phrasal tones:

- a) rate of agreement as to the presence or lack of tone
 - categories: yes (tone is marked), no (no tone marked)
- b) rate of agreement as to the tone marked
 - categories: H-H%, H-L%, L-L%, L-H%, H-, L-, none (no tone)
- c) rate of agreement as to the strength of the boundary
 - categories: full boundary (boundary tone), intermediate boundary (phrase accent), none (no tone marked)

3. break indices:

- rate of agreement as to break index marked
 - categories: 0, 1, 2, 3, 4, none (no break index marked)

A note should be made about the use of *none* as a valid category in both sets of tonal analyses (b) and in the break index analysis. As in this experiment, both Pitrelli et al. (1994) and Grice et al. (1996) divide their tonal analyses into two elements: the presence or lack of a tone on a word, and the decision as to which tone in particular was selected. The analysis of agreement on presence of tones is done the same way as in this study—by examining all tones marked on all words by all transcribers. However, in their analysis of agreement as to the exact tone marked, both Pitrelli and Grice examine only those words that two transcribers agreed should have an accent or a boundary tone. The problem that arises from this method comes when two transcribers from the group label a word as H* (for example) and the rest of the group leave the word unlabelled. In this situation two transcribers have agreed that the word is accented,

and, following the method of Pitrelli and Grice, their agreement should be examined further. Both transcribers have labelled the word in question H^* , therefore the rate of agreement as to which tone is marked should be 100%. However, if there are more than three transcribers in the group, these two who have marked the word as accented are in the minority: if the group of transcribers totals 10 (for instance), then the pair who marked the accents are only one pair out of 45 pairs of transcribers, and their agreement of 100% is rather a false positive. While it is valid to say that the two transcribers agreed with each other, they did not agree with the rest of the group; their agreement as to which tone is marked is no more representative of 100% agreement of the group than it would be if the other eight transcribers labelled the word with L^*H instead of no tone at all. In fact, because they include all labels marked, it is the ToBI and GToBI analyses of *overall* agreement in each category rather than their analyses of choice of tone which correspond to the analyses of choice of tone in this study of GlaToBI, and which will be used in any comparisons made.

As far as the break index analysis is concerned, while technically there is not an option for no break indices to be placed, there were a number of transcribers who omitted occasional break indices. While this does constitute a grammatical error, the number of break indices omitted was small enough to justify analysing those utterances with missing break indices, and adding a null category to deal with them.

Following the analyses performed on the original ToBI system, the downstepped pitch accents: $!H^*$, $!L^*!H$, and the downstepped phrasal tones: $!H-H\%$, $!H-L\%$, $!H-$, were all collapsed into their respective non-downstepped counterparts (i.e. $!H^*$ was considered to be H^*). In addition, break indices with the ‘p’ disfluency diacritic, and with the ‘-’ uncertainty diacritic were also collapsed into their unmarked counterparts.

In this study phrasal tones were analysed three times. The first two analyses—rate of agreement as to the presence of a phrasal tone, and rate of agreement as to the actual choice of tone—were the same as the analyses performed for pitch accents (and as the phrasal tone analyses done on the original ToBI system). The third, following the analysis by Grice et al. (1996) for GToBI, examines the rate of agreement as to the strength of the phrasal tone (i.e. either full—boundary tone—or intermediate—phrase accent).

Chapter 3

Results, Analysis and Discussion

3.1 Pairwise analysis

3.1.1 Results

Table 3.1 shows the results of a pairwise analysis of transcriptions totalled across all utterances. Overall, the level of agreement as to which pitch accent was present on a word was the lowest across all transcribers (62% for novices, 69% for experts), while the rate of agreement as to the presence or lack of phrasal tones was extremely high (92%–93%). The novice transcribers did not differ greatly in their level of agreement from the expert transcribers, although in general they tended to agree slightly less. The only aspect of the system in which the novices did make more consistent choices than the experts was the break index tier (77% as compared to 72%).

