TRANSCRIBING SPEECH
errors in corpora and experimental settings

1

ISABELLA CHIARI
UNIVERSITÀ LA SAPIENZA DI ROMA

ISABELLA.CHIARI@UNIROMA1.IT
WWW.ALPHABIT.NET

Transcription of speech

Functional practice and linguistic act

- Contexts
  - Administrations
  - Government organs (parliament)
  - Judiciary courts
  - News reports
  - Podcast transcripts
- Linguistics
  - Ethnography
  - Conversation analysis
  - Corpus linguistics
  - Computational linguistics
- Transcription is a linguistic act itself, governed by its own strategies

What kind of transcription errors?

3

Errors and repairs at the basic level of transcription

- Not orthographic or grammatical errors
- Not linguistic annotation errors
- Not non-linguistic events

Errors in the pure identification of the spoken words only: MISTRANSCRIPTIONS

Errors that appear in various phases of the transcription process

- not easily detectable with automatic post-editing procedures

Transcription of speech and errors

4

Questions	Transcription process

• What happens when we are involved in a transcription task?
• What kind of errors do we make? Are there any patterns in errors?
• Are there possible explanations for these errors?
• Are these errors predictable? Are they avoidable?
• Can we improve transcription accuracy?
	• Transcription involves
	  ○ Interpretation and choices
	  ○ Selective process
	  ○ Filtering of the transcriber
	• Transcripts containing errors are generally
	  ○ Grammatical
	  ○ Meaningful

Goals

5

Psycholinguistic goals

- Errors as evidence of listening and transcribing processes and strategies
- Error patterns based on **meaning** and on **form**
- Similarities and differences in typologies and relative frequencies in experimental and corpus data

Practical goals

- Transcriber's guidelines
  - **Evidence about common errors** made during transcriptions, of their frequency and typology
  - improved planning of **instruction manuals** supplied to transcribers
  - improvement in the **correction and revision phases**

Transcription of spoken Italian

6

Corpus investigation ✓

- Variable input
  - Variable sequence lenght
  - Variable nr. of repetitions
  - Indefinite performance lenght
  - Indefinite nr. of transcribers
  - No infos about transcribers and revisors
- Various different settings
- Audio self-administered

Experimental research ✓

- Controlled input
  - Fixed spoken sequence lenght
  - Nr. of repetitions
  - Fixed experiment lenght
  - Given nr. of transcribers
  - Transcriber's infos
- Given setting
- Audio administered by experimenter

Structural Change (1): **SUBSTITUTION**

7

An element is switched with another at any linguistic level.

CORPUS DATA

Ex1. "l'attore americano , Tom Cruise , ha presentato , presso un tribunale di Los Angeles l'istanza di separazione dalla moglie Nicole Kidman"

istanza di divorzio > ***istanza di separazione***

"application for divorce" > *"application for separation"*

Ex2. "è opportuno lavare la cassetta una due volte alla settimana utilizzando acqua"

lavare la cassettina > ***lavare la cassetta***

"wash the litter box" (little box in Italian) > *"wash the box"*

EXPERIMENTAL DATA

Ex1 . "e spiega che questo governo ha avviato un grande cambiamento"

un profondo cambiamento > ***un grande cambiamento***

"a deep change" > *"a great change"*

Ex2. "Berlusconi parla ad un congresso di Confindustria"

parla a un convegno di > ***parla a un congresso di***

"talks at a convention" > *"talks at a congress"*

Structural Change (2): **ADDITION**

8

An element is inserted into the original sequence of words.

CORPUS DATA

Ex1. "che gli affreschi che erano stati messi sopra, cioè che avevano intonacato"

cioè avevano intonacato > ***cioè che avevano intonacato***

"in other words they plastered" > *"in other words that they plastered"*

Ex2. "che concepì in una idea di governo mondiale i giorni longitudinali, quelli che vanno dal nord al sud"

che vanno > ***quelli che vanno***

"that go" > *"those that go"*

EXPERIMENTAL DATA

Ex1 . "Fermarci ora sarebbe il colpo di grazia per l'economia. Sentiamo ora l'inviato"

