Words, texts and genres: Applications of corpus linguistics

Ramesh Krishnamurthy &
Emmanuelle Labeau

Section 1

The Taming of the Shrew

1. WHAT?

- The primacy of aspect hypothesis

	STATES 'had'	ACTIVITIES 'played'	TELIC EVENTS 'taught x to y'	PUNCTUAL EVENTS 'broke in two'
1	*tiene*	*juega*	*enseña*	*se parte*
2	*tiene*	*juega*	*enseña*	*se partió*
3	*tenía*	*juega*	*enseña*	*se partió*
4	*tenía*	*jugaba*	*enseño*	*se partió*
5	*tenía*	*jugaba*	*enseño* *enseñeba*	*se partió*
6	*tenía*	*jugaba* *jugó*	*enseño* *enseñeba*	*se partió*
7	*tenía*	*jugaba* *jugó*	*enseño* *enseñeba*	*se partió* *se partia*
8	*tenía* *tuvo*	*jugaba* *jugó*	*enseño* *enseñeba*	*se partió* *se partia*

Testing of the AH

- Learners use perfective past marking (e.g. Preterite) first on telic [situation with an inherent endpoint] events and they later extend its use to verbs from other lexical aspectual classes
- The imperfective marker (e.g. Imperfect) will appear later than the Preterite in association with atelic events (states and activities), eventually extending to telic events
- The use of periphrastic Progressive will initially appear in association with activity verbs and then extend to telic events
- The use of periphrastic Progressive will not overextend to stative verbs. (Shirai & Kurono 1998: 248–9)

Problems of the AH

- In general:
  - Only 4/8 stages supported by Andersen's data (1991:313-4)
  - Relevance of the sequence for advanced learners once any verb can be associated with any morpheme? (Kihlstedt 1998:43ff)
  - Application to acquisition in academic settings where input to identify prototypical markers of lexical aspect may be lacking (Salaberry 1998:532)
- For French: No specific marker of progressivity (unlike English and Spanish) $\rightarrow$ use of:
  - Periphrases (e.g. *être en train de)*
  - IMP

  $\Rightarrow$ Is the AH able to account for tense-aspect development of L2 French at advanced levels in academic settings?

2. HOW?

- **Research Design**
  - 61 learners at 3 levels of a degree
  - Control group of 6 native speakers
  - Pseudolongitudinal over 3 years
  - Range of data:
    - Written & oral film narratives
    - Cloze tests based on literary texts, learner's production and description of cartoons
    - Acceptability judgement tests

Pros of the design

- Choice of *Modern Times:*
  - Comparability:
    - Across subjects
    - With previous studies
  - Check of pragmatic accuracy
  - No interference from dialogues
  - Mixture of narration and description
- Variety of tasks (influence of medium)
- Cloze test:
  - Prevents avoidance (Bardovi-Harlig 1992a)
  - Provides a set of tokens across verb types (Salaberry 2000)
  - Offers directly comparable native data (Bardovi-Harlig 1992a)
  - Tests the difference between implicit and explicit knowledge

3. RESULTS

- Limitations of the AH:
  - PC stable across levels and near-native
  - Overuse of IMP in combination with perfective verbs at level 1
  - No conclusive evidence on progressive
- Discoveries
  - Influence of cotext (lexical environment): object, adverb, neighbouring verbal forms…
  - Influence of the context (medium, genre…)

If only…

The analysis was done without recourse to electronic parsing systems that could have sped:

A. The quantitative analysis as all tokens needed to be counted manually. It would have been handy to have **automatically sorted lists of forms**

B. The qualitative analysis of the cotext:

- Were certain **collocations** frequent (for example, *il y avait*)?
- Were **verbs in the imperfect** used with **direct object** that would make the use non native like (e.g. Le tram avait un accident)?
- Did some correlations between **tenses** and **adverbs** occur? (e.g. finalement il tombait)

And what about…? Further issues raised by Emmanuelle's PhD

In addition, access to electronic parsing could have **pedagogic advantages**:

• For the **language teacher**, tagging each verbal form for correctness of form and of function could have provided a direct insight into features of interlanguage and could have inspired remedial work.

