Exploring word order in learner corpora: The production of postverbal subjects in L2 English

CARLSeminar
School of Humanities, University of Southampton
2/5/2007
Cristobal Lozano
Universidad de Granada

Amaya Mendikoetxea,
Universidad Autónoma de Madrid/Lancaster University

http://www.uam.es/woslac

AIMS OF THE PRESENTATION

- To present the WOSLAC project: (i) its motivation and objectives, (ii) data collection, (iii) annotation and query software and (iv) data analysis.
- To inform on the results of a preliminary study on the production of inverted subjects in non-native English (Spanish/Italian learners).

FIRST PART

The WOSLAC Project: objectives

To determine the properties that constrain word order in non-native grammars (L2):

- Spanish L1 – English L2 & English L1 – Spanish L2.

  a) **Lexicon-syntax interface**: how the lexical properties of verbs are represented in the syntax (syntactic realization of arguments and adjuncts).
  b) **Syntax-discourse interface**: the relevance of information structure notions such as topic (given/old/retrievable information) and focus (new/non-retrievable information) in word order in L2 grammars.

ENGLISH and SPANISH differ in devices employed for constituent ordering: English ‘fixed’ order is determined by lexico-syntactic properties and Spanish ‘free’ order is determined by information structure, syntax-discourse properties.

DATA COLLECTION (1): WriCLE

WriCLE “Written Corpus of Learner English”

- L1 Spanish - L2 English
- Target: 1 million words
- So far: 400 essays = 600,000 words

  - Learners: 1st and 3rd yr. students of English at the UAM.
  - Essays: around 1,500 words written for the EAP course.

  - Data gathered:
    - a) Essay,
    - b) Learner profile,
    - c) Essay profile and
    - d) Oxford Quick Placement Test
DATA COLLECTION (2): CEDEL2

CEDEL2 “Corpus Escrito del Español L2”
- L1 English - L2 Spanish
- All prof. levels
- Target: 1 million words
- So far: 228,658 words
  - Learners: Words: 204,305, Essays: 820
  - Natives: Words: 24,353, Essays: 43
- Learners: University students of Spanish in USA, UK, Australia & Spain.
- Essays: descriptive and argumentative essays up to 500 words.
- Data gathered: online collection of:
  A) Learning background, B) Spanish Placement Test (Wisconsin) and C) Essay

SOFTWARE: UAM CorpusTool

- UAM CorpusTool (Mick O’Donnell) can be used as a coder and a searcher.
- The tool allows a analyst to select a text from the corpus, and annotate it in various ways. For instance, the analyst can highlight a segment (e.g., an it-cleft) and then assign features to that segment. The tool produces an XML-encoded version of the text file, including the features assigned to the segments.
- Because hand-annotation is slow, the tool will allow the analyst to associate lexico-syntactic patterns with each feature, allowing the tool to automatically detect instances of the pattern.
  - For instance, a pattern like “it be# NP that” would match sentences in the corpus like “It was John that we saw”, and tentatively mark them with the feature it-cleft.
  - The tool would then ask the user to eliminate false matches. This approach eliminates much of the corpus annotation effort.

SOFTWARE: UAM CorpusTool

Project window

Download UAM-Corpus Tool for FREE: http://www.wagsoft.com/CorpusTool/index.html
SOFTWARE: UAM Corpus Tool

Creating annotation schemes:

SOFTWARE: UAM Corpus Tool

Annotating text segments

SOFTWARE: UAM Corpus Tool

Searching the corpus

SOFTWARE: UAM Corpus Tool

Statistics
DATA ANALYSIS: STRUCTURES

Word-order phenomenon
- Left periphery
  - Preposing
  - Left dislocation
- Right periphery
  - Postposing
  - Right dislocation
- Other
  - Passive
  - Inversion
  - There-construction
  - Dative alternation
  - Phrasal verb
  - Cleft
  - Cleft
  - Extraposition

DATA ANALYSIS: FRAMEWORK

- Comparative Framework: to determine the role of L1 in L2 acquisition (transfer) in the areas under study:
  - L1 properties
  - L2 properties
  - Universal Grammar

- We adopt some methodological aspects of CIA: Contrastive Interlanguage Approach (see, e.g., Granger 1996 and Gilquin 2001)
  (a) NNS vs. NS: non-native vs. native data.
  It involves a detailed analysis of linguistic features in native and non-native corpora to uncover and study non-native features in the speech and writing of (advanced) non-native speakers. This includes errors, but it is conceptually wider as it seeks to identify overuse and underuse of certain linguistic features and patterns.

