HG3051 Corpus Linquistics

Contrastive and Diachronic Studies

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 9
https://github.com/bond-lab/Corpus-Linguistics

Overview

- Revision of Case Studies
  - Possessive Pronouns
  - Classifiers
- Contrastive Studies
- Diachronic Studies

Revision of Case Studies

Possessive Pronouns in Japanese contrasted with English

- Introduction
- Possessive Expressions in Japanese and English

(1)

Kanji:	私は	舌を	噛んだ
Jap:	*watashi-wa*	*shita-wo*	*kanda*
Gloss:	I-TOP	tongue-ACC	bit
Eng:	'I bit my tongue'		

- Differences in Noun Phrase Structure
- Pragmatic Analysis
- Application to Machine Translation
  Proposed method for generating possessive pronouns
  Experimental Results
- Conclusion

Application to Machine Translation

- Mark nouns that head English noun phrases with possessive determinatives where there is no possessive expression in the Japanese in the lexicon (possessed-nouns)
  - 205 different possessed-nouns (MT test set)
  - heading 825 noun phrases
  - 359 (44%) translated with possessive pronouns

- Mainly nouns that denote `kin, body parts, work, personal possessions, attributes` and `people defined by their relation to another person`

- Which nouns need to be marked is language specific, and probably register and domain specific as well.

Corpus-based Study of Distribution

Type:		MT Test set		News reports	
		No.	%	No.	%
I	English Idiomatic Possessive	105	16%	35	19%
II	Possessive Expression in Japanese	193	30%	5	3%
III	No Possessive in Japanese	359	54%	176	78%
Total:		657		181	

- Two Corpora
  - NTT MT Test set (6,200 sentences, 15,000 NPs)
  - Nikkei News Reports (1,382 sentences, 8000 NPs)

- Matched English:
  `[Mm]y|[Yy]our|[Hh]is|[Hh]er|[Ii]ts|[Tt]heir|[Oo]ur`
  Then hand checked Japanese for translation (on paper with colored pens!)

English:

1. Possessive determinative contrasts with articles
   — equivalent effort
2. Use of indefinite article implicates not owned
   — unless 'possession' predicated by verb
3. Use of definite article implicates more restricted reference
4. $\Rightarrow$ Use possessive determinative if relevant
   — unless 'possession' predicated by verb (don't be more informative than is required)

Japanese:

1. Possessive expression requires extra effort

2. Don't use by default
   — interpretation is that subject is antecedent

3. $\Rightarrow$ Use possessive expression to contradict default

4. $\Rightarrow$ Use possessive expression to emphasize default

Translating NPs headed by possessed-nouns

1. A noun phrase that fulfills all of the following conditions will be generated with a default possessive determinative with deictic reference determined by the modality of the sentence it appears in.
   (a) The noun phrase is headed by a possessed-noun that denotes `kin` or `body parts`
   (b) The noun phrase is the subject of the sentence
   (c) The noun phrase is referential
   (d) The noun phrase has no other determiner
2. A noun phrase that fulfills all of the following conditions will be generated with a default possessive determinative whose antecedent is the subject of the sentence the noun phrase appears in.
   (a) The noun phrase is headed by a possessed-noun
   (b) The noun phrase is not the subject of the sentence
   (c) The noun phrase is referential
   (d) The noun phrase has no other determiner
   (e) The noun phrase is not the direct object of a verb of `possession` or `acquisition`

Experimental Results

Results of the generation of all noun phrases headed by possessed-nouns in the MT test set (Total 752 noun phrases).

