A corpus-based study on the distribution of complements and adjuncts in learner language: will we reveal <major findings> <in this study> or will we reveal <in this study> <minor findings>?

Javier Pérez-Guerra (jperez@uvigo.es)
Ana Elina Martínez-Insua (minsua@uvigo.es)

***L**anguage **V**ariation and **T**extual **C**ategorisation* Research Group
University of Vigo

lvtc

LCR 2013, Bergen

The outline

- The goal
- The background
- The case study
- The data
- The analysis of the data
- The concluding remarks
- For further research
- The references

lvtc

The goal

• check the distributional tendencies affecting the placement of **adjuncts/modifiers** and **complements/arguments** in verb phrases (predicates)

Examples:

(1) Now I will deal [with the construction] [in a way which will lead to odd results].

(2) Now I will deal [in a way which will lead to odd results] [with the construction].

• determine whether the production of such constituents in English by non-native speakers is influenced or not by their first language (Spanish)

lvtc

3

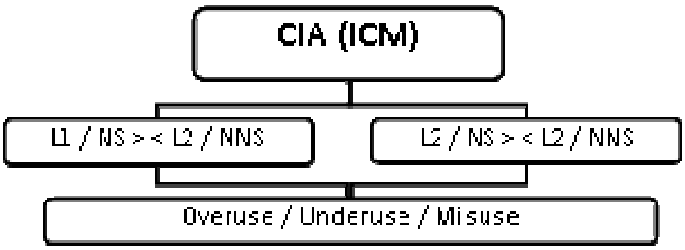
The background

- learner corpora as bases for studies of learner interlanguage (from late 80s / early 90s): new corpora, new tools, new annotation conventions, new theoretical approaches
- 'interlanguage' approaches (Eubank et al. 1997)
- 'approximative linguistic systems' (Nemser 1971)
- in general, Granger's (1996) comprehensive 'Integrated Contrastive Model' (ICM; see Gilquin 2008: 6–8) of interlinguistic analysis:
  - component where **reference** (native/proficiency) data from one language is compared with **reference** data from another language
  - an additional module where **reference** language is compared with an **interlanguage** variety of the language

lvtc

4

The background



Contrastive Interlanguage Analysis (CIA) (Granger 2002, 2009)

- CIA and the issue of comparability (later)

5

The case study

- previous studies on word-order alternations in L1 Spanish and L1/L2 English: Lozano and Medikoetxea (2008, 2010), on subject-verb inversion.

  They consider:
  - type of predicate (verb)
  - end-weight
  - information structure

  They show that the differences between L1 Spanish and L2 English are not so striking.

6

The case study

(1) Now I will deal [with the construction]$_{\textbf{complement}}$ [in a way which will lead to odd results]$_{\textbf{adjunct}}$.

versus

(2) Now I will deal [in a way which will lead to odd results]$_{\textbf{adjunct}}$ [with the construction]$_{\textbf{complement}}$.

lvtc

7

The case study

- **Complements**
  - semantically selected or subcategorised
    - Matthews (2007: 187): "unit in a construction either required or specifically taken by an individual member of a lexical category"
    - Matthews (1981: 124-127): impossibility of dropping (if dropped, then latent)
    - exclusion when the pattern is saturated
  - syntactic dependencies:
    - lexical restrictions or formal determination (Greenbaum et al. 1996: 76)
      {*deal*} + *with*-PP
      {*assume*, *hypothesise*} + *that*-clause
- **Adjuncts**
  - loose semantic connection between the adjunct and the head => not required

lvtc

8

The case study

- **Competing forces:**

  (i) syntactic: complements-first

  (ii) processing: end-weight

lvtc

The case study

(i) Syntactic force:

– Quirk et al. (1985: 49-50): 'Complements first'

– Hawkins (2007): 'Arguments precede X'

lvtc

The case study

(ii) Processing force:

– Quirk et al. (1985: 1398): End-weight

– Hawkins' (2004) 'Minimize Domains' or MiD:

> Given two or more categories A, B, [...] related by a grammatical rule R of combination and/or dependency, the human processor prefers to minimize the distance between them within the smallest surface structure domain sufficient for the processing of R. (Hawkins 2000: 234)

and Hawkins (2007) hypothesises that MiD is relevant especially to examples of complementation.