  (b) NNS vs. NNS: different non-native data.
  By comparing learner data from different L1 backgrounds, we can gain a better understanding of interlanguage processes and features, such as those which are the result of transfer or those which are developmental, common to learners with different L1.

- Descriptive and inferential statistics

DATA ANALYSIS: FRAMEWORK

- Formal and functional features interact in the structures under consideration.

- Formal and functional approaches are therefore essential for the understanding of SLA data.

- At the same time, data from non-native grammars is potentially significant for the understanding of linguistic phenomena in native grammars

CONTRIBUTIONS TO THE FIELD

- Linguistic Theory: better understanding of interfaces (lexicon-syntax, syntax-discourse and syntax-phonology).

- L2 acquisition: better understanding of transfer and non-transfer phenomena.

- Corpus studies: use of corpora for the study of formal features. Creation of the first Spanish learner corpus.

- Pedagogy: better understanding of word order errors
SECOND PART

Postverbal subjects in learner English


- Postverbal subjects
- L1 Spanish/L1 Italian – L2 English
- ICLE (International Corpus of Learner English)
- Interfaces:
  - lexicon-syntax
  - syntax-discourse
  - syntax-phonology

What are the conditions under which learners produce inverted subjects, regardless of problems to do with syntactic encoding?

Word Order in L1 English (1)

- Fixed SV(O) order- Restricted use of postverbal subjects:
  - a) XP V S
    - (i) XP is an adverbial element, typically expressing time or place and linking the sentence to the prior discourse
    - (ii) V is an intransitive verb, typically expressing existence or appearance on the scene (= unaccusative)
    - (iii) S is often syntactically/phonologically ‘heavy’ consisting of a noun and a variety of pre and/or postmodifiers, which introduce new information in the discourse.

1. Michael puts loose papers like class outlines in the large file-size pocket. He keeps his checkbook handy in one of the three compact pockets. The six pen and pencil pockets are always full and in the outside pocket goes his schedule book, chap stick, gum, contact lens solution and hair brush. [Land’s End March 1989 catalog, p. 95] (Birner 1994: 254)

Word Order in L1 English (2)

b) There-constructions

(2) a. Somewhere deep inside [there] arose a desperate hope that he would embrace her [FICT]

b. In all such relations [there] exists a set of mutual obligations in the instrumental and economic fields [ACAD]

c. [There] came a roar of pure delight as…. [FICT] [Biber et al. 1999: 945]

Word order in L1 English

(summary)

- Lexicon-syntax interface (Levin & Rappaport-Hovav, etc):
  - Unaccusative Hypothesis (Burzio 1986, etc)
    - *There sang four girls at the opera. [unergative verb]
    - There arrived four girls at the station. [unaccusative verb]

- Syntax-discourse interface (Biber et al, Birner 1994, etc):
  - Postverbal material tends to be focus (new info)
    - We have complimentary soft drinks and coffee. Also complimentary is red and white wine.

- Syntax-Phonological Form (PF) interface (Arnold et al 2000, etc)
  - Heavy material is sentence-final (Principle of End-Weight, Quirk et al. 1972):
    - That money is important is obvious.
    - It is obvious that money is important.

