Slips and errors in spoken data transcription

Isabella Chiari

Università La Sapienza di Roma
Dipartimento di Studi Filologici, Linguistici e Letterari (DSFLL), p.le Aldo Moro, 5 Roma (Italy)
isabella.chiari@uniroma1.it

Abstract

The present work illustrates the main results of an experiment on errors and repairs in spoken language transcription, with significant relevance for the evaluation of validity, reliability and correctness of transcriptions of speech belonging to several different typologies, set for the annotation of spoken corpora. In particular, we dealt with errors and repair strategies that appear on the first drafts of the transcription process, that are not easily detectable with automatic post-editing procedures. 20 participants were asked to give an accurate transcription of 22 short utterances, repeated from one to four times, belonging two non-spontaneous (10) and spontaneous conversation (10). Error analysis suggest a general preference for meaning preservation even after the alteration of the original form, and for the preference for certain error patterns and repair strategies.

1. Introduction

Transcription of spoken language is becoming a common practice in corpus linguistics, computational linguistics, besides being a regular activity in administrative, parliamentary and judiciary acts. Even though transcription of speech and conversation entails complex linguistic annotation systems (such as those included in the recommendations guidelines of *Text Encoding Initiative* and *Corpus Encoding Standard*), there is a basic transcription level dealing with a preliminary phase of rough orthographic transcription. Such basic transcription is generally followed by grammatical annotation and, in some cases, phonetic and phonological tagging. Although some automatic transcription systems, based on speech recognition (ASR), already exist for some languages, the majority of spoken corpora employ human transcribers to carry out this task. Recent literature has often been centered on transcription system design (e.g. Du Bois, 1991; Edwards, 1992; Du Bois et alii, 1993; Gumpertz & Berenz, 1993; Cook, 1995; Leech et alii, 1995), on reviewing and comparing different transcription systems (e.g. Psathas, & Anderson, 1990; Edwards, 1995; Cook, 1995; Johansson, 1995; Chafe, 1995; O'Connell & Kowal, 1995a, 1995b), and on errors and inconsistencies in linguistic annotation (e.g. Oppermann, Burger and Weilhammer, 2000). However, a consistent amount of errors and repairs occur even at the basic level of transcription, when the mere sequence of spoken words are heard and transcribed. Some of these errors are corrected in further stages of annotation (especially when phonetic and phonological labeling is required), but some others remain undetected in the revision process.

The present experiment is focused on the phase of mere orthographic transcription of the first draft (deliberately excluding further linguistic tagging, such as grammatical or paralinguistic annotation which require specific skills to be learned and developed) of spontaneous speech carried by not specifically trained individuals. The aim of the experiment is similar to Lindsey & O'Connell (1995), but is furthermore meant to investigate the understanding and re-productive acts involved in the transcription process.

Transcription errors and slips of the ear made while listening and transcribing (or repeating) spontaneous spoken material, have been observed and analyzed, with special attention devoted to those made with no awareness on the part of the subject producing them. The experiment is based on the presentation of spontaneous speech to transcribers (dialogues turns and monologue utterances in spoken Italian). Results show patterns in error typologies, compensation and repair strategies and eventual self-correction.

2. Data analysis

Speech from two different typologies was selected to be included in each test: type A includes accurate read or controlled speech (from television broadcast news or public speeches), while type B includes spontaneous speech (in various ordinary situations, from real-tv shows). All recordings where digitalized (acquired directly from tv source in February 2006), segmented into turns (utterance turns or dialogue turns), and saved in wav format to be heard on a compact player or from pc speakers.

All recording were selected on the base of the highest quality of audio sound (with least background noise possible and no superimpositions). Each turn contains only one speaker's voice, and is a complete utterance, brief sequence of utterances or a meaningful portion of a long utterance. Length varies from around 1.5 sec to 13 secs.

Utterances selected for each test typology were chosen to be belonging to the same "spoken text" where possible, as to preserve the listener's ability to rely on what has been previously heard.

Before each of the two series of hearing exposures, participants were presented with a test for volume adjustment with utterances not belonging to their test type. Before the first series two utterances were added (without telling participants) as a training, and were not computed in the results.