sentiamo l'inviato> ***sentiamo ora l'inviato (2)***

"Lets listen to the reporter" > *"lets listen now to the reporter"*

Ex2. "Mi dispiace molto quindi anche lei adesso le ho parlato e le ho chiesto scusa"

ho parlato le ho chiesto> ***ho parlato e le ho chiesto***

"I talked to her I apologized" > *"I talked to her and I apologized"*

Structural Change (3): DELETION

9

An element is cancelled from the original sequence of words

CORPUS DATA

Ex1. "e tutte le attrezzature necessarie per panifici pasticcerie gelaterie e pasta fresca"

pasticcerie pizzerie gelaterie> ***pasticcerie gelaterie***

"bakeries, pizzerias, ice cream shops" > "bakeries, ice cream shops"

Ex2. "immagino che voi abbiate vissuto un pizzico di scetticismo"

Immagino che anche voi > ***immagino che voi***

"I imagine that you too have experienced" > "I image that you have experienced"

EXPERIMENTAL DATA

Ex1 . "E un quasi decalogo di consigli pratici è arrivato dal ministero delle attività produttive"

è arrivato anche dal ministero > ***è arrivato dal ministero***

"has arrived also from the ministry" > "has arrived from the ministry""

Ex2. "Ma non erano meglio le caffettiere delle mamme delle nonne* che quando facevano il caffè si sentiva pure al sesto"

Il caffè al primo piano si sentiva > ***il caffe si sentiva***

"when they made coffee on the first floor you could smell it" > "coffee you could smell it"

Structural Change (4): MOVEMENT

10

One or more elements misplaced in the order sequence

CORPUS DATA

Ex1. "i Vergine che hanno soprattutto un'attività indipendente sotto questo punto di vista dovrebbero essere soddisfatti"

soprattutto che hanno> ***che hanno soprattutto***

"Virgos especially that have an indipendent activity" > "Virgos that have especially"

Ex2. "provate a scegliere voi quello che vi piace di più"

provate voi a scegliere > ***provate a scegliere voi***

"try yourself to choose" > "try to choose yourself"

EXPERIMENTAL DATA

Ex1 . "Ancora denunce di bottiglie manomesse. Finora otto casi accertati"

Otto casi finora accertati > ***finora otto i casi accertati****

"eight cases until now ascertained" > "until now eight cases ascertained"

Ex2. "Ma non erano meglio le caffettiere delle mamme delle nonne che quando facevano il caffè si sentiva pure al sesto"

delle nonne delle mamme > ***delle mamme delle nonne***

"of grandmothers of mothers" > "of mothers of grandmothers"

CLIPS corpus of spoken Itatalian

Corpora e Lessici di Italiano Parlato e Scritto

- www.clips.unina.it, last accessed 8 July 2007

CLIPS MEDIA subcorpus

- 50% radio broadcasts and 50% television broadcasts
- national and local networks
- sixty minutes of national broadcasts (thirty minutes of radio and thirty of television), and about eighteen minutes for each of the fifteen cities where the recordings took place
  - (Bari, Bergamo, Bologna, Cagliari, Catanzaro, Florence, Genoa, Lecce, Milan, Naples, Palermo, Parma, Perugia, Rome, Venice)
  - **330 minutes** (5.5 hours)

CLIPS media transcripts

Orthographic transcripts have been produced by different transcribers

- (a total of 29 transcribers for the whole 100 hours corpus)
- and subsequently revised by different researchers

A smaller section of the corpus has also been phonetically annotated

- thus leading to a further revision of the full transcripts
- No explicit trace of the number of revisions is given in the public documentation.

