Corpus linguistics and its applications

Wolfgang Teubert
University of Birmingham
Email: teubertw@bham.ac.uk

Corpus linguistics and its applications

- Corpus linguistics and lexicography
- Corpus linguistics, parallel corpora and translation studies
- Corpus linguistics and grammar
- Corpus linguistics and language teaching/learning
- Corpus linguistics and critical discourse analysis
- Computational linguistics and the corpus (human language technology)

Corpus linguistics (CL): What is different?

- CL investigates the discourse and not people's minds.
- The discourse consists of all the texts of a discourse community.
- The focus of CL is on meaning.
- Meaning is in the discourse.
- The word is not the core unit of meaning.
- What is a lexical item (i.e. a unit of meaning or a translation unit) depends on our goals.
- The discourse has a diachronic dimension.
- The discourse is unpredictable; meaning is always provisional and never stable.
- The discourse is auto-referential: we are told what lexical items mean.
- The corpus is a sample of the discourse that suits a given purpose. There is no all-purpose corpus.

Corpus linguistics and lexicography (I)

The monolingual dictionary

- The COBUILD project
- The Western concept of the word
- Sinclair's choice vs. co-selection principle
- Collocation and the corpus: statistical significance and semantic relevance
- The lexical item is the unit of meaning
- The disappearance of ambiguity
- A dictionary of lexical items?

Corpus linguistics and lexicography (II)

The bilingual dictionary

- Ambiguity from the target language perspective
- Translation into a non-native language
- Parallel corpora as translation practice: translating unambiguous units
- The lexical item is the translation unit
- The translation unit is unambiguous
- For each translation unit, there is only one equivalent
- The dictionary of translation units and their target language equivalents: the *TranslationBase*

Corpus linguistics and lexicography (III)

Meaning and language use: collocation profiles

Travail: **work; labour**

Collocation profile: The statistically most significant context words in a window of –5 / +5

- **Collocation profile: travail: work**
- **Collocation profile: travail: labour**

Corpus linguistics and lexicography (IV)

travail: work		travail: labour	
Programme	410	Marché	747
Commission	255	Ministre	170
Conseil	212	Marchés	151
Cours	123	Sociales	125
Organisation	122	Affaires	117
Préparatoires	113	Emploi	88
Vue	109	Forces	65
Groupe	108	Normes	60
Temps	99	Femmes	60
Securité	97	Sociale	50

Corpus linguistics and lexicography (V)
Translation using collocation profile

work: La réforme du fonctionnement du Conseil soit opérée indépendamment des TRAVAUX préparatoires en vue de la future conference intergouvernementale.

labour: La Comité permanent de l'emploi s'est réuni aujourdhui sous la présidence de M. Walter Riester, ministre fédéral du TRAVAIL et des affaires sociales d'Allemagne.

Corpus linguistics and lexicography (V)

The reflexivity of the discourse: meaning is paraphrase

- Meaning is in the discourse
- Paraphrase: explanation, discussion, definition of lexical units and the discourse objects for which they stand
- Paraphrase: neologisms, concepts under discussion
- The example of google hits for ‘friendly fire means’

Corpus linguistics and lexicography (IV) Paraphrase for 'friendly fire means'

"Enemy fire" means bombs that come from the enemy. "Friendly fire" means bombs that come from the soldier's own army.

In military terms, 'friendly fire' means that you've caused damage to your own troops.

The military is legendary for its euphemistic lingo. "Friendly fire" means shooting your own troops.

Collateral Damage means "to accidentally blow up something of theirs." Friendly Fire means "to accidentally blow up something of ours."

Corpus linguistics and lexicography (VII)

Research topic:

The grammar of paraphrases

Corpus linguistics, parallel corpora, and translation studies (I)

- The issue of translation equivalence
- An ontological given or something created by a discourse community?
- Parallel corpora as repositories of translation equivalence
- Good translations and bad translations?
- Why are there so few parallel corpora?