Each test consisted of 22 different utterances: the first two were the training utterances, followed by 10 utterance form non-spontaneous controlled speech (news and public speeches) and 10 utterances from spontaneous speech (single dialogue turns with only one speaker talking). A total of 100 different utterances were presented.

Participants were given a brief sociolinguistic questionnaire and paper for drafts. Subjects were asked to transcribe in handwriting the spoken sequences they heard (choosing their own jotting strategies: online or offline),

and then to copy their drafts in an ordered form at the end of data exposure. They were also told to write down only the words spoken (excluding vocal activities, noises and pauses) and not to clean up text, in particular signaling repetitions they heard and not correcting errors produced by speakers. After the data exposure phase participants were not allowed to correct their first draft.

The administration of spoken data was conducted by the experimenter with the aid of a computer with speakers. Before each utterance, participants were told how many times they were to ear it (one to three times depending of length of sequence). The entire duration of the experiment lasted about 30 minutes for each participant.

3. Error Analysis and Results

Sample spoken material consisted of 100 different utterances (50 in controlled speech and 50 in spontaneous speech), plus two control utterances added at the beginning of the test. The total amount of utterance token presented to the subjects was 400. Utterances ranging one to five seconds were presented once, from five to eight seconds twice, and those lasting more than eight seconds were run three times.

Different tests were presented to 20 participants (12 women and 8 men), whose age ranged from 18 to 62 years old, with an average of 28, all having obtained at least an high school degree.

The 20 utterances belonging to each test were analyzed in order to obtain a full list of errors, where the participant's transcription differed from a supervised transcription (always checked with audio). Missing words or misperception of the first word and last word of each utterance has not been computed, since they involve a certain amount of surprise and voice lowering. Given that participants were not themselves managing repetition of utterances it would have been misleading.

A total amount of 455 errors have been reported, with an average of 22.7 errors per participant (about 1.13 errors per utterance heard). 5.75 errors per utterance type were reported in the whole experiment.

A slight different in frequency differentiates the two text typologies selected. Controlled speech induces errors in 48.4% of the total, while spontaneous speech covers 51.6%. In this specific case since utterances in controlled speech were selected from television news and speeches there is probably an error effect due to fast speech rate. While generally spontaneous utterances were relatively shorter in duration, and still gathered more errors.

	Frequency	%
Controlled speech	220	48.4
Spontaneous speech	235	51.6
Total	*455*	*100.0*

Table 1: Errors per Speech Typology

Errors were also analyzed to observe more precisely what kind of change occurred in transcriptions: substitution, addition, deletion, movement (see Table 2).

Substitutions were an element is switched with another at any linguistic level occurred 205 times (45.1%). Examples are utterances where *un profondo cambiamento* is transcribed as *un grande cambiamento*. Among substitutions 52.7% of occurrences involve lexical elements, 19% function words and 16.6% verb conjugation errors. Target grammatical categories involved in word level substitutions are maily verbs 21.5% (44), prepositions 13.2 (27), pronouns 11.% (23) and nouns 9.8% (20). Substitutions in the great majority of cases involve elements belonging to the same grammatical category. Regarding content preservation in word level substitutions, in 38.7% of cases meaning is preserved completely, in 22.6% is partially preserved, while in 38.7% a complete misunderstanding occurs.

	Frequency	%
substitution	205	45.1
addition	40	8.8
deletion	199	43.7
movement	11	2.4
Total	*455*	*100.0*

Table 2: Type of Change

Addition or insertion of words is relatively rare (8.8%), and can be generally seen as a repair device where subjects try to give a written textual form to the spoken material (adding conjunctions for examples instead of reporting direct coordination in a sequence of sentences). The far commonest addition is that of the conjunction *e* ("and"), that occurs in nearly half of the cases (45%). Additions generally affect function words (in 72.5% of the cases, with conjunctions – *e* – and articles – *la* – inserted in the textual material), while lexical units are added in 25% of the errors of this kind. While from the semantic point of view additions rarely change utterance meaning. Meaning is preserved in 90% of the cases, and partially preserved in 7.5%.

Among deletion of words is common misdetection of repetitions (21.6% of deletion cases), especially of function words not playing any role other than fillers (*fa la parte di quello che che mi prende in giro*, instead of *che che*). Deletions occur in 43.7% of errors (199 cases).