• For the **language student**, access to parallel native narratives and to samples of various genres (history, obituaries, novels…) referring to past time would inform them about native usage. They could discover divergent practices in written and oral and a variety of tenses to express past time, such as the *présent historique* or the *futur des historiens…*

Section 2

Great Expectations

As if in answer to a maiden's prayer… Corpus Linguistics: a brief history

- **A young discipline** (well, c. 50 years old)
- **Computers for text**, not just for numbers
- **1960s – small corpora** (1m words) – language description – **grammatical studies**
- **1980s – medium-size corpora** – (20m words) – language description – **lexical studies**

  (UK ELT dictionaries)
- **1990s – large corpora** (100m+ words) – many languages, genres, varieties, domains – general and specific – language learning and teaching – computational linguistics (information retrieval, machine translation, automatic summarization, text and speech generation)

Corpus Linguistics: basic methodology

- **Tokenization**: to identify the objects of study (words, phrases, etc)
- **Frequency lists**: to place the objects in a hierarchy (frequency = importance)
- **Concordances**: to examine the detailed behaviour and co-text of the objects
- **Collocation**: to quantify the co-text of the objects, observe sequences and patterns

Corpus Linguistics: discoveries in English lexis: impact on EFL dictionaries

- Most frequent lexical words: '***time***', '***people***'
- lexical verbs used in 'delexical' senses
  (e.g. ***take a bath, make a decision***)
- '***thing***' and pronouns **'this' and 'that'** mostly refer to an abstract entity (e.g. a proposition or argument), not to a physical object: *A strange thing happened… Is that why you had a few days off?... This is why I'm opposed to the plan* (Willis)
- '***see***' mostly means 'understand' (esp in spoken: '***I see***')
- '***lamely***' is only used for excuses (not legs), '***crisply***' mainly with verbs of speech

Corpus Linguistics: discoveries in English grammar: impact on EFL teaching

- '***of***' is not a preposition (Sinclair) – e.g. rarely used in adjuncts

- **'*would*'** is used generally to talk about hypothetical events:

e.g. *I think The Tempest* ***would*** *make a wonderful film*.
This makes up almost half of its corpus occurrences. As a sub-category, **would** is used in conditional sentences:
e.g. *It would surprise me if sterling strengthened*.
But EFL courses present **would** as a part of the second conditional. (Willis)

- **the 'rules' for reported speech** ('present simple becomes past simple, present perfect and past simple become past perfect' etc) are totally unnecessary. Differences in tense, person, phrases of time and place occur because we are taking a different standpoint from the original writer or speaker. **This is a feature of language as a whole, Not a feature of reported speech**. (Willis)