3.1.2 Discussion

Three aspects of the pairwise results are worth examining more closely. The very high rate of agreement for the presence or lack of phrasal tone on a word (and, to a slightly lesser extent, the rate of agreement as to the choice of phrasal tone) is in fact deceptive. Included in the category of phrasal tones are both phrase accents and boundary tones. However, while the placement of some phrase accents will depend on how the transcriber believes the utterance is sub-divided, the decision as to whether or not to place the final phrase accent and boundary tone combination is dictated by the structure of the system. Although there may be some exceptions when transcribing spontaneous speech, by rule all utterances, as complete intonational phrases, should end with a phrase accent–boundary tone compound; in addition, as a phrasal tone marks the *end* of a phrase, an utterance cannot begin with a phrasal tone, or have

phrasal tones on two consecutive words. Unless the transcriber has misunderstood or forgotten these rules there will be very little choice, and therefore little variability, in the placement of phrasal tones.

The rate of agreement as to the choice of pitch accent was affected in the opposite direction—the overall rate of agreement was nearly 30% lower than the rate of agreement as to the presence or lack of phrasal tones, and nearly 10% lower than the agreement as to choice of break index. In fact, the reason for this low consistency is,

	All Transcribers (%)	Experts (%)	Novices (%)
Pitch accents			
a) Rate of agreement as to the presence or lack of pitch accent.	78	81	75
b) Rate of agreement as to the choice of pitch accent.	65	69	62
Phrasal tones			
a) Rate of agreement as to the presence or lack of phrasal tone.	92	93	92
b) Rate of agreement as to the choice of phrasal tone.	82	82	82
c) Rate of agreement as to the strength of the phrasal tone.	88	88	88
Break indices			
Rate of agreement as to choice of break index.	74	72	77

Table 3.1. Table of results for pairwise analysis.

like the high agreement in the use of edge tones, due to the structure of the system. While the placement of both phrasal tones and break indices is dictated at least in part by the theories underlying the system (i.e. both a boundary tone and break index 4 must go at the end of an intonational phrase), the labelling of pitch accents is the one aspect of the system where the transcriber has the most freedom. With few exceptions, any pitch accent may be placed on any word in any sequence, which means that the agreement between transcribers will likely be lower than if there were restrictions on the use of pitch accents.

As mentioned above, the level of agreement within the novice transcribers was generally slightly lower than the agreement within the experts. This difference in performance is not unexpected: all of the novice transcribers were new to the system, all but one were new to ToBI-style transcribing as a whole, and all of them received only one day's training. The experts, on the other hand, were all system developers, and as such should have had a more intuitive understanding of the system. However, the differences in the levels of agreement are very small (at most only 7%), indicating that with the appropriate training GlaToBI can be quickly and easily learned, up to very high levels of consistency.

3.1.3 Results of pairwise analysis compared with ToBI and GToBI

Although the scale of this evaluation study is much smaller than the studies carried out on the original ToBI system, and the version developed for German, it is useful to compare previous system results with the levels of agreement reached by GlaToBI. The exact numbers involved in each study were as follows: ToBI—489 word corpus, 26 transcribers; GToBI: 733 word corpus, 13 transcribers; GlaToBI—273 word corpus, 7 transcribers. There are some other differences in the studies that should be noted: the original ToBI analysis examines phrase accents and boundary tones separately, while both GToBI and GlaToBI treat them as one category with two different strengths. In addition, the evaluation of the level of agreement as to the choice of break index in the original ToBI was done without break index 4s on utterance-final words, on the basis that they are obligatory. The current study *includes* break index 4s in the calculation of the same agreement, on the basis that utterance-final words in spontaneous speech may not in fact require a break index of 4, as they do in read speech. Table 3.2 below lists the GlaToBI results and the test results from ToBI and GToBI. All percentage agreements in this table were calculated with diacritic-marked tones (downstep, disfluency etc.) collapsed into their corresponding non-diacritic-marked tones.