Corpus linguistics, parallel corpora, and translation studies (II)

- I went down yesterday to the Piraeus with Glaucon. I wanted to make my prayers to the goddess. [D. Lee]
- I went down to the Piraeus yesterday with Glaucon, to make my prayers to the goddess. [F.M. Cornford]
- I went down to the Peiraeus yesterday with Glaucon. I wished to make my prayers to the goddess. [A. D. Lindsay]
- I went down yesterday to the Piraeus with Glaucon that I might offer up my prayers to the goddess. [F. M. Jowett]
- I went down to the Peiraeus yesterday with Glaucon, to pay my devoirs to the goddess. [W. H. D. Rouse]
- I went down yesterday to the Peiraeus with Glaucon, to pay my devotions to the goddess. [P. Shorey]

Corpus linguistics, parallel corpora, and translation studies: Topics

- Resolution of ambiguity in translation using the translation unit approach: A study in English–Greek translation
- Translation equivalence: a study of 12 English translations of Plato's *Republic*
- Translating EU legal documents into new languages: issues of consistency and standardisation

Corpus linguistics(CL) and grammar (I)

- CL and the laws of universal grammar
- CL and the rules of natural languages
- The arbitrariness of rules
- The issue of POS-tagging and syntactic annotation
- From a rule-based to a list-based approach
- From general grammar to local grammar
- Example: the use of prepositions

Corpus linguistics(CL) and grammar (II)
The use of the preposition *on*

Spatial use: *on* 'if someone or something is on a surface or object, the surface or object is immediately below them'

But:

- the paint on the wall
- he hit his head on the wall
- kiss her on the mouth
- ride on the bus / in a taxi
- on the road / in the street
- on the land / in the country

Corpus linguistics(CL) and grammar (III)
The use of the preposition *on*

CL evidence: lexical items preceding *on* (complements)

- *Tips on growing garlic*
- *The impact on the business*
- *She concentrated on the matter*
- *It is tough on young players*

But:

- *The belief in corpus linguistics*
- *The discussion about semantics*
- *The fight against SARS*
- *My arrangement with her*

Corpus linguistics(CL) and grammar (IV)
The use of the preposition *on*

CL evidence: lexical items following *on* (adjuncts)

- *She has ended on a high note*
- *I carry a penknife on any holiday*
- *A new club on the coast*
- *A trial on Monday morning*

But:

- *my ride in the taxi*
- *Her visit to her sister*
- *The path along the river*
- *He came under a wrong impression*

Corpus linguistics(CL) and grammar (V)
The use of the preposition *on*

CL evidence: lexical items containing *on*

- *The list goes on and on*
- *You want me to turn on the light*
- *They put on this air of normalcy*

But:

- *He put up with her*
- *He put her off*
- *He goes about his business*

Corpus linguistics and language teaching/learning

Using target language reference corpora (e.g. BNC)

- What is being used in the target language?

Using learners' corpora

- Underuse / overuse of features (e.g. modality / connectors / prepositions / idioms)

Using parallel corpora

- Analysing and explaining contrasts

Corpus linguistics (CL) and Critical Discourse Analysis (CDA) (I)

- CDA studies language as a cultural and social practice
- CDA attempts to discover the attitudes, beliefs and ideologies expressed in contributions to the discourse
- CDA investigates the political and economic conditions of participation in the discourse
- CDA is often reproached for its inherent subjectivity

Corpus linguistics (CL) and Critical Discourse Analysis (CDA) (II)

In CL, paraphrases will unravel underlying attitudes beliefs and ideologies.

By investigating intertextual clues, CL can identify ideological structures.

By investigating the traces texts leave in subsequent texts, CL can detect power structures .(Only 'powerful' texts leave traces.)