Deletions often regard entire constituents (41.7% of cases), and are generally more dangerous for meaning preservation: 50.3% of cases are not affected semantically by the error, while 16.6% are partially affected and 33.2% lead to misunderstanding. Grammatical categories affected by deletion are mainly pronouns (14.1% of deletions), adverbs (9.5%), verbs (9%) and prepositions (8.5%).

Movement is the least frequent phenomenon with only 2.4% of total error occurrences. Movement rarely changes the overall meaning of the utterance (18.2% of movement cases), and always involves entire sentence fragments and not single words (*sull'appennino centrale e sul medio versante* instead of *sul medio versante e sull'appennino centrale*).

Looking at all the different phenomena together we observe a general tendency at preserving the overall meaning of the sentence (45.9%), especially when single

words are affected (and not whole constituents) (55.1% preservations, and 20.7% partial preservations).

	Frequency	%
yes	209	45.9
partial	76	16.7
no	170	37.4
Total	*455*	*100.0*

Table 3: Meaning preservation on total errors

Looking more closely at error types, those occurring at word level imply deletions as the most frequent phenomenon, followed by lexical substitution and addition (see Table 4).

	Freq.	%
lexical switch	53	20.7
sing/plur switch	5	2.0
switch substituent with lexical element	14	5.5
function word substitution	40	15.6
insertion of words	41	16.0
missing words	76	29.7
verb conjugation error	25	9.8
phonetic variant	2	.8
Total	*256*	*100.0*

Table 4: Error types at word level

The presence of an error (especially those that imply substitution of verb tense or person, and singular/plural switching) often produces the occurrence of other errors in the following words, since the transcriber tends to repair textual cohesion signals. For example, since the transcriber has erroneously perceived a singular subject (*il corridore*) in the utterance (*I soccorritori avrebbero avuto problemi*), the rest is conjugated with a verb agreement in the singular form (*avrebbe avuto problemi*).

It is interesting to note that participants who were given the same utterances to transcribe tended to make the same errors and repairs (tipical is the deletion of *anche* in the utterance *E un quasi decalogo di consigli pratici è arrivato* anche *dal ministero delle attività produttive*).

4. Discussion

The experiment was both meant to provide hints on human understanding and creative repair in a linguistic re-production task and suggest specific error typologies that can and do occur in linguistic corpora transcription and that are not easily detectable in automatic post-editing procedures without direct access to the spoken audio material.

The most striking finding regards the amount of repair that does not rely of linguistic form but on creative unconscious reconstruction made by the transcriber, that generally tends to preserve utterance meaning.

Lexical substitutions cannot be thus attributed to misunderstanding or slips of the ear (see Voss, 1984; Bond, 1999; Chiari, 2005), but to subsequent interventions relying on what the hearer has actually understood. From this point of view we can see repair strategies as:

1) a coherent re-creation of the spoken text
2) as textual reproduction of written conventions to the spoken material (deletion of repetition, especially those representing hesitation)
3) as error correction (as in the redundant expression *a me mi dispiace* becoming for the transcriber *a me dispiace*).
4) a consequence of the "volatility" of the form of the utterance.

The errors analyzed imply that there are errors patterns, common errors and repairs suggesting that there might be weak elements in a spoken discourse which are more often subject to deletion or repair.

It is also interesting to note that actual misunderstanding tend to occur more at a sentence level than at word level, implying a difficulty in the general segmentation and detection of the spoken material, especially occurring in a context of unpredictability and surprise.

A possible interpretation of this findings is that ordinary understanding practices are strictly focused on meaning rather than form, so that, even with the best possible audio quality, when trying to concentrate attention on the reconstruction of linguistic form, we tend to shift and rely on our understanding strategies, that lead us to re-create text in a plausible way.

A stress factor was established by the small number of repetition that were administered by the experimenter (and not autonomously by the participants). As was already said, the total length of the experiment was 30 minutes, far less than transcription times for professional corpus transcribers, thus reducing factors like drops of attention, and fatigue that often influence transcription accuracy.

Further research should be addressed to specific corpus transcription error analysis, to a more natural setting and audio management, and to a more precise evaluation of performance in relation to explicit instruction to participants.