Overall, while the rates of agreement for GlaToBI tended to be slightly lower than

	ToBI (%)	GToBI (%)	GlaToBI (%)	
Pitch accents				
a) Rate of agreement as to the presence or lack of pitch accent.		80.6	87	78
b) Rate of agreement as to the choice of pitch accent.		72.4	74	65
Phrasal tones				
a) Rate of agreement as to the presence or lack of phrasal tone.	(phrase accent) (boundary tone)	89.8 93.4	n/a	92
b) Rate of agreement as to the choice of phrasal tone.	(phrase accent) (boundary tone)	85.3 90.0	86	82
c) Rate of agreement as to the strength of the phrasal tone.		n/a	86	88
Break indices				
Rate of agreement as to choice of break index.		70.4	n/a	74

Table 3.2. Comparison of pairwise results.

those achieved in the ToBI and GToBI tests, the difference was never more than 7%. In addition, the very high levels of agreement for presence of phrasal tones, and low levels for choice of pitch accent seen in the analysis of GlaToBI can also be seen in both ToBI and GToBI's results. This would seem to be a further indication that these extreme levels of agreement are the result of the structure of the systems, rather than of any changes made specifically for Glasgow English. Taking into account the differences between the design of the GlaToBI study and the two previous studies, it is clear that the levels of agreement reached by transcribers using GlaToBI are comparable to those seen in both the original and the German ToBI systems.

3.2 Kappa analysis

3.2.1 Results

The results of the kappa analysis of the transcriptions totalled over all utterances are as found in table 3.3, which follows.

Carletta proposes that values of K which may be considered to be reliable should fall between 0.67 and 0.80 or higher (see section 2.6.2). Keeping this in mind, the results of this kappa analysis seem at first glance to indicate that the GlaToBI system is not as stable as the pairwise analysis suggests. The highest values of K achieved were for the rate of agreement as to the presence of phrasal tones: 0.78 for all transcribers, 0.81 for experts and 0.75 for novices, and for the rate of agreement as to the strength of the phrasal tone: 0.70 for all transcribers, 0.71 for experts and 0.66 for novices. While these values do fall within Carletta's reliability scale, it has already been seen that the analysis of phrasal tones produces deceptively high results. The results (for all transcribers) for the rest of the analyses ranged from $K = 0.41$ (rate of agreement as to choice of pitch accent) to $K = 0.55$ (rate of agreement as to presence of pitch accent/rate of agreement as to choice of phrasal tone), none of which come near the bottom end of the reliability scale. As will be seen in the following section, however, these low values are not indicative of a flaw in the GlaToBI system, but in the method of evaluation itself.

3.2.2 Discussion

As mentioned in section 2.6.2, the kappa statistic was chosen as an additional method of analysis for this evaluation study because it takes into account the possible *chance* behaviour that would be expected from the same transcribers and the same number of

categories. In other words, kappa evaluates the level of agreement that could be expected by chance, and gives a measurement of how much better than chance the levels of agreement actually are. However, as discussed in that section, the kappa statistic depends on independence—the judges must be able to categorise each object independently of every other categorisation they have made. Unfortunately, the categorisation of words in a ToBI-style transcription is not made independently: the placement of a

	All Transcribers	Experts	Novices
Pitch accents			
a) Rate of agreement as to the presence or lack of pitch accent.	0.55	0.60	0.48
b) Rate of agreement as to the choice of pitch accent.	0.41	0.47	0.36
Phrasal tones			
a) Rate of agreement as to the presence or lack of phrasal tone.	0.78	0.81	0.75
b) Rate of agreement as to the choice of phrasal tone.	0.55	0.58	0.52
c) Rate of agreement as to the strength of the phrasal tone.	0.70	0.71	0.66
Break indices			
Rate of agreement as to choice of break index.	0.50	0.50	0.51

Table 3.3. Table of results for kappa analysis.

boundary tone on a word, for example, automatically precludes the placement of another boundary tone on the following or preceding word. The placement of that same boundary also dictates the placement of a phrase accent on the same word, and the placement of at least one pitch accent on the preceding words. In fact, as can be seen, far from being independent objects for categorisation, every word in an utterance is intonationally dependent on the words around it. Statistically this leads to category skew: the placement of words predominantly in one category and not in another (for instance: in the calculation of the number of phrasal tones on words in an utterance, almost no words other than phrase final ones will be marked with a boundary tone). The effect of this skew can be seen in the very low values of K seen in the kappa analysis results above.

The confusion surrounding the use of this statistic to analyse the results of the GlaToBI evaluation arises from the “building block” nature of Pierrehumbert’s method of intonational description. Pierrehumbert’s work is based on the idea that intonation is a sequence of salient pitch events, joined together by sections of non-salient pitch contours. While this is not a description of intonation in terms of the whole tune of the utterance, the pitch events do not occur in isolation from each other, as would be necessary for a kappa analysis to be appropriate. Prosody acts at the level of the whole utterance, not on just the words in the utterance. Therefore, while pitch events may be associated with only one word in an utterance, it is the relationship between one event and another, and one event and all of the events, that is intrinsic to the understanding of the intonational message.