11

The case study

- **So...**

(1) Now I will deal [with the construction] [in a way which will lead to odd results].

(2) Now I will deal [in a way which will lead to odd results] [with the construction].

(1) is claimed to be a better performance solution, on **syntactic** grounds, than (2) because the complement *with the construction* follows the verb (and precedes the adjunct).

(1) is claimed to be a better performance solution, on **processing** grounds, than (2) because of the amount of structure which has to be processed (between the head category and the second constituent in the (local) phrase).

12

The data

- difficulties of exploring syntactic strategies in spoken language:
  - **simpler**: Brown et al. (1984: 17-18): "the spoken language produced by the majority of young people, as indeed by the majority of the population, consists of relatively simple sentence structures – often just sentences and incomplete sentences, strung together"
  - **restricted**: Miller and Weinert (1998: Chapter 3): study of (syntactic) constructions, many of them subordinate => in spontaneous oral language a number of these constructions which appear in written texts were missing from their oral data or appeared in a very low percentage

lvtc

13

The data

Learner corpus:

- **VICOLSE** (Vigo Corpus of Learner Spoken English), produced by Spanish University students of English (Tizón-Couto 2013), 100,000 words

lvtc

14

The data

L1 corpora:

- **LOCNEC** (Louvain Corpus of Native English Conversation, Université Catholique de Louvain –Centre for English Corpus Linguistics–); 162,000 words; as the English native control corpus
- **ADESSE** (University of Vigo, http://adesse.uvigo.es): syntactic database of (native) Spanish; 1.5 million words; as the Spanish native comparable database
- (**PPCMBE** (Penn Parsed Corpus of Modern British English), 1 million words; 1700-1914; written English)

15

The data

- **VICOLSE**:
  - elicitation oral techniques, as regards medium:
    - retelling a story (they were asked to read the first part of *Little Red Riding Hood* and then tell the remaining part of the story in their own words and with their own ideas)
    - describing a picture-based real-world scene (first picture: John Barnet's front room, August 5th at 11am; second picture: the same place at 12 noon on the same day)
    - commenting on a familiar/current topic (current topics such as education, a famous TV show, a film or a book, music, etc.) and giving personal opinion
  - broad trawl, as regards the kind of tasks and variety of the data expected
  - narration and argumentation, as regards genre

16

The data

- **VICOLSE**:
  - participants:
    - 86 undergraduate students of English, University of Vigo
    - consented recording
    - sociological-cultural questionnaire => XML header

lvtc

17

Variable	Value	#	%
Age	19	34	40%
	20-21	32	37.64%
	> 21	19	22.09%
Gender	Female	70	81.39%
	Male	16	18.60%
Mother Tongue	Spanish	46	52.87%
	Galician	7	8.04%
	Both	32	36.78%
	Other	2	2.29%
Speak Galician	Never	7	8.13%
	Sometimes	60	69.76%
	At home/university	15	17.44%
	Always	4	4.65%
Place of residence	Vigo	45	52.32%
	Other	41	47.67%
Years Learning English	< 10	15	17.44%
	10-15	62	72.09%
	> 15	9	7%
Study other foreign languages	French	60	58.82%
	German	19	18.62%
	Other	15	14.70%
	None	8	7.84%
Have visited English-speaking country	Yes	43	50%
	No	43	50%
Have English native friends or relatives	Yes	36	41.86%
	No	50	58.13%
Reason or motivation to study English	I like it	77	81.91%
	Crucial	14	14.89%
	Other (e.g. work)	3	3.19%

lvtc

18

The data

- **VICOLSE**:
  - transcription (c100,000 words)
    - text
    - sound indications (laughing, breathing, vocalic clicks, hesitations, etc.)
    - LINDSEI conventions

lvtc

19

<stdnt 01>

[Task 1]

