I. Chiari

SPOKEN CORPORA AND TRANSCRIPTION ERRORS

1. Introduction

Transcription of spoken language is an ordinary practice in modern linguistics (particularly in corpus linguistics, computational linguistics) and in administrative, parliamentary and judiciary acts. Recent literature has often been centred on transcription system design, on reviewing and comparing different transcription systems and on errors and inconsistencies in linguistic annotation[1].

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

[1] *Du Bois J.W.* Transcription design principles for spoken discourse research // Pragmatics. 1991. № 1. P. 71–106; *Edwards J.A.* Design principles in the transcription of spoken discourse // *Svartvik J.* (ed.). Directions in corpus linguistics: Proceedings of Nobel Symposium. Berlin, Germany, 1992. № 82. P. 4–8; *Du Bois J.W., Schuetze-Coburn S., Cumming S., Paolino D.* Outline of discourse transcription // *Edwards J.A., Lampert M.D.* (eds.). Talking data: Transcription and coding in discourse research. Hillsdale, NJ, 1993. P. 45–89; *Gumperz J.J., Berenz N.* Transcribing conversational exchange // *Edwards J.A, Lampert M.D.* (eds.). Talking data: Transcription and coding in discourse research. Hillsdale, NJ, 1993. P. 91–121; *Leech G., Myers G., Thomas J.* (eds.). Spoken English on Computer: Transcription, Markup and Applications. Harlow, 1995; *O'Connell D.C., Kowal S.* Basic principles of transcription // *Smith J.A., Harre R., Van Langenhove L.* (eds.). Rethinking methods in psychology. London, 1995. P. 93–105; *O'Connell, D. C., & Kowal, S.* Transcription systems for spoken discourse // *Verschueren J., Ostman J.O., Blommaert J.* (eds.). Handbook of pragmatics. Amsterdam, 1995. P. 646–656; *Oppermann D., Burger S., Weilhammer K.* What are transcription errors and Why are they made? // Proceedings of the Second International Conference on Language Resources and Evaluation (LREC–2000). 2000. P. 409–441.

further stages of annotation (especially when phonetic and phonological labelling is required), but some others remain undetected in the revision process since they are not easily detectable with automatic post-editing procedures. These kinds of errors generally derive from the transcribers' involuntary creative reconstruction of the spoken material heard and thus result in perfectly grammatical and meaningful sentence. Errors at this basic level of transcription have been rarely analyzed[1], mainly because they often remain unnoticed in further stages of annotation.

In the present work results from an experiment conducted on errors and repairs in spoken Italian language transcription will be illustrated briefly and discussed. The experiment was 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 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.

Some of the questions addressed are: What kind of errors transcribers make? Are there any patterns in error typologies? Are human being reliable listeners? Are there possible explanations of the various transcription errors? Is there a way of avoiding those errors? Is there a way of correcting them in additional stages of processing? Can we improve transcription accuracy?

---

[1] *Lindsay J., O'Connell D.C.* How do transcribers deal with audio recordings of spoken discourse? // Journal of Psycholinguistic Research. 1995. № 24. P. 10–115.

2. Experiment method and procedure

A brief account of experimental procedure will be given in the following paragraph[1]. Each of the 20 participants was submitted to the hearing of 22 different utterances to transcribe (2 training utterances; 10 utterances from controlled speech and 10 from spontaneous speech[2]). Utterances were recorded from television source, selected only with least noise and no superimpositions, best audio quality and subsequently segmented into turns. Length of utterance turns varies from around 1,5 sec to 13 seconds. Participants were asked to transcribe in handwriting the spoken sequences they heard, only the words spoken (excluding vocal activities, noises and pauses), trying not to clean up text. 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).

On the total amount of 400 utterance presented to the subjects 455 errors have been reported, with an average of 22,7 errors per participant (about 1,13 errors per utterance heard).

---

[1] More details about the methodology used and results analysis will be found in: *Chiari I.* Slips and errors in spoken data transcription // Proceedings of 5th International Conference on Language Resources and Evaluation LREC–2006. Genova, 2006. P. 1596–1599; *Chiari I.* What do we do when we transcribe speech? Typologies in lexical substitutions // *Pusch C.D., Raible W.* (eds.). Romanistische Korpuslingustik III: Korpora und Pragmatik. 2007. Tübingen, in print.

[2] An example of controlled speech is: *L'Italia nella morsa del freddo. Temperature in picchiata da nord a sud, miglioramento previsto da mercoledì* (R26: 5.52 secs). An example of spontaneous speech is: *Quando ieri è stata fatta la spesa e si poteva fare qualche altra cosa* (R1018: 2.59 secs).

3. General overview of results

A comparison of different textual typologies was conducted in order to find out if there are any differences in error rate in controlled versus spontaneous speech. Data does not provide any special insight. A slight variation 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 of news broadcast reading habits. Usually spontaneous utterances were relatively shorter in duration, and still gathered more errors.

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).

Errors were further analyzed to observe more specifically what kind of change occurred in transcriptions. Simple structural categories common in slips and error research were used: substitution, addition, deletion, movement. The most common type of errors were substitutions (205 cases, 45,1%) and deletions (199 cases, 43,7%), while cases of addition (40 cases, 8,8%) and movement (11 cases, 2,4%) were fairly rare.

4. Discussion and Guidelines

Main findings suggest that listeners are not particularly reliable transcribers, unless their main task is meaning or content centred. Even when explicitly asked (and trained) to concentrate on form (and on the sequence of exact words to reproduce), the attitude of the transcriber turns toward meaning-centred practices. A possible interpretation of this findings might be that ordinary understanding behavior is 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. Errors in these cases derive from understanding rather than misunderstanding.

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. Nevertheless, even trained transcribers tend to make mistakes of which they remain unaware.

Thus the observation of naturally occurring errors in transcription should suggest best practices and guidelines to be modeled as to include specific training in detecting a certain amount of weak elements. General suggestions include: making transcribers aware of common errors made during transcriptions, of their frequency and typology; manuals supplied to transcribers should include specific chapters on common errors; making transcribers acquire a default attitude of doubt toward first heard sequences (even when short and sounding meaningful); assuring at least three different revisions of the transcription process, with direct access to the original audio material; revisors should be different individuals from transcriptors.

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. Experimental and analytic research on error typologies in different languages would reveal new insight into hearing and understanding processes, in listeners' strategies and in language similarities and differences.