3.3 Evaluation of GlaToBI innovations

In adapting the original ToBI for Glasgow English several fundamental changes were made to the system. Naturally, while the main goal of this study is the evaluation of the GlaToBI system as a whole, it is informative to examine the performance of these altered aspects of the system in more detail. The first of the two main changes made was the addition of the compound accent L*H to mark a Glasgow English pitch rise, and the subsequent removal of the original rising pitch accents, L*+H and L+H*. These two original tones were not used in the GlaToBI inventory because they describe rises which occur either before or after the stressed syllable—examination of Glasgow English by the GlaToBI designers showed only evidence of a rise which occurs *on* the stressed syllable. However, there are rises other than an L*H-type rise which are used in urban north British dialects, including some which involve a pitch movement up to or down from a stressed syllable. An examination of the use of the L*H accent by the

transcribers in this study could give some indication as to whether Glasgow English does use only one type of rise, or if the addition of other types of rise should be made to the tonal inventory.

Of the 273 words transcribed by the 7 subjects in this study, 105 were labelled L*H by at least one transcriber. Of these 105, only two were labelled L*H by all seven transcribers (in general, however, pitch accents are not unanimously labelled—only one word was labelled unanimously L*, and one other unanimously H*). The confusion over the use of L*H on the remaining 103 words can be divided into two types: confusion between L*H and no accent, and confusion between L*H and other accents. These two types of confusion can essentially be seen as two different issues. The first, where there is a division between transcribers labelling a word L*H and those not marking any accent at all, can be seen simply as an issue of perception of prominence—it is a fact of intonational transcription that some levels of prominence may not be heard in the same way by all transcribers. The second issue, however, that of confusion between L*H and different accent labels, gives a good indication of the use being made of the GlaToBI rising pitch accent. 90 of the 105 words labelled L*H were also marked with a different accent by at least one transcriber (included in this count were words which had not been unanimously marked as having a pitch accent—i.e. those words with L*H, another accent, and no accent marked by different transcribers). 40% of these words were confusions of L*H with both L* and H*, which, given the nature of the L*H tone is not an unexpected result. An accented syllable labelled with L*H aligns with the *movement* between the L and H tones, rather than either the high or the low tone. This means that if L*H is confused with tones to which the accented syllable aligns directly (i.e. L*, H*), it is equally likely to be confused with an accented syllable aligned with a low tone (L*) as an accented syllable aligned with a high tone (H*). More interesting, however, is the fact that 43% of the words labelled L*H were confused with H* only, indicating disagreement between the transcribers as to whether the tone was a rise, or a simple H aligned with the accented syllable. Examining this information, the question that naturally arises is what, if anything, is causing these confusions—is there a phonetic or phonological cause, or have the transcribers been conditioned into the use of one accent over another in certain situations? While this study is not large enough to go into a detailed statistical analysis of the use of L*H, an impressionistic examination of the utterances and their corresponding transcriptions could indicate whether more extensive studies should be done.

When L*H is used unanimously, or confused with only “no accent”, it tends to be in the nuclear position in the utterance. As well, in general it is followed by an H–phrase accent, and usually by an H–L% phrase accent–boundary tone combination. In

other words, the times when there is least confusion regarding the choice of the L*H accent are those when the pitch most resembles the L*H H-L% structure—i.e. the prototypical Glasgow English final rise fall.

However, outside of the use of L*H in prototypical (or stereotypical) constructs, there do not seem to be any factors which are conditioning the confusions between L*H and other accents (examples of conditioning factors might be the confusion of L*H with H* only in pre-nuclear positions, or with L* only after a high phrase accent). An examination of those words which were labelled L*H and H* and/or L* shows that confusion occurred in both nuclear and pre-nuclear positions. In all situations the accents were followed with relatively equal likelihood by low and high tones, and visually the f0 contours did not seem to reach the high point of the rise either significantly early or significantly late in relation to the stressed syllable. In addition, L*H was confused with H* and/or L* on both mono- and multi-syllable words, and on words with both long and short vowels in the accented syllable.

