"*Not as often as I thought*" – the problems of frequency and transcription in corpora

Richard Chapman,

University of Ferrara

The importance of corpus linguistics

* ***The impact of corpora in the field of applied linguistics over the last twenty years has been enormous, transforming both how we understand and how we study language*** – Hyland, Huat and Handford (2012)

* ***[corpora] have given the researcher access to huge amounts of real language in use and have enabled us to identify patterns of form and meaning on an unprecedented scale*** – Prodromou (2008)

Significant recent developments

* specific as well as general language corpora
* technological advances: memory and hardware costs
* increasingly refined methods of data analysis
* the era of '*big data*': *c.f.* the use of Google and changes in the balance of human/computerised communication
* use in many social/political situations (teaching, CDA)
* developing critical discourse in corpus linguistics:

  e.g. Widdowson; Swayles; Stubbs; Flowerdew; Baker; Hunston *et al*.

Some methodological developments

* **corpus-based *vs.* corpus-driven investigation**
* **levels of annotation**
* **increased interest in spoken corpora**
* **very large, general language corpora *vs.* small-scale, specific corpora**

Critical issues in corpus linguistics

* ***token vs. type*?**	**– hidden conventions**
* **difficulties in counting**	**– some subjective elements**
* **text collection / selection**	**– omissions**
* **levels of objectivity**	**– claims and criticisms**
* **the *monolingual fallacy***	**– in various guises**
* ***langue* or *parole*?**	**– stated research aims**
* **binary oppositions**	**– erasure?**

“Recurrence of words isn’t repetition. Ever ”

– Glyn Maxwell

* “Each text (as an utterance) is individual, unique, and unrepeatable, and herein lies its entire significance”
* “…the reproduction of the text …… is a new, unrepeatable event in the life of the text”
* “Two or more sentences can be absolutely identical…. But as an utterance (or part of an utterance) no one sentence, even if it has only one word, can ever be repeated”

– all from Bakhtin 1986, pp. 105-107

Repetition, again

* ***"Never, never, never, never, never!"***

  King Lear, *V. iii. 324*

* ***"What looks like repetition isn't repetition"***
* ***"There is no repetition in poetry"***
* ***"What's intervened … is the need to say the same again"*** - Glyn Maxwell

A *type* or just another *token*? An example

LemmyCautions Trabzonlu
08 October 2014 4:42pm

Recommend
0

I guess to you Turks and your logic: There was no Armenian genocide, millions of Armenian just vanished.

There are no 50 millions Kurds living in a land called Kurdistan.

And ISIS are Angeles.

And the moon is made of Chess.

Not to mention that Pigs fly.

Date: 08/10/2014; Guardian website, weblog/comments following article about Kobane:

http://www.theguardian.com/world/2014/oct/08/new-air-strikes-isis-pressure-win-kobani-syria-kurds-turkey

Repetition as cohesion:

"The reactive text being created" (Bakhtin)

"Something always has been said before" (Teubert)

WyldeWolfe LukeRijnhurt

21 July 2014 8:21am

Recommend

117

> , Israel agreed to cease fires 3 times,

It's not a ceasefire but surrender demand but even if it was negotiating a ceasefire with anyone other than your opponent is incredibly dimwitted.

> Hamas also increased its invasions, Hamas invasions to Israeli villages near the border don't count?

No it doesn't. They didn't do so until after Israel invaded. And I hate to break it to you sunshine but the idf are prefectly legit targets and hitting them from behind is a perfectly legit tactic.

Repetition is not a binary

* risk of assuming simplistic view of linguistic behaviour
* Derrida's process of *erasure*
* Saussure's *Platonism* and Chomsky's *Cartesian Linguistics*
* *e.g.* the written *vs.* spoken dichotomy:
  – now all the more in doubt because of digital voice recording, emails, *sms* and tweets
  – but also the case with speeches, play scripts *etc.*

Transcribing the indescribable?

"Any kind of transcription is, by definition, an abstraction in which a complex phenomenon is reduced to some of its constitutive features and transformed for further analysis"

"Another problem with standard orthography is that it does not do justice to certain paralinguistic phenomena, including sound play" – Duranti, 1997

"A transcript, however, is never equivalent to the speech event itself. … we have already moved from speech to written language" – Teubert, 2007

* transcription is the first step in analysis: it includes ***selection*** and ***simplification***
* no-one denies the challenges of transcribing spoken language
* ***Conversation Analysis*** might give substantial insights and methodologies for corpus data annotation

An ethnographic element to transcription?