CLIPS media errors overview

13

Minutes of recordings	**330**
Errors reported	**135**
Avrg. nr. errors per minute	**0.41**

		Frequency	**%**
	Radio	66	**48,9**
	Television	69	**51,1**
	Total	135	**100,0**

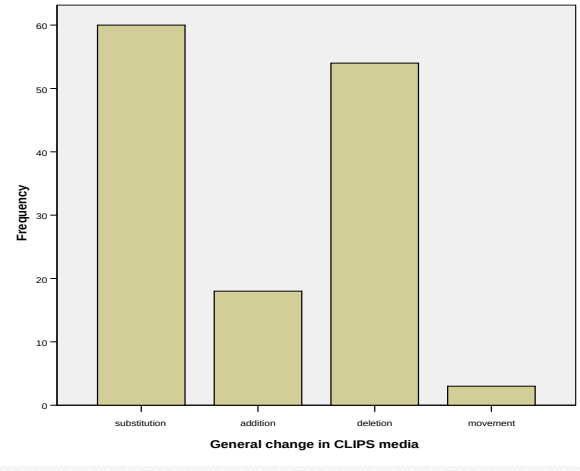
l'iscrizione sul ("on") *registro degli indagati*
is transcribed as
> *l'iscrizione nel* ("in") *registro degli indagati*

nei suoi riguardi
is transcribed as
> *nei suoi confronti*.

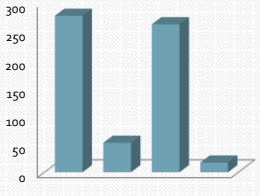
MEANING PRESERVATION	Frequency	**%**
yes	67	**49,6**
partial	21	**15,6**
no	47	**34,8**
Total	135	**100,0**

Structural change in CLIPS media

14



substitution
(60 cases, 44.4%),
deletions
(54 cases, 40%),
addition
(18 cases, 13.3%)
movement
(3 cases, 2.2%)



EXPERIMENTAL METHOD

15

Material

- Each participant was submitted to the hearing of 22 different utterances to transcribe
- Two speech typologies
  - A. Controlled speech
  - B. Spontaneous speech

Samples

- Material acquired from tv broadcasts
  - Segmented in turns (utterance turns or dialogue turns)
  - Highest sound quality with least possible noise
  - No superimpositions
- Length varies from around **1.5 sec to 13** seconds.

Tests and participants

16

- Test: 22 utterances (2 training, 10 type A, 10 type B)
- 100 different utterances were tested (50 in type A speech and 50 in type B): total 400
- Administration of audio was performed by the experimenter
- Before each utterance, participants were told how many times they were to ear it (one to three times depending of length of sequence).
- Mean length of total experiment: 30 minutes
- Listening material: 2 minutes (1 minute of controlled speech 1 minute of spontaneous speech)

Participants and Test summary

17

Participants Nr.	20
Sex	Female: 12 Male: 8
Educ. degree	All attending university
Utterances analyzed	400
Errors reported	**455**
Avrg. nr. errors per participant	22.7
Avrg. nr. Errors per utterance	1.13
Nr errors per minute of listening	**11.38**

Errors overview

18

SPEECH TYPOLOGY

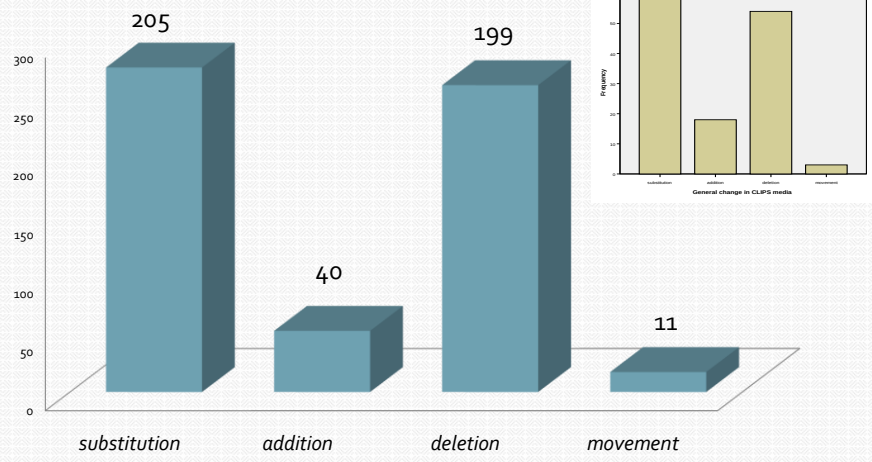
	Freq.	%
A. Controlled speech	220	48.4
B. Spontaneous speech	235	51.6
Total	**455**	**100**