As far as an impressionistic evaluation of the actual speech in the utterances is concerned, there does seem to be a difference in sound between those pitch movements marked unanimously L*H and those marked L*H and some other accent. For those marked L*H and H*, and those marked L*H and L*, the actual movement of the pitch seems very small. In addition, the rise seems to start quite high (for those confused with H*) or low (for those confused with L*) relative to the pitch in the rest of the utterance. By contrast, for the words labelled unanimously L*H (or L*H and “no accent”) the pitch movement seems very pronounced, starting perceptibly low and with quite an audible glide up to a high point. In fact, there is often a dip down in the pitch just before the low of the L*H, as if to emphasise the low by making it lower than the preceding pitch, and to give room for the movement up.

If these investigations had found any conditioning factors for the confusion of the accents there would be a strong argument for keeping the GlaToBI tonal inventory the same and simply emphasising different aspects in the training session. However the seeming lack of conditioning factors for anything but the unanimous use of L*H, on top of the impression given that the rises are *perceived* differently, would seem instead to indicate that there is more than one type of rising pitch in Glasgow English. What exactly the phonetic/phonological cues are to these differences is not clear—the scope of this study is not large enough to come to anything but an impressionistic conclusion—however the small amount of evidence presented does make a case for further investigations into Glasgow English rises.

The second major alteration made to the system was the elimination of the up-step cuing property of the H- phrase accent. This change, made so that H-L% could be used

to describe a fall in pitch at an intonational phrase boundary, forced a readjustment to the phonetic realisation of three of the four phrase accent–boundary tone combinations. The result of this readjustment was that the pitch at the end of an intonational phrase could be described (grossly) in terms of a low plateau followed by a rise, a high plateau followed by a fall, a high plateau or a low plateau. Unfortunately, however, this leaves GlaToBI transcribers unable to describe two phrase–final pitch contours which they *were* able to describe using the original ToBI system: a high plateau followed by a rise, and a low plateau followed by a fall. Again, it is not clear from this study if these contours are actually found in Glasgow English, however given the UNB preference for rising tones, there is a possibility that at least the high plateau–rise might be used, and would therefore need a symbolic representation in the GlaToBI tonal inventory.

3.4 Evaluation of the study

While this evaluation study was carried out under very similar conditions to the studies performed on the original ToBI system and the GToBI system, there were some aspects which were altered to better accommodate the GlaToBI system. Because it is possible that the alterations made could have had an effect (positive or negative) on the outcome of the study outside of the performance of the system itself, it may be useful to examine the changes in more detail.

3.4.1 The relationship between the training and the test

The corpus of 30 utterances (described in section 2.2) on which the GlaToBI system was tested, were obviously not all transcribed to the same level of accuracy—the percentage agreement between the transcribers varied greatly across individual utterances. An impressionistic evaluation of the utterances that were transcribed poorly (below the average agreement) shows that there seem to be three factors which affected the transcription: the clarity of the pitch contour, the level of disfluency and, inasmuch as it affects the potential for disfluency, the length of the utterance.

The effect that the state of the pitch contour has on the ease of transcription highlights an issue of ToBI–style systems in general: the problem of reconciling the actual speech with the f_0 contour. In the transcription of utterances which have not been scripted for the purpose of illustrating intonational patterns, the speech and the f_0 contour may contradict each other: speech which is mumbled may have a surprisingly clear pitch contour, or more commonly, speech which seems to have very clear intonation patterns may have an almost flat pitch contour. ToBI–style transcription depends on the interaction between the speech and the pitch contour—the decision as to the

choice and placement of a heard pitch prominence is often made on the basis of the visual representation of that prominence. If the two contradict each other, or are in some way not compatible, the transcription is made that much more difficult.

The other main factor in the ease or difficulty of a transcription is the fluency of the utterance. Pauses (filled and unfilled), hesitations, restarts and corrections all result in the disturbance of the intonation contour. At any point where there is a disfluency in an utterance, the transcriber must make a decision as to what has happened to the intonation. The pitch of an utterance can travel relatively uninterrupted across a pause in the utterance, or it can be diverted by it. In addition, the break index aspect of a ToBI-style system forces the transcriber to decide whether a break in the utterance flow was an intended one, signalling the possible end of an intonational phrase, or an unintended one, caused by a hesitation. In practice these classifications may not be any more straightforward than the reconciliation of the speech and the pitch contour.

