A Corpus-based Contrastive Study of Optional Syntactic Omission in Two Varieties of Institutional Academic English

Federico Gaspari

fgaspari@sslmit.unibo.it

SITLeC Department

University of Bologna (Forlì campus), Italy

Outline

- Theoretical background
- Motivations and aims of the study
- Corpus description
- Patterns of optional omission
  - Results and discussion
- Conclusion and future work

Theoretical background

- From *translation* universals…
  - Baker (1993), Laviosa (2000; 2003), Olohan (2001; 2002)
    - normalization, levelling out, simplification, explicitation (Olohan & Baker, 2000: reporting *that* after *say* and *tell*)
- **Monolingual comparable** corpus consisting of translated language vs. native-speaker writing (e.g. TEC + BNC)
  - In the same language (English)
  - Originals and translations (from multiple SLs)
  - As closely comparable as possible
    - but differences in genre, register, level of formality, etc. inevitable

Motivations and aims of the study[4]

- … via contrastive (variational) linguistics…
  - Cortes (2004), Hyland (2008)
  - Biber (2006): registers of university language

- … to *mediation* universals
  - Ulrych & Anselmi (2008), Ulrych & Murphy (2008)

    Gaspari & Bernardini (2010)

Motivations and aims of the study

- Look at optional syntactic omission/explicitation in institutional, i.e. not (only) disciplinary, academic language
- The case of *that* as a complementizer
- Patterns of optional syntactic omission in mediated vs. native/original varieties of English
- Compare
  - mediated language, i.e. a conflation of L2, translated and variously edited English vs.
  - native/original language as a benchmark

Corpus description

- **acWaC**: monolingual comparable corpus of institutional academic English (Bernardini et al., 2010)

– **ENUNI**

(native/original benchmark):

- UK universities (Russell)
- Irish universities (all)

– **ITUNI**

(mediated variety):

- English sections of universities in Italy (preliminary manual search for English content)

Corpus description

- Building procedure
  - BootCaT (Baroni & Bernardini, 2004)
  - Google language identifier + filetype filter (- PDF)
  - Max 300 pages per website
  - POS-tagged and lemmatised (TreeTagger) + CWB-ed

- Size

	ENUNI	**ITUNI**
Tokens	5,435,855	4,228,841
Types	125,089	165,037
Documents	7,271	6,754
Websites	28	55

Optional omission

	ENUNI		ITUNI	
Occurrences of "that" complementizer	Abs. no.	pMw	pMw	Abs. no.
	13,824	2,542	1,790	7,575

Occurrences of pronouns	Abs. no.	pMw	pMw	Abs. no.
	74,799	13,775	9,483	40,117

Optional omission

- Verbs taking optional *that*-complement clauses in post-predicate position (Biber et al., 1999:661ff)

admit	confirm	hope	read	state
agree	decide	imagine	realise/ize	suggest
announce	demonstrate	indicate	recognise/ize	suppose
argue	doubt	insist	recommend	tell
assume	ensure	know	remember	think
believe	expect	mean	report	understand
bet	feel	notice	require	wish
check	find	observe	say	
clarify	guess	predict	see	
conclude	hear	prove	show	

Optional omission

- ≥ 100 occurrences pMw in both ENUNI and ITUNI

expect	know	recognise/ize	see	think
find	mean	require	show	understand
indicate	read	say	suggest	wish

- Focus on a specific pattern involving pronouns:

I) *that* retention (explicitation) within variable 4-slot window

A) lemma_verb + contiguous *that*_comp + PRON + VERB

B) lemma_verb + up to intervening 4 slots + *that*_comp + PRON + VERB

C) lemma_verb + up to 2 slots + *that*_comp + up to 2 slots PRON + VERB

II) *that* omission (zero connective) within 4-slot window

D) lemma_verb + Ø [NO *that*_comp] in a 4-slot span + PRON + VERB

Optional omission

expect	know	recognise/ize	see	think
find	mean	require	show	understand
indicate	read	say	suggest	wish

No occurrences at all of any of A-D patterns

Very severe data sparseness

Data sparseness

Each of the four patterns clearly represented

Results

"mean"	ENUNI		ITUNI	
Overall occurrences of the verb	Abs. no.	pMw	pMw	Abs. no.
	854	157	133	564
Retention (patterns A, B, C)	39	7.17 (4.56%)	3.07 (2.30%)	13
Omission (pattern D)	27	4.96 (3.15%)	1.41 (1.06%)	6
		— **1.41%**	— **1.24%**	

Results

"think"	ENUNI		ITUNI	
Overall occurrences of the verb	Abs. no.	pMw	pMw	Abs. no.
	1,299	238	111	470
Retention (patterns A, B, C)	11	2.02 (0.84%)	1.89 (1.70%)	8
Omission (pattern D)	52	9.56 (4.01%)	7.32 (6.59%)	31
		+ 3.17%	**+ 4.89%**	

Results

"say"	ENUNI		ITUNI	
Overall occurrences of the verb	Abs. no.	pMw	pMw	Abs. no.
	2,103	386	210	889
Retention (patterns A, B, C)	35	6.43 (1.66%)	6.38 (3.03%)	27
Omission (pattern D)	51	9.37 (2.42%)	10.63 (5.06%)	45
		+ 0.76%	**+ 2.03%**	

Results

"know"	ENUNI		ITUNI	
Overall occurrences of the verb	Abs. no.	pMw	pMw	Abs. no.
	1,818	334	292	1,237
Retention (patterns A, B, C)	20	3.67 (1.09%)	2.36 (0.80%)	10
Omission (pattern D)	28	5.14 (1.53%)	2.36 (0.80%)	10
		+ 0.44%	**=**	

Discussion

	Combined occurrences of the verb pMw in ENUNI + ITUNI	*Retention vs. omission, normalized*	
		ENUNI	**ITUNI**
"mean"	290	**- 1.41 %**	**- 1.24 %**
"think"	349	**+ 3.17 %**	**+ 4.89 %**
"say"	596	**+ 0.76 %**	**+ 2.03 %**
"know"	626	**+ 0.44 %**	**=**

Conclusion and future work

- Focus on same-subject in matrix and subordinate clause
  - does it play a role in the *that* omission vs. retention choice?
  - equal effect on native/benchmark and mediated language?
- Investigate in more detail
  - patterns involving NPs instead of pronouns
  - other corpora
  - fine-grained distinction across L2 / translated / edited language
- Extend to other SLs/L1s (with optional complementizer)
- Involve to other syntactic explicitation patterns, e.g.
  - *seem* / *appear* + [to be] + Adjective
  - *help* + [to] + Verb (McEnery & Xiao, 2005)

A Corpus-based Contrastive Study of Optional Syntactic Omission in Two Varieties of Institutional Academic English

Thanks!

Federico Gaspari

fgaspari@sslmit.unibo.it

SITLeC Department
University of Bologna (Forlì campus), Italy