Subjects which are focus, long and complex tend to occur postverbally in those structures which allow them.
Word Order L1 Spanish/Italian (1)

- Postverbal subjects are produced ‘freely’ with all verb classes (as part of the cluster or properties associated with the Null Subject Parameter):

(3) a. Ha telefoneado María al presidente. (transitive).
   Has phoned Mary the president

b. Ha hablado Juan. (unergative)
   has spoken Juan

c. Ha llegado Juan. (unaccusative)
   has arrived Juan

Word Order L1 Spanish/Italian (2)

- Inversion as ‘focalisation’:
  - preverbal subjects are topics (given information)
  - and postverbal subjects are focus (new information) (Belletti 2001, 2004, Zubizarreta 1998)

(4) ¿Quién ha llegado/hablado?
Who has arrived/spoken?
   i. Ha llegado/hablado Juan
   ii. #Juan ha llegado/hablado

- The occurrence of postverbal subjects in Spanish is determined by:
  - syntax-discourse properties (they are focus)
  - and syntax-phonology properties (heavy subjects show a tendency to be postposed – a universal language processing mechanism: placing complex elements at the end reduces the processing burden; Hawkins 1994)

Previous L2 findings


- L1 Spanish/Italian/Arabic – L2 English:
  6. ...it arrived the day of his departure ...
  7. And then at last comes the great day.
  8. In every country exist criminals
  9. ...after a few minutes arrive the girlfriend with his family too.

- Only with unaccusative verbs (never with unergatives).
  - Unaccusatives: arrive, happen, exist, come, appear, live...
  - Unergatives: cry, speak, sing, walk ...

- Explanation: syntax-lexicon interface (Unaccusative Hypothesis)

Hypotheses

GENERAL HYPOTHESIS:
- Conditions licensing VS in L2 Eng are the same as those in Native Eng, DESPITE differences in syntactic encoding.

SPECIFIC HYPOTHESIS:

- H1 [LEXICON]: Lexicon-syntax interface:
  - Postverbal subjects with unaccusatives (never with unergatives)

- H2 [WEIGHT]: Syntax-PF interface:
  - Postverbal subjects: heavy (but preverbal light)

- H3 [FOCUS]: Syntax-Discourse interface:
  - Postverbal subjects: focus (but preverbal topic)
Method

- Learner corpus: L1 Spa – L2 Eng; L1 Ital – L2 Eng
  - ICLE (Granger et al. 2002)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of essays</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICLE Spanish</td>
<td>251</td>
<td>200,376</td>
</tr>
<tr>
<td>ICLE Italian</td>
<td>392</td>
<td>227,085</td>
</tr>
<tr>
<td>TOTAL</td>
<td>643</td>
<td>427,461</td>
</tr>
</tbody>
</table>

(Data: proficiency level?)

- WordSmith v. 4.0 (Scott 2004)
  - Concordance queries can be performed automatically with WordSmith, by targeting specific verbs but there is a lot of manual work (filtering out unusable data, coding data in Excel, analysing data in SPSS, etc).

- SPSS v. 12.0

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>V type</th>
<th># usable concordances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>Unergative</td>
<td>153</td>
</tr>
<tr>
<td>Italian</td>
<td>Unaccusative</td>
<td>640</td>
</tr>
<tr>
<td>TOTAL</td>
<td>Unaccusative</td>
<td>794</td>
</tr>
</tbody>
</table>

Data analysis (cont’d)

- MAIN FILTERING CRITERIA:
  - The verb must be intransitive (unergative or unaccusative).
  - The verb must be finite, active voice.
  - The subject can appear either postverbally (VS) or preverbally (SV).
  - The subject must be an NP.
  - The sentence can be either grammatical or ungrammatical in native English.

- OTHER FILTERING CRITERIA (TOTAL=28)

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>V type</th>
<th># usable concordances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>Unergative</td>
<td>153</td>
</tr>
<tr>
<td>Italian</td>
<td>Unaccusative</td>
<td>640</td>
</tr>
<tr>
<td>TOTAL</td>
<td>Unaccusative</td>
<td>794</td>
</tr>
</tbody>
</table>
Data coding/analysis: EXCEL

Data analysis: SPSS (descriptive & inferential statistics)

H1: syntax-lexicon

Example: syntax-lexicon (unaccusatives only)

- UNACC: Mostly S-V: [92% Italian, 97% Spanish]
  - The real problem appears when they have to look for their first job.
  - These people should exist.