Result	Not generated	Generated
Good	I hit him in the face	I hid my face
Bad	I scratched a face	I lost my face

Result	Possessive determinative	MT-93 NPs	MT-93 %	MT-94 NPs	MT-94 %
Good	Not generated	429	57%	346	46%
	Generated	0	0%	263	35%
	— Total	429	57%	609	81%
Bad	Not generated	323	43%	60	8%
	Generated	0	0%	83	11%
	— Total	323	43%	143	19%

Conclusions

1. Possessive determinatives are used in English even when there is no equivalent possessive expression used in Japanese

2. This can be explained by the fact that in English possessive determinatives function as determiners, while in Japanese the possessive construction is an optional modifier phrase

3. 'possessed-nouns' can be identified in English that act (imperfectly) as cues

4. Implementing an algorithm that uses possessed-nouns in the Japanese-to-English MT system **ALT-J/E** generated possessive pronouns with an accuracy of 81% (up from 57%) and precision of 88%.

5. Should also be applicable to other under-specified generation: AAC.

How do we count Email in Japanese?

- Japanese has two main classifiers for counting messages:
  - 通 *tsuu*: used for letters
  - 件 *ken*: used for incidents
  - 本 *hon*: used for phone calls
  - コール *call*: used for phone calls

Year	1996	1998	2000	1002	2004
Email Usage	5%	11%	34%	81%	86%
Classifier	通, 本	通	通	通, 件	通, 件
SMS Usage	—	39%	45%	67%	76%
Classifier		通, コール	通	通	件, 通

Change in Classifier use with increased familiarity
Classifiers listed in frequency order

Contrastive Studies

Contrastive linguistics

- a linguistic approach that seeks to describe the differences and similarities between a pair of languages

- Some of the many applications:
  - to avoid interference errors in foreign-language learning
  - to assist interlingual transfer in the process of translating texts from one language into another
  - to find lexical equivalents in the process of compiling bilingual dictionaries

- generalized to the differential description of one or more varieties within a language, such as styles (contrastive rhetoric), dialects, registers or terminologies of technical genres

Three Common Approaches

- Look at examples in a translation lexicon
  - well aligned sample
  - x no frequency information
- Look at examples in comparable corpora "similar genre but not bitext"
  - good for getting general trends
  - x hard to measure **comparability**
- Look at examples in aligned corpora "bitext"
  - can get numbers at a very detailed level
  - x translated text differs from monolingual text
  - x translations can be very free

A Corpus based Comparison of Satellites in Chinese and English

- two balanced corpora (BNC and Academia Sinica Balanced Corpus of Modern Chinese) were used to compare satellites in Chinese and English

- Satellites:

  E verb+particle *go out, look up*
  C verb+complement *fei chulai* "fly out from, lit: fly exit-come"

- more satellites used in Chinese than in English

- mainly due to use as aspect markers

Comparable Corpora

- British National Corpus (BNC)
  - 100 million words
  - balanced
  - 10% speech
- Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus)
  - 5 million words
  - balanced
  - 10% speech
- Compared 1,000 sentences randomly selected from each
  - 500 Chinese satellites
  - 300 English satellites

English and Chinese Satellites

English	Frequency	Chinese	Gloss	Frequency
out	55	***lai***	come	75
up	31	***qu***	go	55
in	28	***chu***	exit	52
back	27	***chulai***	exit come	36
down	23	***dao***	arrive, achieve	35
into	20	***shang***	ascend	32
on	17	***qilai***	rise come	32
through	14	***zou***	walk, away	21
away	13	***qi***	rise	18
off	13	***zhu***	hold	on
		xia	descend	16
		kai	open	14
		shangqu	ascend go	12

Types with Token frequency > 10

Discussion

- More double satellites in Chinese
  None in the English top ten
- More varied semantics for the Chinese:
  fulfillment, underfulfillment, overfulfillment, antifulfillment and other events
- The satellites themselves are only vaguely comparable
- FCB comments:
  - I would liked to have seen some examples from a parallel corpus

Task

- In a language you speak, try to come up with examples of verb+satellite
- Try to find some examples of verb-satellite from the NTU-MC (or another corpus such as OPUS)
  - try to find at least one example with verb-satellite translated as verb-satellite

Contrastive connectors in English and Chinese

- a comparative study of ***however*** and its Chinese counterparts
- in two translation corpora
  - the HLM parallel corpus
  - the Babel English-Chinese Parallel Corpus
- a good example of comparison with parallel text
  - a deep analysis of a small number of examples