Little Red Riding Hood *ah* continued ahead (tch) and *ah* through the forest and then he arrived to a little house a wooden house (tch) which was her grandma's one *ahm* once she was there she knocked on the door (tch) and *ahm* her grandmother's voice said *eh* come in . she went in and *mm* she (..) kissed her grandmother and then she said *eh* here you have a gift some things I've bought for you like honey and *ah* cakes and whatever *ah* the grandmother was very happy and *ah* yeah *ah* the problem was that she was in bed cause she was quite ill and *ah* Little Red Riding Hood *ah* (..) started to ask her why was she so so ill in bed (tch) *ahm* (..) then the the grandmother said *aom* I'm very ill because *m* I have a cough (tch) and the the little *child child *started saying *mm* yeah I *?? I think I'm starting to notice some kind of strange features on on your face like for instance you're your eyes they they *really look bigger than they usually *aren't they and then the grandmother said *mm* yes but you know is (tch) they are for for seeing you better and the child and the child continued and he said ok *mm* what about your ears they rel= they are really big and the grandmother said yes because you know I I've bought them in order to hear my beloved niece better (tch) *ah* that's great but and about your hair is (..) h= it is not white as as usually

lvtc

20

The data

- **LOCNEC**:
  - 162,000 words
  - spoken production by 50 British (most undergraduate) students, aged 18-30, at Lancaster University
  - LOCNEC as the native control corpus for LINDSEI
  - LOCNEC as the comparable corpus to VICOLSE

lvtc

21

The data

- comparability in CIA: VICOLSE – LOCNEC
  - same design of tasks and topics
  - practically identical transcription conventions (inherited from the LINDSEI project)

but…

  - as regards participants, the subjects in VICOLSE are (Spanish) non-native university students of English and those in LOCNEC are British native university students => LOCNEC as the comparable database
  - as regards modality, in LOCNEC the tasks are carried out in the form of an interview, while VICOLSE consists of monologic recordings

lvtc

22

Parameter	VICOLSE	LOCNEC	Degree of Comparability
Speech type/elicitation	Voluntary, untimed, unprepared	Voluntary, untimed, unprepared	HIGH
Size (words)	100,663 (76,337 words without story telling)	118,398 (B turns only)	AVERAGE
Age of stdts (Average age)	19-32 (20.59)	18-30 (??)	HIGH
Year of studies	First, second, third and fourth year	Mainly first and second, also third, fourth and postgraduate	AVERAGE
L2 proficiency	Intermediate - Upper intermediate –advanced	Native	AVERAGE
Tasks	Picture description, storytelling, set topic, free discussion	Picture description, set topic, free discussion	AVERAGE
Topics	A favorite book or film, a famous TV show, your education, music, important decisions in your life, etc.	A country you have visited, an experience which has taught you something, a book or play you have liked/disliked	AVERAGE
Genre	Monologue	Guided monologue (Dialogue)	LOW/ AVERAGE
Time of completion	2001-2008	1995-2006 (unreleased)	AVERAGE

23

The data

- **ADESSE**:
  - syntactic and semantic database of Present-day written and spoken Spanish
  - annotation of arguments and adjuncts:
    - syntactic function (subject, object, etc.)
    - syntactic category (noun phrase, adjective phrase, clause, etc.)
    - semantic information (animate, concrete, abstract, etc.)
    - semantic/thematic role (agent, theme, patient, etc.)
  - This paper: spoken Spanish (in Spain): 207,948 words.