The observation of naturally occurring errors in transcription should also suggest best practices and guidelines to be modeled as to include specific training in detecting a certain amount of weak elements. Reliable transcriptions should always be subjects to revisions by different transcribers, and most of all should always involve direct access to audio material. The observed tendency in making the same transcription mistakes by different participants suggest the need of a specific revision phase focused on repair pattern that often remain undetected, because of their semantic plausibility.

Issues of relevance, selection, interpretation, memory, recall and self-monitoring should be addressed as further evidence of speech production and understanding processes. Better knowledge of transcription errors allows improved planning of instruction manuals supplied to transcribers (training the ears and training the mind towards formal and superficial linguistic elements) and improvement in the correction and revision phases during corpus processing and annotation.

5. References

Bond, Zinny S. (1999). *Slips of the Ear. Errors in the perception of casual conversation.* New York: Academic Press.

Chafe, W. (1995). Adequacy, user-friendliness, and practicality in transcribing. In G. Leech, G. Myers, & J. Thomas (eds.), *Spoken English on computer:Transcription, mark-up, and application*, Harlow, England: Longman, pp. 54-61

Chiari, I. (2005). Condizioni, limiti e analogie del lapsus linguae e del lapsus auris. In A. Frigerio e S. Raynaud (eds), *Significare e comprendere: la semantica del linguaggio verbale*, Roma: Aracne, pp. 129-44.

Cook, G. (1995). Theoretical issues: transcribing the untranscribable. In Leech, G., Myers, G. and Thomas, J. (eds.) *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman, pp 35-53.

Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. *Pragmatics*, 1, pp. 71-106.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & Paolino, D. (1993). Outline of discourse transcription. In J. A. Edwards & M. D. Lampert (eds.), *Talking data:Transcription and coding in discourse research*. Hillsdale, NJ: Erlbaum, pp. 45-89.

Edwards, J. A. (1992). Design principles in the transcription of spoken discourse. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium* 82, Stockholm, 4-8 August 1991. Berlin, Germany: de Gruyter.

Edwards, J. A. (1993). Principles and contrasting systems of discourse transcription. In J. A. Edwards & M. D. Lampert (Eds.), *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Erlbaum, pp. 3-31

Edwards, J. A. (1995). Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up, and application*. Harlow, England: Longman, pp. 19-34

Gumperz, J. J., & Berenz, N. (1993). Transcribing conversational exchange. In J. A.Edwards & M. D. Lampert (Eds.), *Talking data: Transcription and coding in discourseresearch*. Hillsdale, NJ: Erlbaum, pp. 91-121

Johansson, S. (1995). The approach of the Text Encoding Initiative to the encoding of spoken discourse. In Leech, G., Myers, G. and Thomas, J. (eds.) *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman. pp. 82-98.

Leech, G., Myers, G., & Thomas, J. (1995). *Spoken English on computer: Transcription, mark-up, and application*. Harlow, England: Longman.

Lindsay, J., & O'Connell, D. C. (1995). How do transcribers deal with audio recordings of spoken discourse? *Journal of Psycholinguistic Research*, 24, pp. 101-115.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), Developmental pragmatics (pp. 43-72). New York: Academic Press.

O'Connell, D. C., & Kowal, S. (1995a). Basic principles of transcription. In J. A. Smith, R. Harre, & L. Van Langenhove (Eds.), Rethinking methods in psychology(pp. 93-105). London, England: Sage.

O'Connell, D. C., & Kowal, S. (1995b). Transcription systems for spoken discourse. In J. Verschueren, J.O. Ostman, & J. Blommaert (Eds.), *Handbook of pragmatics*. Amsterdam, The Netherlands: John Benjamins, pp. 646-656.

Oppermann, D., S. Burger and K. Weilhammer (2000). What are transcription errors and Why are they made? In *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC 2000), Athens, Greece, pp. 409-441.

Psathas, G., & Anderson, T. (1990). The 'practices' of transcription in conversation analysis. *Semiotica*, 78, pp. 75-99.

Voss, B. (1984). *Slips of the ear: Investigations into the speech perception behaviour of German speakers of English*, Tübingen: Narr Verlag.