English-Chinese Parallel Corpora

- HLM Parallel Corpus: *Hóng Lóu Mèng*
  - Two complete English translations
    - "The Story of the Stone" David Hawkes and John Minford (literal)
    - "A Dream of Red Mansions" Gladys Yang and Yang Hsien-yi (free)

- Babel Corpus (no longer online)
  - 327 English articles and their translations in Mandarin Chinese
  - 544,095 words (253,633 English words and 287,462 Chinese tokens)
  - half from *World of English* and half from *Time* (2000–2001)

HLM Parallel Corpus

- Took 20 English sentences with ***however*** and compared to the Chinese translation and then the other English Translation
  - 13 no translation of ***however***
  - 2 到底 *daodi* "however"
  - 2 卻 *que* "of course"
  - 1 雖 ...卻 *sui ...que* "though ...never the less"
  - 1 任憑是什麼好的 "however good they are"
  - 1 人來客往 "however many guests"

Discussion

- 75% of the Chinese contrastive connectors are implied
- 90% of connectors are used between sentences; only 10% at clausal level
- The positional distributions of the contrastive connectors in these two languages differ considerably.
  - 85% of the Chinese contrastive connectors occur in the beginning of the sentence or clause
  - 52.5% do in English only; second initial position is also common.
  - The lack of contrastive connectors in Chinese compared to English is likely to be related to the frequent omission of subjects

Babel Parallel Corpus

- 101 sentence pairs
  - 4 no translation of ***however***
  - 38 然而 *raner*
  - 26 不过 *bu guo*
  - 11 但是 *danshi*
  - 7 但 *dan*
  - 4 可是 *keshi*
  - 6 others

Discussion

- 96% of the Chinese contrastive connectors translated
  - difference between literary and news style
- The positional distributions of the contrastive connectors in these two languages differ considerably.
  - 94.1% of the Chinese contrastive connectors occur in the beginning of the sentence or clause
  - 49.5% sentence initial, 38.6% second position.
- word order differences should be emphasized in teaching
- FCB comments:
  - different samples show very different results: sample wisely

A Parallel Corpus-based Study of Translational Chinese

- Compared English Text (EST), Original Chinese Text (OCT), Translated Chinese Text (TCT)

- Translational Chinese has the following features
  - TCT uses fewer monosyllabic words than OCT does
  - TCT tends to expand the normal load capacity of some Chinese constructions, which leads to longer sentence segments
  - compared with OCT, TCT uses more function words
  - TCT can change or expand the compositionality of some words or morphemes in Chinese.

- TCT use more types and longer segments than OCT. This does not support the hypothesis of lexical and syntactic simplification in translation.

Diachronic Studies

Historical Linguistics

- All Historical linguistics is corpus linguistics: we can not ask Old English speakers for grammatical judgments
- The texts of a historical period or a dead language form a closed corpus
  - can only be extended by the (re-)discovery of previously unknown texts
- For some languages you can use (almost) all of the closed corpus of a language for research
  - the *Theasurus Linguae Graecae* corpus contains most of extant ancient Greek literature.

Corpus-based historical linguistics

- Comparable corpora sampled over different times have made it possible to quantitatively study language change
  - An early, influential corpus is the Helsinki Corpus of English Texts: Diachronic and Dialectal
    The Corpus contains a diachronic part covering the period from c. 750 to c. 1700 and a dialect part based on transcripts of interviews with speakers of British rural dialects from the 1970's.
    http://khnt.hit.uib.no/icame/manuals/hc/index.htm
  - Recently the Corpus of Historical American English has been created
    COHA allows you to quickly and easily search more than 400 million words of text of American English from 1810 to 2009.
    http://corpus.byu.edu/coha/
  - Also http://ngrams.googlelabs.com/

The Helsinki Corpus

- approximately 1.6 million words of English dating from the earliest Old English Period to the end of the Early Modern English period
  - Old English (before AD 850)
  - Middle English
  - Early Modern English (to 1710)
- Each period is subdivided into 100 or 70-year sub periods
- The Helsinki corpus covers a range of genres, regional varieties and sociolinguistics variables such as gender, age, education and social class
- Two satellite Corpora:
  - early Scots English
  - early American English