24

The analysis of the data

- Examples (from VICOLSE):
  - verb + complement + adjunct:
    - they **have** ehh eh [important cds or important records] [in the shops]
    - I **knew** [a a good eh English teacher] [eh in in my sixth year eh of primary school]
    - they **earn** [a lot of money] [with the taxes that the the the sell selling of tobacco]
  - verb + adjunct + complement:
    - I think that no-smokers **complained** [all the time] [about this this theme]
    - he decided to wai-- **wait** [there] [to for for Little Red *Riding Hood]
    - I just **met** [a couple of weeks ago] [ah one of my ah friends of school]

lvtc

25

The analysis of the data

(i)
Syntactic principle
'Complements-first'

lvtc

26

The analysis of the data

- Distribution of complement-first/last constructions in the spoken corpora:

	compl-first	compl-last		%(compl-first)
Vicolse	263	46		85,11326861
Locnec	342	4		98,84393064
Adesse	306	195		61,07784431

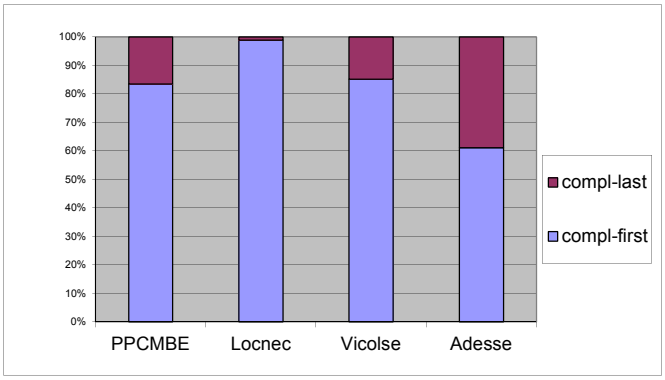
- Comparison with a comprehensive written Modern English corpus (PPCMBE; Pérez-Guerra and Martínez-Insua 2010)

	compl-first	compl-last		%(compl-first)
PPCMBE	13084	2579		83,53444423

lvtc

27

The analysis of the data



•Statistical significance for variation LOCNEC-VICOLSE: yes (P<.0001)
•Statistical significance for variation VICOLSE-ADESSE: yes (P<.0001)
•Statistical significance for variation PPCMBE-LOCNEC: yes (P<.0001)

lvtc

28

The analysis of the data

- Syntactic principle of complements-first:
  - very strong (99%) in native spoken English (LOCNEC)
  - –strong (85%) in non-native (Spanish-L1) spoken English (VICOLSE)
  - actually not strong (61%) in native spoken Spanish (ADESSE)
  - strong (83%) in native written English (PPCMBE) [written modality is not a positive factor]
- So… both English syntax and Spanish syntax exert significant influence on learners' productions (99%>85%, 61%>85%)

lvtc

The analysis of the data

(ii)
Processing principle
'End-weight'

lvtc

30

The analysis of the data

Times the second constituent is longer than the first one in the whole corpora:

Length_first_constituent (whole corpus)				
	min	max	mean	
Vicolse	1	10	1.94174757	
Locnec	1	10	2.10115607	
Adesse	1	8	2.15677966	
Length_second_constituent (whole corpus)				
	min	max	mean	#_2nd>1st
Vicolse	1	19	3.68608414	2.52572816
Locnec	1	19	3.37540453	2.18835673
Adesse	1	18	4.0403397	2.45954146

lvtc

Apparently, the global results are similar.

The analysis of the data

Times the second constituents is longer than the first one in complement-first cxns:

Length_first_constituent (complement-first)				
	min	max	mean	
Vicolse	1	10	1.90874525	
Locnec	1	10	2.10526316	
Adesse	1	8	2.2543554	
Length_second_constituent (complement-first)				
	min	max	mean	#_2nd>1st
Vicolse	1	19	3.53612167	2.51070976
Locnec	1	19	3.39473684	2.19640769
Adesse	1	18	4.1048951	2.32891986

lvtc

The analysis of the data

Times the second constituents is longer than the first one in complement-last cxns:

Length_first_constituent (complement-last)					
	min	max	mean		
Vicolse	1	6	2.13043478		
Locnec	1	3	1.75		few ex
Adesse	1	8	2.00540541		
Length_second_constituent (complement-last)					
	min	max	mean	#_2nd>1st	
Vicolse	1	12	4.54347826	2.6115942	
Locnec	2	3	2.25	1.5	few ex
Adesse	1	15	3.94054054	2.66218147	

lvtc

The analysis of the data

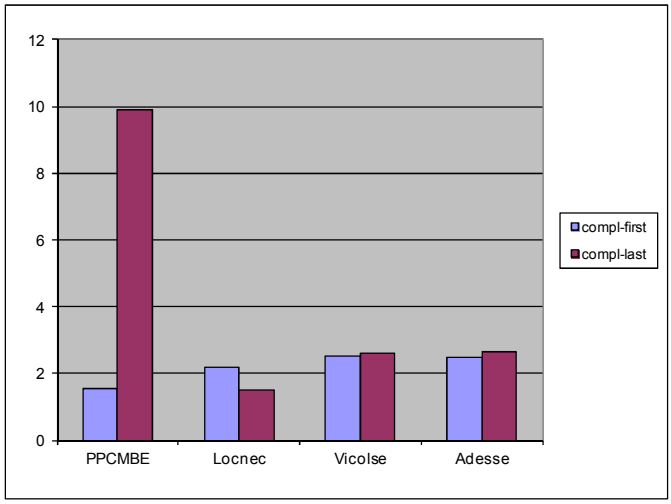
Summing up…

times the second constituent is longer than the first one:

	compl-first	compl-last
PPCMBE	1,55	9,91
Locnec	2,19640769	1,5
Vicolse	2,51070976	2,6115942
Adesse	2,45954146	2,66218147

lvtc

The analysis of the data



lvtc

The analysis of the data

- L1 Spanish (VICOLSE, ADESSE):
  - end-weight explains complement-first/last cxns
  - no differences between complement-first/last cxns
- L1 English:
  - end-weight explains complement-first/last cxns
  - spoken (LOCNEC): no (statistically significant) differences between complement-first/last cxns
  - written (PPCMBE): end-weight is very strong in complement-last cxns => end-weight explains the distribution when the syntactic principle of complements-first is not at work, that is, in complement-last cxns, in native written English

lvtc

The analysis of the data

- So…
  - end-weight is at work in learner and native corpora
  - transfer is not relevant to the results of spoken language
  - in L1 English: written modality is a positive factor in favour of end-weight

lvtc

The concluding remarks

- Syntactic principle (complements-first):
  - Transfer issue:
    - L1 Spanish: 61,07%
    - Learner: 85.11%
    - L1 English: 98,55%
  - Written modality is not a positive factor in L1 English.

lvtc

38

The concluding remarks

- Processing principle (end-weight):
  - At work in learner and native data
  - No transfer differences
  - Written modality is a positive factor in L1 English in complement-last cxns =>
    complements-first > end-weight

lvtc

39

The concluding remarks

- So…

  In native English…

  "the biggest single predictor of relative orderings (…) is syntactic weight" (Hawkins 2000: 232)

  … is **not** strictly correct since, according to the (written) data, syntax (complements-first) seems to be a bigger predictor (end-weight comes into play only when the syntactic principle fails).

  ["in general the light-heavy distribution [end-weight] is no longer a major factor in English word order" (Traugott 1992: 276)]

lvtc

40

For further research

- Analysis of written L1 Spanish:
  - Hypothesis I: There are minor differences between complement-first/last cxns in L1 written Spanish than in L1 written English:
    "The relative 'free' word order in Spanish (...) means that the principle of end-weight may be less noticeable" (Lozano and Medikoetxea 2008: 96)
  - Hypothesis II: elaboration (written versus spoken) determines end-weight compliance both in L1 (i.e., input-based elaboration) and L2 (transfer-based elaboration) codes:
    "correlation between complexity and proficiency: (...) phrasal composition increases in complexity with developmental level" (Lozano and Mendikoetxea 2010: 491)

lvtc

For further research

- Data from a Present-Day English corpus
- Analysis of other types of phrase:
  (3) the author [of this book] [from London] //
  the author [from London] [of this book]
  (4) keen [on music] [to a large extent] //
  keen [to a large extent] [on music]
- Influence of information structure (given-new)

lvtc

42

The references

Brown, Gillian, Anne Anderson, Richard Shillcock and George Yule. 1984. *Teaching talk: strategies for production and assessment*. Cambridge: Cambridge University Press.