- UNACC: But many V-S: [8% Italian, 3% Spanish]
  - There exist positive means of earning money.
  - So arises the Saint Inquisition.
  - In the main plot appear the main characters: Volpone and Mosca.
  - …it has appeared some cases of women that have killed their husbands.
  - …and from this moment begins the avarice.
  - …and appeared a lot more theatres.
Result: VS and specific unaccusative verbs

Results: Unaccusative: grammaticality of VS

- Locative inversion:
  - In the main plot appear the main characters: Volpone and Mosca.
- There-insertion:
  - There exist positive means of earning money.
- AdvP-insertion:
  - ... and here emerges the problem.

- * It-insertion:
  - *In the name of religion it had occurred many important events…
- *XP-insertion:
  - *In 1760 occurs the restoration of Charles II in England.
- *Ø-insertion:
  - *...because exist the science technology and the industrialisation.

Results: Unaccusative: grammatical vs. ungrammatical VS

Figure 1. Proportion (in %) of grammatical vs. ungrammatical unaccusative VS

<table>
<thead>
<tr>
<th>Group</th>
<th>VS Spanish ICLE</th>
<th>VS Italian ICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical</td>
<td>53.3%</td>
<td>46.7%</td>
</tr>
<tr>
<td>Ungrammatic</td>
<td>46.7%</td>
<td>53.3%</td>
</tr>
</tbody>
</table>

Result: Unaccusative: Type of VS structures
**Examples: syntax-phonology**

**SV:** typically LIGHT

...but they may appear everywhere.
...since the day eventually came...
...these people should exist, ...

**VS:** typically HEAVY

Against this society drama emerged an opposition headed by Oscar Wilde and Bernard Shaw.

...so came the decline of the theatre.

Then came the necessity to earn more.

---

**Result 3: syntax-discourse**

**EXAMPLES: syntax-discourse**

**VS:** FOCUS

...there also exists a wide variety of optional channels which have to be paid.

So arised the Saint Inquisition.

In 1880 it begun the experiments whose result was the appearance of the television some years later.

**SV:** typically TOPIC

I use the Internet ... I find windows ... if they press on any of these windows ... these windows cannot appear because a child could enter easily...

...the world of drugs: mafias ... problems with mafias finished ... dangerous people making money ... no reason why these people should exist.

---

Discourse status (topic/focus) has to be measured manually by establishing theoretical criteria and then by checking the context (or even the essay) manually.
Thank you!

ADDITIONAL SLIDES TO FOLLOW:

Unaccusativity Hypothesis

VS in native Spanish
VS en español (cont'd)

**Foco presentacional (interfaz sintaxis-discurso)**

**Inerativos: VS**
A: ¿Quién gritó anoche en la reunión?  
B: # Un hombre gritó.  
Gritó un hombre.

**Inacusativos: VS**
A: ¿Quién llegó anoche a la reunión?  
B: # Un hombre llegó.  
Llegó un hombre.

---

**Datos de análisis (cont'd)**

- **CONCORDANCIAS: 6 CRITERIOS DE FILTRO BÁSICOS:**
  1. El verbo debe ser **intransitivo** (ingenerativo o no transaccivo).
     - V: En el programa de televisión, un hombre llegó a la reunión. [unac]
     - V: Leontes habla de cómo ha creado varias opiniones. [unerg]
     - X: Este gobierno ha modificado varias opiniones. [trans]

  2. El verbo debe ser **finito, con(out) auxiliar**.
     - X: Also, also exists the psychological aggression... [finite no aux]
     - V: The cases of men mistreated do not appear in the media. [finite aux]
     - X: This contradiction could disappear [finite modal]
     - X: There's no reason for it to exist. [for clause + to infinitive]
     - X: Poor people cross borders to escape from poverty. [to-inf clause]
     - V: Let's pass... ['let' constructions]
     - V: ...make everyone's life go ahead [causativo + infinitivo]
     - V: ...they go away in order to escape to France. ['in order to' clauses]
     - X: ...women have to live with the aggressor [have to/ought to/able to]
     - V: ...prudence was beginning to disappear. [aspetual/periphrases]
     - X: Before entering the argumentation... [small clauses]
     - X: ...instead of following... [complement of P]
     - X: ...likely to happen... [complement of A]
     - X: The tests to enter the army are quite difficult now. [complement of N]