The *by*-agent in English

- Peitsara (1993) used four subperiods from the Helsinki corpus to calculate the frequencies of different prepositions introducing agent phrases
  - In late Middle English (*c.* 1350) ***of*** and ***by*** were in roughly equal distribution (10.6:9)
  - By the fifteenth century ***by*** was three times more common than ***of***
  - By 1640 ***by*** was eight times as common

- There was marked influence of text type: statutes and official documents were much more likely to use ***by***
  - This is probably due to bilingual influence from French

Peitsara, K. (1993) "On the development of the by-agent in English", in *Early English in the Computer Age*. Rissanen, Kytö and Palander-Collin eds, 1993 pp 217-33, Berlin, Mouton de Gruyter.

The case of it's and 'tis

Peitsara, K. (2004) "Variants of contraction: The case of it's and 'tis" *ICAME* **28** pp 77-94

- two contracted forms of ***it is***
  - ***'tis*** (***tys, 't is, t is, t' is***)
  - ***it's***

- Contractions are more often used in speech than text
  - Their use or absence in text may reflect the editor's or printer's choice not the writer's actual usage
  - "the forms and structures of speech are best reflected in text categories that imitate speech, are addressed to a less educated readership, or are written in the less formal register and by less educated writers" (Kytö and Rissanen 1993: 12)

The case of it's and 'tis

'tis vs ***it's*** in the Early Modern English part of the Helsinki Corpus

	EModE1 (1500–1570)		EModE2 (1570–1640)		EModE3 (1640–1710)		Total	
	N	%	N	%	N	%	N	%
'tis	5	100	41	93.18	67	65.68	113	74.83
it's	—	—	3	6.82	35	34.31	38	25.17

The case of it's and 'tis

'tis vs *it's* in the prose part of the LION corpus (all spelling variants)

	-1600	-1650	-1700	-1750	-1800	-1850	-1900	1900-
'tis	82	286	2302	1950	2137	4641	1964	3
%	100.00	96.30	98.33	98.19	81.75	28.78	15.78	1.67
it's		11	39	36	479	11485	10484	177
%		3.70	1.67	1.81	18.25	71.22	84.22	98.33

- Big break between 1800 and 1850
- Considerable variation in individual writers
  - Thomas Hardy 90% *'tis*
  - American writers use *'tis* relatively more

Why the change?

- General trend from proclisis (initial) to enclisis (embedded)
  - OE and ME proclitic negation (***nam, nis, neren,*** etc.)
  - ModE enclitic contractions (***isn't, ain't, weren't,*** etc.).
- Perhaps in analogy to personal pronouns with be (***I'm, he's, she's, we're***, etc.), supported by other enclitic contractions (***we'll, he'd, etc.***).
- Note: ***'tis*** survives in the south-western varieties of British English and also in Newfoundland English

Issues with Historical Linguistics

Rissanen (1989) identifies three main problems associated with using historical corpora

- The philologist's dilemma —the danger that the use of a corpus and a computer may supplant the in-depth knowledge of language history which is to be gained from the study of original texts in their context

- The God's truth fallacy — the danger that a corpus may be used to provide representative conclusions about the entire language period, without understanding its limitations in the terms of which genres it does and does not cover

- The mystery of vanishing reliability — the more variables which are used in sampling and coding the corpus (periods, genres, age, gender etc) the harder it is to represent each one fully and achieve statistical reliability. The most effective way of solving this problem is to build larger corpora (if possible)

Rissanen, M. (1989) "Three problems connected with the use of diachronic corpora", ICAME Journal 13: 16-19.

Acknowledgments

- Historical Linguistics examples taken from Chapter 4 of Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. CUP

- Several papers from Richard Xiao, Lianzhen He, and Ming Yue (2008) *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008)*, Alberta
  www.lancs.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/