Widespread current use of Corpus for grammar, language students, and teachers

Google searches: Nov 2005

(a) “corpora” + “grammar”: 407,000 hits

(b) “corpora” + “language learning”: 172,000

(c) “corpora” + “language teaching”: 81,000

(d) Even on very specific topics…

“learner corpora” + “collocation errors”: 14

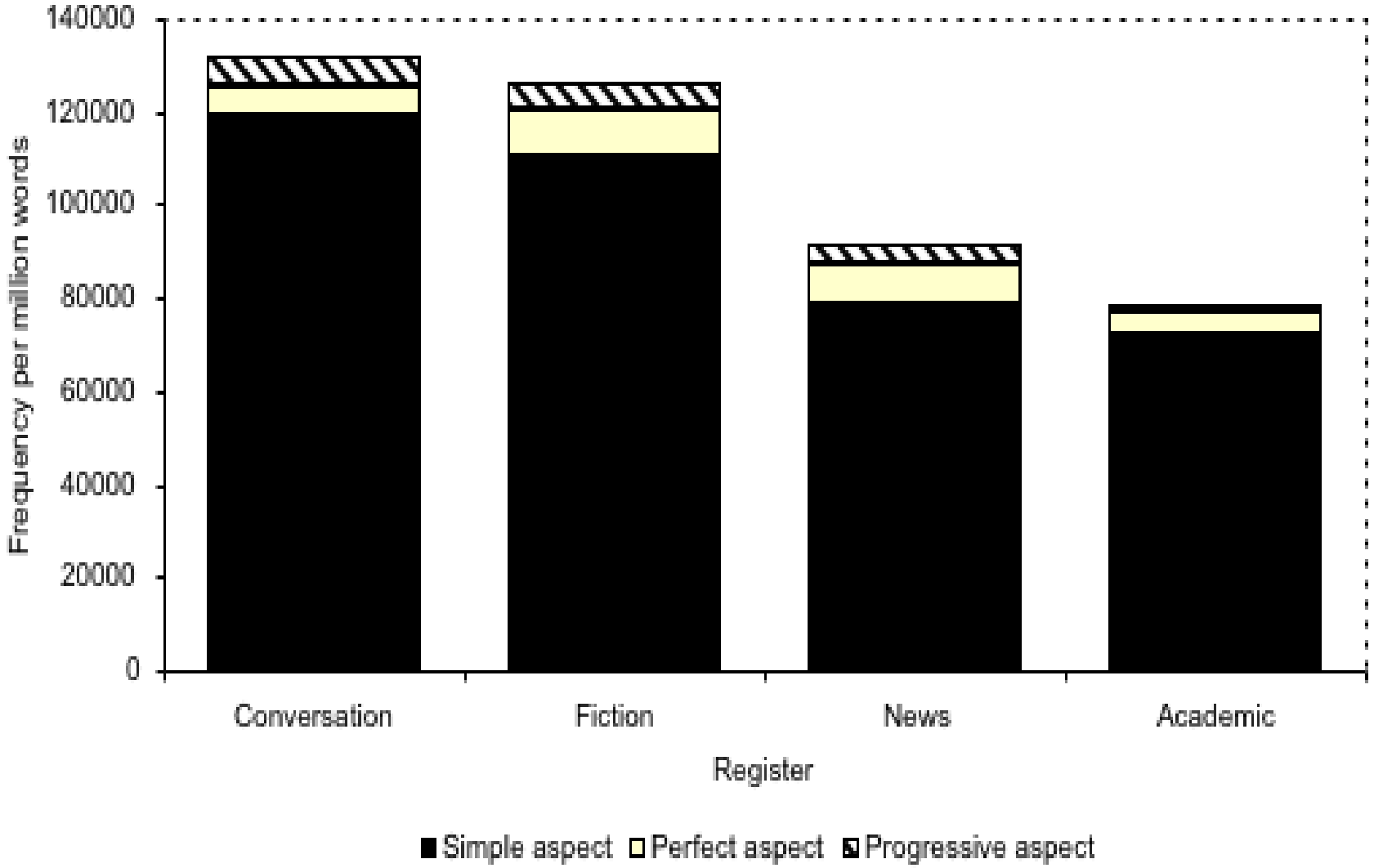
(a) Corpus used for grammar: ongoing discoveries (Biber, Nov 2005)

- As language teachers and professionals, we often have strong intuitions about language use… ***Corpus-based research, however, shows us that our intuitions are often completely wrong.*** In many cases, we simply do not notice the most typical grammatical features because they are so common.

- Many English teachers (including myself!) have had the **intuition** that ***progressive aspect verbs are the normal choice in conversation,*** and thus they should be ***more common than simple aspect verbs*** in that register.

- In this case, **corpus research shows that our intuitions are dramatically wrong**: As Figure 5 shows, **simple aspect verb phrases are more than 20 times as common as progressives in conversation.**

Figure 5: Frequency of simple, perfect, and progressive aspect in four registers (based on Biber et al, 1999, Figure 6.2)



(b) Corpus used by English language students

Mike McDonald (MSc TESOL participant):

"I was very impressed by the richness of their responses. I don't know how much they have internalised, but they seem to have **noticed an amazing number of details."**

'because'

- **many words after *because* are pronouns**
- 10 examples of ***because of***; often used after a main clause
- **main patterns**:
  - *[Fact] because (of) [reason]*
  - ***Because (of) [reason], [fact]***
  - ***Why . . . ? Because [reason]***

'however'