Half of the utterances used for the novice transcribers' training session came from the original ToBI training corpus, the other half came from the Map Task corpus. In general, the ToBI training material (and specifically those utterances chosen for the GlaToBI training) consists of short scripted utterances with a high vowel count (to minimise pitch perturbation), which are read, for the most part, with exaggerated intonation. The resulting pitch contours are therefore very stylised, easy to read and easy to train with. The utterances in the Map Task corpus, on the other hand, are specifically not scripted, and as a result suffer from pitch perturbation, disfluencies and general long-windedness. Ostensibly, the primary criterion for choosing utterances from the Map Task corpus to use in the training was that they illustrate the important issues of the GlaToBI system. However, in an effort to avoid throwing the novice transcribers in at the deep end, as it were, the utterances were also selected for their comprehensibility and clarity of intonation.

Unfortunately, however, while these utterances may have been clearer and more prototypical, they may not have been the most appropriate training material for the test that was to be performed. A large portion of the speech on which the original ToBI system and the GToBI system were tested was read speech (Pitrelli et al. 1994, Grice et al. 1996)—in other words the use of read speech in the training of new transcribers was good preparation for the actual test transcription. The GlaToBI system, on the other hand, was tested entirely on spontaneous speech, but the speech chosen for the training was as far away from the messiness of spontaneous speech as possible while still being spontaneous. The training utterances were shorter than those seen in the actual test, ranging in general from 4 to 11 words (one utterance was 16 words long), and averaging 7 words per utterance. In addition, the utterances were very fluent,

with very few re-starts or pauses, and with markedly pronounced pitch contours. The training utterances were in point of fact all *good* examples of Glasgow English speech, from an illustrative point of view. It is questionable, however, whether they were *typical* examples of the kind of speech found in the bulk of the Map Task corpus.

3.4.2 The current ToBI set-up

The most significant problem with the set-up of ToBI used for this study was the lack of a grammar checker. Grammaticality is quite a large issue in ToBI-style transcription systems. ToBI systems are designed so that the transcribers feel that they have a fair amount of freedom in the placement of tones and break indices. Unfortunately the complicated manner in which many of the elements of the system interact, and the redundancy built into the different levels of transcription, mean that this freedom is slightly illusory—a boundary tone cannot appear without a break index of 4 and an intermediate phrase cannot exist without a pitch accent, even if the transcriber is convinced there isn't one. The transcribers testing the original system were provided with a software tool that flagged illegal or ungrammatical transcriptions and allowed them to make alterations. Unfortunately this programme was not in place for the current study, and as a result a significant number of the 50 utterances given to the transcribers had to be rejected because of a high rate of errors. In addition, one of the novice transcribers had to be eliminated from the final test because she completed the pre-test with more than the allowed number of grammatical errors per utterance. In point of fact, a number of tonal disagreements in the 30 utterances that *were* accepted for analysis could also have been caused by inadvertent errors as opposed to deliberate choices—something that a grammar checker would have been able to catch.

Chapter 4

Conclusions

4.1 Recommendations for the evaluation study and suggestions for future analysis

In the process of carrying out this evaluation study, deficits in the design of the study itself were highlighted. If this evaluation were to be carried out again, it is clear that the relationship between the training material and the test data must be taken into account. The set of utterances chosen for illustrative purposes must include *both* clear examples of prosodic phenomena *and* enough typical examples of the type of speech to appear in the test for the users to become accustomed to all the idiosyncrasies of transcription.

The issue of grammaticality in ToBI-style systems would also need to be taken into account. The fact that GlaToBI has structured compositional rules, which cannot be broken, means that when it is used accurately the levels of agreement will be very high simply by virtue of the system structure. However, the reverse is also true. GlaToBI, like all other ToBI systems, has the potential to break down and become unreliable, not because of any intrinsic design flaw, but because of misuse or misunderstanding of the system rules.