* Bateson's ***tone of behaviour***: human posture, gesture, intonation and laughter
* visual anthropology and technology may make a fuller description of linguistic behaviour possible
* **"It is now universally accepted that in face-to-face interaction what humans say to one another must be understood vis-à-vis what they do with their body and where they are located in space" – Duranti, 1997**
* we need information as to eye-gaze, gesture etc. in real time simultaneity with utterances

Issues in text selection

* ***balance*** is ultimately a subjective judgement
* ***criteria of text selection*** must be transparent, but …

  * commercial problems and self-censorship;
  * the observer's paradox;
  * poetic uses of language;
  * short texts and interchanges;
  * in science sampling is rarely truly representative (*c.f.* the 'basket of groceries' in economics), but aims at statistical reliability
  * should texts be **weighted** ? (not just in *design*, but in *analysis*)

The hermeneutics of text selection

Difficulties and limits: subjectivity in data collection

* the sheer quantity of text compels selection
* selection risks being arbitrary – an interpretation
* chosen texts aim to represent the infinite
* both the selected body of texts and their analysis must remain limited and provisional
* is our approach to text merely ***emic*** or also ***etic***?
* ***"It is the researcher who in the end decides whether two texts are semantically linked or share a certain function"***

Teubert, 2007

It's been a hard day's search

* The British National Corpus is a general, monolingual British English corpus:

  a snapshot of British English in the 1990s

* includes novels and 'imaginative writing'
* a quick, simple search of the corpus using the titles of Beatles songs produced very limited solutions:
* ***It's been a hard day's night*** – 1 (in a conversation)
* ***I am the walrus*** – 2 (indirect references)
* ***Yellow submarine*** – 8 (all indirect references)

The historical dimension:

"so accentuating yesterday's truth to make it appear today's" (Voloshinov)

* most, though by no means all, corpora describe themselves as ***synchronic*** (*e.g.* the BNC)
* but we should never ignore their essentially ***historical*** character:
  – probably produced a number of years after data collection
  – collected over an extended period ('*the 1980s' etc.*)
  – often 'topped up' at a later date (*e.g.* BNC in 1990s and 2014)
  – corpora tend to have a long 'shelf-life' (BNC still in use)

Modest proposals for future corpora

 # 1 – description and *a priori* aspects

* greater transparency and display of corpus design
  – especially ***selection criteria***, ***balance*** and ***exclusions***
* description of corpora by *text*- rather than *word*- count
* hedges and modest claims (tendencies not facts)
* beware of 'general language' as a corpus description
* care with the term 'corpus-driven'

Modest proposals for future corpora

 # 2 methodological aspects

* the need for increased triangulation:
  – with other corpora; with *Google* (?!!); with intuition
* negotiation of the relationship between ‘real language’ in context and authentic text for data analysis
* losing our commercial/legal virginity?
* small, specific and contextualised corpora to work alongside grand, large-scale (big budget) corpora
* improved annotation
* video-based corpora?

Modest proposals for future corpora

 # 3 methodological aspects continued

* size can make a difference
* wider collection of data, especially email/blogs/tweets
* development of the critical discourse concerning CL
* awareness of the complexities of higher-level language analysis and CL – pragmatics and CDA *etc.*
* embracing the ethnographic element, despite difficulties
* recognition of varieties, repertoires and systems of use (Bakhtin: *raznorecie*; Blommaert: *repertoires*)
* understanding the theoretical basis (*empiricism* etc.)

Modest proposals for future corpora

4 some conclusions

* need for a *hermeneutics* for corpus analysis (Stubbs)
* CL tells us what is **probable** in language, but we also need to *interpret* these probabilities
* do we need to re-evaluate the conventions of written language?
* presentation of interrogations should rarely be nodes and accompanying *n*-items to left and right
* ***"We must monitor our own procedures"*** – Duranti, 1997

A brief bibliography

* Bakhtin, M.M. *Speech Genres and Other Late Essays.* 1986, University of Texas Press
* Blommaert, J. *Discourse*. 2005 CUP
* Bruner, J.S. *Acts of Meaning.* 1990, Harvard University Press
* Duranti, A. *Linguistic Anthropology*. 1997, Cambridge University Press
* Hoey, M.,Mahlberg,M., Stubbs, M., Teubert,W. *Text, Discourse and Corpora.* 2007, Continuum
* Hyland, K., Chau, M.H., Handford, M. (eds.) *Corpus Applications in Applied Linguistics*. 2012, Bloomsbury
* McEnery, T., Xiao, R. and Tono, Y. *Corpus-based Language Studies*. 2006, Routledge
* Prodromou, L. *English as a Lingua Franca. A corpus-based analysis*. 2008, Continuum

Some quotations

* *A transcript is, however, never equivalent to the speech event itself. We lose the immediacy of the situational setting, the embeddedness in the lifeworld, the intonation, the facial expressions and the gestures, the acts of deixis and of ostentation.*
* *"The prime object of corpus linguistics is, in my view, 'la parole'*
* *"By having to rely on a limited set of texts rather than on all the texts fulfilling our criteria, we lose a possibly critical amount of intertextual links."*
  – Teubert

* *"As researchers, we must understand the advantages and limitations of the analytical procedures we employ."*
* *"The extent to which a given phenomenon is seen as an occurrence of a more general category is partly due to our interpretative frame."*
  – Duranti