Corpus linguistics (CL) and Critical Discourse Analysis (CDA): Topics

- British Eurosceptic discourse
- Emotions in contrast: The English concept of sadness and its equivalents in Japanese
- US and Turkish diplomatic discourse: unilateralism vs. multilateralism
- Evaluation in the US Department of Defense discourse
- The concepts of property in the social encyclicals of the Catholic Church

Textual stance (ideology) and hermeneutics

- Ideology can be recognised only in comparison with other stances.
- This is why we have to relate texts to other / previous texts to which they refer.
- This is the hermeneutic art of interpreting texts.
- I should imagine the name Hermes has to do with speech, and signifies that he is the interpreter (*ermeneus*), or messenger, or thief, or liar, or bargainer: all that sort of thing has a great deal to do with language. (Plato: *Cratylus*)

Hermeneutics and the monitor corpus of social Vatican encyclicals: *property*

Private property, as we have seen, is the natural right of man. "It's lawful," says St. Thomas Aquinas, "for a man to hold private property, and it is also necessary for the carrying on of human existence" [1891, *Rerum novarum* § 22]

The natural right itself of owning goods ought always to remain intact and inviolate, since this indeed is a right that the state cannot take away. [1931, *Quadragesimo anno*, § 49]

Every man has in principle the right to use all the material goods of this earth, and this right can by no means be abolished, not even by other rights. [1941, Whitsun address].

The right to private ownership of goods has permanent validity. [1961, *Mater et magistra*, § 109]

Private property does not constitute for anyone an absolute and unconditional right. [1967, *Populorum progressio*. § 23]

The violation of the human right to ownership of property leads to lawlessness. [1991, *Centesimus annus*, § 24]

Referring to previous texts: The dangers of attribution

Private property, as we have seen, is the natural right of man. "It's lawful," says St. Thomas Aquinas, "for a man to hold private property, and it is also necessary for the carrying on of human existence" [1891, *Rerum novarum* § 22]

But:

Thomas Aquinas: "The distribution of property is not a matter of natural law". [1266-73 : *Summa theologica* Qu. 66, 2]

Computational linguistics and the corpus (Human language technologies)

- Knowledge management

1. Information retrieval
2. Knowledge building
3. Artificial Intelligence

- Machine translation

1. Statistics-based MT
2. Example-based MT

- Speech recognition

Knowledge management

- Corpora needed as testbeds
- Making sense of documents
- Building corpus-based ontologies for information retrieval
- Gauging knowledge building and innovation

Corpus linguistics and knowledge building

- Knowledge as discourse objects and what is said about them
- Is there discourse-external knowledge?
- Knowledge building and the discourse
- Knowledge building and innovation
- Knowledge building and emergent terminology
- Transfer from research genre to patent and textbook genre

Example-base MT

- Extraction of examples from parallel corpus (re-use of previous translations)
- Based on n-grams
- Normally without linguistic input (e.g. word order, POS-defined patterns, lemmatisation)
- Based on surface similarity
- Combines features of classical MT with TM

Statistics-based MT

- Require training on huge parallel corpora
- Parallel corpora are sentence-aligned and lexicon-aligned
- Normally rejects linguistic input
- Not concerned with meaning
- So far cannot produce high quality translations

MT based on *translation units* (I)

- Translation units: units translated as a whole, unambiguous
- Translation equivalent: the target language equivalent of a translation unit
- Translation units and their equivalents extracted from large parallel corpus
- Using linguistic knowledge (POS, phrases, fixed expressions, collocation etc.)
- Provides solution to problem of ambiguity

MT based on *translation units* (II)

- MT of the *Periodico de Catalunya*
- Translation of unrestricted text
- Nearly 100% satisfactory results
- Closely related languages
- Replaces source language phrases by target language phrases (units of up to six words)
- Huge database of translation units and their target language equivalents
- Requires large team of lexicographers (at least initially)

Conclusions

- CL looks at language in a new way
- CL can produce better monolingual dictionaries
- CL can help us with translation
- CL gives us a new perspective on grammar
- CL can improve language teaching
- CL provides a methodology for discourse-oriented social and cultural studies
- CL provides solutions to the problems of human language technology