The final aspect of this evaluation study which did not live up to expectations was the analysis in terms of the kappa statistic. In principle, the addition of an alternative method of examining data produced by the study is a good idea. The kappa statistic in particular seemed appropriate because it takes into account the chance performance of the system users, an aspect which is not examined in a simple pairwise analysis. Unfortunately, the kappa statistic can only be used when the objects being categorised are independent of each other; the objects examined in this study—the words in an utterance—are inherently not independent. The kappa statistic need not be abandoned

altogether, however. The fact that it is the *words* in the utterances which are seen as the objects under consideration is partly because this is the way in which the previous two studies looked at the data, and partly due to the theory behind the system. Pierrehumbert's analysis of English intonation divides the pitch contour into a series of discrete events which occur at the accented syllables of the utterance, or at its phrase endings. In view of this, it is logical to examine the utterance in terms of its smaller parts—i.e. the words on which the pitch events do or do not occur. Another school of thought, however, sees an intonation contour as a tune which is applied to the whole utterance. A kappa analysis of each utterance in terms of the whole tune which has been chosen (via a combination of Pierrehumbert's pitch accents and phrasal tones) would be viable—aside from having a possible connection as a result of the context under which they were recorded, each utterance can be considered to be an independent object. In this type of analysis the transcriptions could be compared for the number of prominences marked, and the agreement of these prominences. Unfortunately the scope of this particular evaluation is not large enough to encompass an additional implementation of a kappa analysis, however as the topic of a future investigation it could be invaluable. A whole tune evaluation of a system that is based on the autosegmental school of thought would highlight the similarities and differences between the two theories, and, more importantly, could give some indications of the validity of both types of analysis. While the implementation of this additional evaluation would entail a shift in theoretical standpoint, from the point of view of intonation as a whole it could be a very informative analysis.

4.2 Evaluation of the GlaToBI system

The results of this evaluation study, and the success of the GlaToBI system, can be examined in terms of the goals set out for the original ToBI system. The fourth goal, that of “capability of being related to current approaches to speech recognition, to parser outputs and to formal representations of semantics and pragmatics” (Silverman et al. 1992) can be said to be intrinsically part of the GlaToBI system by virtue of it having been based on the original ToBI. The other three goals, however, are not automatically fulfilled simply because GlaToBI is a ToBI-style system. If GlaToBI is to be considered to be successful it must be seen to meet the same goals as the system on which it was based.

One goal of a ToBI-style system is “coverage: the system must be sufficiently comprehensive to capture the most important prosodic phenomena in spontaneous speech” (Silverman et al. 1992). In general the changes made to the tonal inventory to

make ToBI more compatible with Glasgow English seem very appropriate. However, in light of the study showing the use made of the new tonal inventory, some additional changes could be made. Examination of the use of the GlaToBI rise, L*H, seems to indicate the possibility of more than one type of rise in Glasgow English. The rise described by L*H is a glide up *on* the accented syllable; the additional rise indicated by the use made of L*H is one in which the accented syllable is aligned with the high point of the rise (i.e. L+H*). This study cannot rule out the possibility that this or other rises are not used in Glasgow English. The recommendation is, therefore, that the two ToBI rises L+H* and L*+H be left in the tonal inventory as possible choices for labelling rising tones. If this is done, however, the training must cover rising tones much more extensively, in order for the non-expert transcribers to become used to hearing and labelling the differences in alignment.

As mentioned in section 3.3, the second change made to the system, the removal of the up-step cuing properties of H-, leaves transcribers unable to describe a final pitch plateau which is high and goes higher. In addition, the re-assignment of the compound tone L-L% to represent a low plateau, means that there is no symbol for a final low pitch plateau followed by a lower boundary. If future investigations find that the two compound tones are in fact used in Glasgow English, a diacritic along the lines of the downstep diacritic, !, could be adopted to indicate an additional final raising of the pitch after a high phrase accent, or lowering after a low accent.

A third goal for the system was that it be learnable “in a relatively short time” (Silverman et al. 1992). Prior to the experiment proper the novice transcribers had one full day of training in which they were familiarised with the concepts behind GlaToBI and with the system itself. In testing, the novices achieved levels of agreement near to (or higher than) those reached by the expert system designers. In addition the novices did not seem to make a great many more grammatical errors than the experts. The overall indication is that, with the appropriate training, GlaToBI is indeed learnable to near-expert levels in a short period of time.