---

**Foco presentacional (interfaz sintaxis-discurso)**

**Inerativos: VS**
A: ¿Quién gritó anoche en la reunión?  
B: # Un hombre gritó.  
Gritó un hombre.

**Inacusativos: VS**
A: ¿Quién llegó anoche a la reunión?  
B: # Un hombre llegó.  
Llegó un hombre.

---

**Fotos de análisis (cont'd)**

- **3. El verbo debe ser en la voz activa.**
  - X: This contradiction could disappear. [active unaccusative]
  - X: This situation has already been happened. [passivised unaccusative]

- **4. El sujeto debe ser una NP.**
  - X: ...it arose [diverse social ranks, the rich and the poor that depended on the property they had]. [inverted NP subject]
  - X: ...it only remains [to add that nowadays we live in a world...]. [extraposition]
  - X: It happened [that the countries which make the weapons are...]. [extraposition]

- **5. El sujeto puede aparecer postverbalmente (VS) o preverbalmente (SV).**
  - X: The real problem appears when they have to look for their first job. [SV]
  - X: So arose the Saint Inquisition. [VS]

---

**Fotos de análisis (cont'd)**

- **OTROS CRITERIOS DE FILTRO:**
  - **Target V + V (coordination):**
    - X: Families without father exist and work well.
  - **Coordinator + target V:**
    - X: ...we can manage to obtain it and live in a better world.
  - **Interrogatives (only if V is the target):**
    - X: How could they live?
    - X: Does exist then a manipulation of television?
  - **Formulaic & Set expressions in English:**
    - X: As sometimes happens...
    - X: ...fall victim to...
    - X: ...the world we live in.
  - **Set expressions transferred from L1:**
    - X: ...it happens the same.
    - X: ...they fall into account that they have treated very badly Mr Hardcastle.
  - **Phrasal verbs:**
    - X: ...a scientist come up with an intention...
  - **Quotes (literary or other):**
    - X: "To what purpose, April, do you return again?"
    - X: "Feminism has to evolved or die", Friedian said in 1982...
WordSmith: query searches:

- For every lemma (e.g., APPEAR, ARISE), we searched for:
  - All possible native forms:
    - appear, appears, appearing, appeared
    - arise, arises, arising, arose, arisen
  - All possible overregularised and overgeneralised learner forms:
    - arised, arosed, arisened, arosened ("So arised the Sain Inquisition")
  - All possible forms with probable L1 transfer of spelling:
    - apear, apearis, apearing, apeared
  - All other possible misspelled forms:
    - appeard, apeard

Extraposition was discarded

- NOTE: extraposition discarded:
  - It only remains [to add that nowadays we live in a world...]
  - It happened [that the countries which make the weapons are...]

Result: VS and (in)definiteness

Figure 1: Production of postverbal subjects according to their definiteness.

INDEFINITE
...some decades ago, it appeared a new invent: the television.
The play was very well performed and also appeared new elements in the stage.
...it has appeared some cases of women that have killed their husbands...

DEFINITE
...because later could appear the real evidence and the real guilty.
...and usually appears the noble young man that either waste or has wasted his fortune.
In the main plot appear the main characters: Volpone and Mosca.

Scale (syntactic weight/complexity)

<table>
<thead>
<tr>
<th>NOMINAL SCALE</th>
<th>ORDINAL SCALE</th>
<th>SYNTACTIC STRUCTURE</th>
</tr>
</thead>
</table>
| LIGHT
 | 0              | (D) ADJ* N         |
|               | (D) ADJ N      |
| HEAVY
 | 1              | (D) ADJ* N         |
|               | (D) ADJ* N* PP* |
|               | (D) ADJ N* IP/CP |

Scale (syntactic weight/complexity)
Topic vs. Focus: Retrievability scale

Discourse old (TOPIC) Discourse new (FOCUS)

1. Evoked (text + fam)
   - Evoked (aud)
   - Evoked (situ)
   - Inferrables (all)
   h-old  d-old

2. Evoked (text - fam)
   h-old  d-old

3. New-anch
   h-new  d-new

4. Brand new
   Information focus
   Present focus intro
   h-new  d-new

5. Unused
   h-old  d-new