- The pattern is often *[Sentence.] However, [contrasting sentence.]*
- ***However* can be used at the end of a sentence (…*make their fortune at any price, however*.)**
- ***However much of a [noun]* is an interesting expression** *[not in Cobuild!]*

(c) Corpus used by teachers to analyse students' errors: general (Fei-Yu Chuang 2005)

- a corpus of 50 essays written by Chinese EAP (English for Academic Purposes) foundation students
- 5232 errors identified. Most frequent errors were:

- (1) Missing definite article 10.1%
- (2) Bare singular count noun for plural 8.8%
- (3) Redundant definite article 8.5%
- (4) Mis-selection of preposition 6.1%
- (5) Lexical misconception 5.8%
- (6) Wrong tense and aspect 3.8%
- (7) S-V non-agreement 2.4%
- (8) Wrong collocation 2.1%
- (9) Missing 'a'/'an' 2.0%
- (10) Comma splice 2.0%

(d) Corpus used by teachers to analyse students' errors: specific (collocation)

(Wible, Kuo, Chien, Liu, Tsao 2001)

- Taiwan Learners Corpus: 1 million words
- Mis-collocations categorized by part-of-speech and frequency

Mis-collocation Type	Frequency
V N	145
Adj N	25
V Adv	5
Adv Adj	2
Total	177

Section 3

Changing Places

Use of corpora in other disciplines

- Language-based
  - Word study: search, frequency, appearance, decline
    - E.g. lexicology, political concepts, use of sociologically relevant features (e.g. feminisation)…
  - Study of collocations
    - E.g. Foreign language learning, translation…
  - Semantic fields
    - E.g. Discourse analysis in literature, politics…
  - Reported speech
    - E.g. linguistic means, discourse manipulation
  - Variations
    - E.g. diachronic, diatopic…

- Genre-related
  - Monolingual or multilingual parallel texts to identify culture-specific structures
    - E.g. culturally appropriate translation or writing
  - Variation of features across genres or sources
    - E.g. registers, analysis of socially- or politically loaded concepts (e.g. How do some minorities refer to themselves vs press or authorities?)

1. Corpus and Literature: poetry

Spring.
Green-shadowed people sit, or walk in rings,
Their children finger the awakened grass,
Calmly a cloud stands, calmly a bird sings,
And, flashing like a dangled looking-glass,
Sun lights the balls that bounce, the dogs that bark,
The branch-arrested mist of leaf, and me,
Threading my pursed-up way across the park,
An indigestible sterility…. (Larkin, 1954)

REVIEW
***After he has established this idyllic but commonplace vision of Nature and humanity in harmony, he shocks us with the image of himself, 'an indigestible sterility'. It sounds awkward and convoluted after the smooth, then buoyant rhythms of the lines that preceded it.* (Hartley 2000)**

Ramesh's analysis

- the negative connotations are signalled from the first word: **'green-shadowed', dangled, arrested, threading, pursed-up**
- '**green**' and '**shadow**' do not collocate in the corpus; **'green'** does collocate with '**shade'**
- **'shadow**' has negative connotations; **'shade**' has positive connotations

Other features found

A lot of 's' sounds: arrested seasons see sings sit spring spring stands sterility sun mist most most best immodest indigestible glass grass gratuitous least lights use pursed (23 out of 99 words)

- Mostly short words compared to corpus
- '**untaught**' is the rarest word (i.e. most of the words are common)
- 4/5 of the **hyphenated forms** are not in the corpus
- 6 corpus **collocates of 'spring'** are in the poem

2. Corpus and Politics:
The word that won the 1997 UK General Election

- derived from a German adjective, meaning ‘from Silesia’, and referring to a thin, cheap cloth
- entered English language in C17
- developed a meaning ‘dirty, disreputable’: of places; then of people: clothes, appearance, character, behaviour
- the noun was coined by back-formation in mid-C20

The word that won the 1997 UK General Election (cont)

- was not in the OED (1971)
- did not occur in the 20m-word BCET corpus (1986)
- was in Collins English Dictionary (1986) and even the tiny Collins Gem (1985-7)
- was in the Bank of English corpus:

1990: 0.6 per 1m	**(c. 50m-words)**
1992: 1.9 per 1m	(c. 100m-words)
1995: 8.7 per 1m	**(211m-words)**

The word was **sleaze**: collocates 1997

***sleaze* collocated strongly with *Tory/Tories, government* (i.e. the Tory one)** … and Labour?

of	455	10.209692
allegations	69	8.264774
tory	67	8.111882
and	329	6.222366
nolan	36	5.987291
scandal	29	5.341044
government	40	5.090982
political	32	5.069198
party	31	4.798490
labour	25	4.576461
inquiry	21	4.470166
tories	20	4.406132

Sleaze 1997 collocate: Labour?

…Labour allegations of Tory ‘sleaze’…

…Labour had no truck with sleaze?...

…Labour's campaigning over sleaze…

…Labour Seeks Anti-Sleaze Act…

…Labour critics of sleaze…

…Blair: I'll sweep out the sleaze; Tony Blair; Labour Party leadership…

…veteran Labour campaigner against sleaze…

Sleaze: collocates 2002

oh dear… now Labour is on the increase!

of	565	11.027169
allegations	95	9.706463
tory	93	9.579130
anti	56	7.307552
scandal	49	6.957746
labour	51	6.808804
sleaze	38	6.158170
and	387	5.542236
party	37	5.269707
government	43	5.240159
tories	26	5.029989

But they still won the 2001 and 2005 elections!

…Now, when the Labour sleaze is appearing…

…With all this sleaze from the Labour Government…

…we can expect yet more Labour sleaze to float to the surface…

…at the height of Labour's sleaze crisis…

…banish Labour sleaze from Scottish local government…

…rows over Labour sleaze…

3. Corpus and Economics

Watch the video clip… and look out for corpus zealots!

4. Corpus and society: ‘ethnic, racial, tribal’ in newspaper articles (1991)

- Yugoslavia: ‘multi-**ethnic** state, six main **ethnic** groups and three major religions’
- Kenya: ‘**tribal** violence, **tribal** fighting, **ethnic** violence, **tribal** fighting, **tribe**smen’
- England: ‘**ethnic** recruiting rate, **ethnic minority** recruitment, Commission for **Racial** Equality, **black** recruits, **ethnic minorities**, **ethnic minority** police officers, **black and Asian** people, **ethnic minorities**
- South Africa: ‘mixed-**race** politician, inter-**racial** marriage’

- ‘tribe’ is never used for Yugoslavia or UK (NB humour)
- If ‘tribe’ is the ‘technical’ word for Kenya, why one reference to ‘ethnic’?
- In UK, ‘ethnic’ means ‘black’ or ‘black and Asian’ (not Irish/Welsh, or Italian/Polish)
- ‘impersonal’ refs to ‘help applications’ not ‘applicants’ in the UK article
- ‘impersonal’ use of numbers and statistics in UK/Kenya articles

COLLOCATION

ethnic:	racial:	tribal:
groups	discrimination	assembly
minorities	non	grand
and	multi	leaders
minority	equality	chiefs
of	and	and
violence	of	groups
in	south	killings
cleansing	africa	navajo
group	commission	in
albanians	ethnic	a

Section 4

All's Well that Ends Well

TWC Course

- Emmanuelle and Ramesh went to Italy for 5-day courses in July 2005
- Emmanuelle: Spoken and Written Corpora
- Ramesh: Creating Corpora from the Web
- We both found our course very intense, stimulating, but not immediately practical
  - Spoken – too technical and expensive
  - Web – technical, plus copyright problems

The ACORN project

- ACORN = **A**ston **COR**pus **N**etwork
- FLDC funding obtained for 1 year
- Aim to collect:
  - 5-10m. words each for English, French, German, Spanish (variety of genres)
  - 1m. words each of original texts+translations
- Design pedagogic software
- Ramesh, Chris, Sylwia, and Stefan
- We will of course be consulting LSS colleagues at every stage and on all aspects of the project, in order to make the data and software as useful as possible for staff and students, for teaching, learning, and research
- We intend to apply for further funding from other sources