Eubank, Lynn, Larry Selinker and Michael Sharwood Smith. 1995. *The current state of Interlanguage*. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle. 2008. "Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation". In Gaëtanelle Gilquin, Szilvia Papp and María Belén Díez-Bedmard eds. *Linking up contrastive and learner corpus research*. Amsterdam: Rodopi, 3–33.

Granger, Sylviane. 1996. "From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora". In Karin Aijmer, Bengt Altenberg and Mats Johansson eds. *Languages in contrast. Text-based cross-linguistic studies*. Lund: Lund University Press, 37–51.

Granger, Sylviane. 2002. "A bird's-eye view of computer learner corpus research". In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson eds. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, 3–33.

Granger, Sylviane. 2009. "The contribution of learner corpora to second language acquisition and foreign language teaching: a critical evaluation". In Karin Aijmer ed. *Corpora and language teaching*. Amsterdam: John Benjamins, 13–33.

The references

Greenbaum, Sidney, Gerald Nelson and Michael Weitzman. 1996. "Complement clauses in English". In Jenny Thomas and Mick Short eds. *Using corpora for language research: Studies in honour of Geoffrey Leech*. London: Longman, 76-91.

Hawkins, John A. 2000. "The relative order of prepositional phrases in English: going beyond manner-place-time". *Language Variation and Change* 11: 231-266.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John A. 2007. "Performance and grammatical varation in the ordering of verb, direct objects and obliques". Delivered at DGfS, Siegen.

Kroch, Anthony, Beatrice Santorini and Ariel Diertani (2010) Penn Parsed Corpus of Modern British English.

Lozano, Cristobal and Amaya Mendikoetxea. 2008. Postverbal subjects at the interfaces in Spanish and Italian learners of L2 English: a corpus analysis. In Gaëtanelle Gilquin, Szilvia Papp and María Belén Díez-Bedmar eds. *Linking up contrastive and learner corpus research*. Amsterdam: Rodopi, 85-125.

Lozano, Cristobal and Amaya Mendikoetxea. 2010. Interface conditions on postverbal subjects: a corpus study of L2 English. *Bilingualism: Language and Cognition* 13/4: 475-497.

Matthews, P.H. 1981. *Syntax*. Cambridge: Cambridge University Press.

The references

Matthews, P.H. 2007. *Syntactic relations. A critical survey*. Cambridge: Cambridge University Press.

Miller, Jim and Regina Weinert. 1998. *Spontaneous spoken language. Syntax and discourse*. Oxford: Oxford University Press.

Nemser, William. 1971. "Approximative systems of foreign language learners". *International Review of Applied Linguistics* 9/2: 115–123.

Nemser, William. 1971. "Approximative systems of foreign language learners". *International Review of Applied Linguistics* 9/2: 115–123.

Pérez-Guerra, Javier and Ana E. Martínez-Insua. 2010. "Diachrony and word order hand in hand: on complementation/adjunction performance solutions in English". Delivered at ICAME 31, Giessen.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Tizón-Couto, Beatriz. 2013, in press. *Clausal complements in native and learner spoken English. A corpus-based study with VICOLSE*. Bern: Peter Lang.

Traugott, Elizabeth Closs. 1992. "Syntax". In Richard M. Hogg eds. *The Cambridge history of the English language. Volume I: The beginnings to 1066*. Cambridge: Cambridge University Press, 168-289.

lvtc

45

A corpus-based study on the distribution of complements and adjuncts in learner language: will we reveal <major findings> <in this study> or will we reveal <in this study> <minor findings>?

Javier Pérez-Guerra (jperez@uvigo.es)
Ana Elina Martínez-Insua (minsua@uvigo.es)

lvtc

Thank you!