MEANING PRESERVATION

	freq	%
Meaning fully preserved	209	45.9
Partially	76	16.7
No meaning preservation (mis-reproduction)	170	37.4
Total	**455**	**100,0**

Structural change in experimental data

19



Substitutions

CORPUS: 60 cases 44,4%, EXP: 205 45.1%

20

- Most substitutions involve content words
  - 10% function words (19% in experimental data)
  - 10% phonetic variants
- Most involve single words
  - 11.7% involve more than one word
  - 13.3% involve proper nouns
- Most affected grammatical categories are:
  - nouns 21.7%, prepositions 16.3%, proper nouns 15%, verbs 8.3% in corpus data
    - Verbs 3.1%, prepositions 19.0%, pronouns 16.8%, nouns 14.6% in experimental data
  - **Grammatical categories in error corresponds to target**
    - Substitution noun for a noun, etc.

Deletions

CORPUS: 54 cases 40%, EXP: 199 43.7%

21

- Deletion involves single words in most cases
  - 20.4% involve more than one word
- Deletions involves
  - 37% function words
  - 5% phatic expressions and 5% repetitions and hesitations
- Most deleted grammatical category at single word level: conjunctions 20.9%, nouns 14% and adverbs 14%
  - In experimental data: adverbs 22.5%, verbs 20%, conjunctions 16.3%

Additions

CORPUS: 18 cases 13.3%, EXP: 40 8.8%

22

- Insertion of words always preserves meaning
  - 100% preservation in corpus data
  - 90% in experimental data
- Addition involves 44% function words
- And regards single words in 83.3% of the cases
- Grammatical categories inserted are:
  - 22,2% conjunctions, 11,1% articles, 11,1% pronouns
    - This corresponds to experimental data
    - *e* "and" being the most common insertion

Movement

CORPUS: 3 cases 2.2%, EXP: 11 2.4%

23

- Movement is rare both in corpus and experimental data
- It generally does not affect utterance meaning
- It generally involves entire phrases or fragments, rarely single words

Conclusions and questions

24

Mis-transcription

- hints on human understanding and **creative repair and filter** in a linguistic re-production task

Repair or editing

- tends to **preserve utterance meaning** (50 to 65% of errors) , but still there is a large amount of misunderstanding in mis-transcription (35-37%)
- Tends to be **meaning-centred**
- produces **grammatical sentences**
- is thus hardly detectable without access to audio
- corpus and experimental data tend to agree in the relative frequencies distribution of structural changes

Conclusions and questions

25

Conversion from speech to writing

- more explicit cohesive markers
  - (deletion of repetition, especially those representing hesitation or insertions of the *e* "and" coordination)
- error correction
  - (agreement reconstruction, or the redundant expression *a me mi dispiace* becoming for the transcriber *a me dispiace*)

Weak elements in a spoken discourse

- more often subject to deletion or repair during transcription
  - *anche* ("also") is systematically deleted, conjunction *e* ("and"), *quindi* ("so"), etc.

Guidelines improvement

- planning of instruction guidelines supplied to transcribers (training the ears and training the mind towards formal and superficial linguistic elements)
- improvement in the correction and revision phases during corpus processing and annotation.

isabella.chiari@uniroma1.it

Thanks

26

slides soon on: ***WWW.ALPHABIT.NET***

REFERENCES

Chiari, I. 2006. "Slips and errors in spoken data transcription", *Proceedings of 5th International Conference on Language Resources and Evaluation* LREC2006, Genova: ELDA, pp. 1596-1599.

Chiari, I. 2006. "Spoken corpora and transcription errors/ Звучащие корпуса и ошибки транскрипции", in А.С. Герд, В.П. Захаров, О.А. Митрофанова (eds.), *Труды Международной научной конференции «Корпусная лингвистика 2006» / Proceedings of the International Conference «Corpus Linguistics 2006»*, – СПб.: Изд-во С.-Петерб. ун-та, pp. 219-223.

Chiari, I. forthcoming. "What do we do when we transcribe speech? Typologies in lexical substitutions", in C. D. Pusch e W. Raible (eds.), *Romanistische Korpuslingustik III: Korpora und Pragmatik*), Tübingen: Gunter Narr Verlag.