The primary goal set by the ToBI designers is that of reliability—specifically, that “agreement between different transcribers [using the system] must be at least 80%”. According to the pairwise analysis performed on GlaToBI, the 7 transcribers who tested the system reached, on average, 78% agreement—very close to this goal. Examining each area of analysis separately, the lowest and highest levels of disagreement (the rate of agreement as to choice of pitch accent: 65%, and the rate of agreement as to the presence or lack of edge tone: 92%) can be seen as resulting from the system design (see section 3.1.2). If these two areas are omitted temporarily, the lowest level of agreement for the GlaToBI system is 74%. In addition, not only is the system reliable in its own

right, but it also compares favourably with both the original system and the version developed for German. The levels of agreement achieved in the tests of the original ToBI system ranged from an average of 70.4% (choice of break index) to an average of 93.4% (presence or lack of phrasal tone). The GToBI system ranged from 74% (choice of pitch accent) to 87% (presence or lack of pitch accent). The GlaToBI results are slightly lower overall, but this is most likely due to the fact that the study is on a much smaller scale than the ToBI and GToBI studies: the same levels of disagreement as seen by the two larger studies will have a greater effect on the smaller GlaToBI evaluation.

As mentioned above, some changes could be made to the GlaToBI system to improve the precision of transcription, but it is clear from the results presented here that even without these additions GlaToBI meets the original ToBI goals of reliability and learnability. The recommended alterations to the system apply to the design goal of coverage; with these changes in place GlaToBI should meet all four of the goals set out for the original ToBI, and would be a comparable system to ToBI for the transcription of Glasgow English intonation.

Bibliography

- Anderson et al. (1991), 'The HCRC Map Task Corpus', *Language and Speech* **34**(4), 351–366.
- Beckman, M. E. & Ayers, G. M. (1993), *Guidelines for ToBI Labelling, Version 1.5*, Ohio State University.
- Beckman, M. E. & Hirschberg, J. (1994), *The ToBI Annotation Conventions*.
- Benzmüller, R. & Grice, M. (1996), *Trainingsmaterialien für GToBI*, University of the Saarland, Saarbrücken.
- Carletta, J. (1996), 'Assessing agreement on classification tasks: The kappa statistic', *Computational Linguistics* **22**(2), 249–254.
- Cruttenden, A. (1986), *Intonation*, Cambridge University Press.
- Cruttenden, A. (1994), Rises in english, in J. W. Lewis, ed., 'Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor', Routledge, pp. 155–173.
- Grice, M., Reyelt, M., Benzmüller, R., Mayer, J. & Batliner, A. (1996), Consistency in transcription and labelling of German intonation with GToBI, in 'Proceedings of the 1996 International Conference on Spoken Language Processing (forthcoming)'.
- Ladd, D. R. (1983), 'Phonological features of intonational peaks', *Language* **59**, 721–759.
- Ladd, D. R. (1996*a*), Intonational phonology, In Press.
- Ladd, D. R. (1996*b*), 'MSc in Speech and Language Processing, prosody module: Lecture notes'.
- Pierrehumbert, J. (1980), The Phonology and Phonetics of English intonation., PhD thesis, MIT.

- Pierrehumbert, J. & Hirschberg, J. (1990), The meaning of intonational contours in the interpretation of discourse, *in* 'Intentions in Communication', MIT Press.
- Pitrelli, J., Beckman, M. & Hirschberg, J. (1994), Evaluation of prosodic transcription labeling reliability in the ToBI framework, *in* 'Proceedings of the 1994 International Conference on Spoken Language Processing', Vol. 1, pp. 123–126.
- Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S. & Fong, C. (1991), 'The use of prosody in syntactic disambiguation', *J. Acoust. Soc. Am.* **90**(6), 2956–2970.
- Siegel, S. & Castellan Jr., N. J. (1988), *Nonparametric Statistics for the Behavioral Sciences*, International edn, McGraw–Hill Inc.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992), TOBI: A standard for labelling English prosody, *in* 'Proceedings of the 1992 International Conference on Spoken Language Processing'.
- Vendetti, J. (1995), *Japanese ToBI Labelling Guidelines*, Ohio State University.