Learner corpora and the acquisition of word order: A study of the production of Verb-Subject structures in L2 English

CORPUS LINGUISTICS 2007
Birmingham
Cristobal Lozano
Universidad de Granada
Amaya Mendikoetxea
Universidad Autónoma de Madrid/Lancaster University

http://www.uam.es/woslac

AIM OF THE PRESENTATION

- To inform on the results of a study on the production of postverbal subjects (VS order) in non-native English.
- Purpose of the study: to characterize the interlanguage of advanced non-native speakers of English L2 (Spanish/Italian L1) by examining their production of both grammatical and ungrammatical VS structures, as represented in the relevant ICLE subcorpora (Granger et al. 2002):
  - (1) ...and here emerges the problem.
  - (2) "In the name of religion it had occurred many important events

FRAMEWORK

- Main question: What are the conditions under which learners produce inverted subjects (VS structures), regardless of problems to do with syntactic encoding (grammaticality)?
- Comparative Framework: to determine the role of L1 in L2 acquisition (transfer) in the areas under study
  Learner corpora vs. native corpora (LOCNESS)

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

Two types of Intransitive Vs: The Unaccusative Hypothesis

- Unaccusative V: S is a notional object (patient or theme) - an entity that comes into existence (3), appears on the scene (4) or undergoes a change of state/location (5)
  - (3) Problems exist
  - (4) Three girls arrived
  - (5) The window broke

- Unergative V: S is a notional subject; it is an agent or has protagonist control over the action:
  - (6) John spoke/cried/laughed/worked...
  John is a subject both notionally and syntactically
**Word Order in L1 English (1)**

Fixed SV(O) order - Restricted use of postverbal subjects:

a) XP V S (Inversion structures with an opening adverbial)

(7) 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 go <in the outside pocket> go <his schedule book, chap stick, gum, contact lens solution and half brushes>, [Land's End March 1989 catalog, p. 95] (Birner 1994: 254)

(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.

**Word Order in L1 English (2)**

b) *There*-constructions

(8) 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]

➢ Roughly the same (sub)class of Vs and same conditions as in XPVS structures.

---

**Word order in L1 English (VS order)**

- **Lexicon-syntax interface** (Levin & Rappaport-Hovav, etc):
  - Unaccusative Hypothesis (Burzio 1986, etc)
  - “There sang four girls at the opera. [unergative verb]” (10) “There arrived four girls at the station. [unaccusative verb]

- **Syntax-discourse interface** (Biber et al, Birner 1994, etc):
  - Postverbal subjects tend to be focus/relatively unfamiliar information
  - (11) We have complimentary soft drinks and coffee. Also complimentary 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) – general processing mechanism (reducing processing burden)
  - (12) One Sunday morning the warm sun came up and - pop! out of the egg came a tiny and very hungry caterpillar. [Eric Carle, The very hungry caterpillar, London: Penguin]

---

**Word Order L1 Spanish/Italian (1)**

Postverbal subjects are produced ‘freely’ with all verb classes:

(13) 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

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

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

(15) Chi è arrivato/parlato?
   I. È arrivato/ A parlato Gianni
   ii. # Gianni è arrivato/a parlato
Word Order L1 Spanish/Italian (VS order)

- **Lexicon-syntax interface**
  No restrictions: postverbal subjects occur with all verb classes

- **Syntax-discourse interface**
  Postverbal subjects in Spanish and Italian are focus

- **Syntax-Phonological Form (PF) interface**
  Heavy subjects show a tendency to be postposed – a universal language processing mechanism: placing complex elements at the end reduces the processing burden (J. Hawkins 1994).

Subjects which are focus, long and complex tend to occur postverbally, with no restrictions at the lexicon-syntax interface.

---

The phenomenon in SLA


Only with **unaccusative** verbs (never with unergatives).

- Unaccusatives: arrive, happen, exist, come, appear, live…
- Unergatives: cry, speak, sing, walk …

- L1 Spanish/Italian/Arabic – L2 English:
  (16) …it arrived the day of his departure…
  (17) And then at last comes the great day.
  (18) In every country exist criminals
  (19) …after a few minutes arrive the girlfriend with his family too.

  > **Explanation**: syntax-lexicon interface (**Unaccusative Hypothesis**)

---

The psychological reality of the Unaccusative Hypothesis

- A number of studies have found that L2 learners are aware of the argument structure distinction between unaccusative and unergative Vs and that they use this as a guiding principle to construct L2 mental grammars.

- However, learners have difficulty in determining the range of appropriate syntactic realizations of the distinction, and this can persist into near-native levels of proficiency (see R. Hawkins 2001: 5.4).

CRUCIAL DIFFERENCE: these previous studies focused on ERRORS, thus emphasising the **differences** between native and non-native structures. By contrast, our study emphasises the **similarities** between native and non-native structures.

---

Hypotheses

**GENERAL HYPOTHESIS:**

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

**SPECIFIC HYPOTHESES:**

- **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)

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, ariseden, arosened (“So arised the Sain Inquisition”)
  - All possible forms with probable L1 transfer of spelling:
    - apear, apear, apear, apearing, apeared
  - All other possible misspelled forms:
    - appeard, apear

Data analysis

- Based on Levin (1993) and Levin & Rappaport-Hovav (1995):
  - Unergatives: cough, cry, shout, speak, walk, dance…
  - Unaccusatives: exist, live, appear, emerge, happen, arrive…

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

- DATA ANALYSIS
  - Corpus Number of essays Number of words
  - ICLE Spanish 251 200,376
  - ICLE Italian 392 227,085
  - TOTAL 643 427,461

- 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, ariseden, arosened (“So arised the Sain Inquisition”)
    - All possible forms with probable L1 transfer of spelling:
      - apear, apear, apear, apearing, apeared
    - All other possible misspelled forms:
      - appeard, apear

CONCORDANCES: RAW OUTPUT

- Thousands of concordances, but approx. ¾ were unusable.
- Filtering criteria had to be applied manually.

- SPSS v. 12.0

- DATA ANALYSIS
  - Corpus Number of essays Number of words
  - ICLE Spanish 251 200,376
  - ICLE Italian 392 227,085
  - TOTAL 643 427,461

- 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, ariseden, arosened (“So arised the Sain Inquisition”)
    - All possible forms with probable L1 transfer of spelling:
      - apear, apear, apear, apearing, apeared
    - All other possible misspelled forms:
      - appeard, apear

CONCORDANCES: RAW OUTPUT

- Thousands of concordances, but approx. ¾ were unusable.
- Filtering criteria had to be applied manually.
Data coding/analysis: EXCEL

H1: syntax-lexicon

Examples: syntax-lexicon (unaccusatives only)

- UNACC: Mostly S-V: [roughly 92% Spanish, 97% Italian]
  (20) a. The real problem appears when they have to look for their first job.
      b. ...these people should exist.

- UNACC: But many V-S: [roughly 8% Spanish, 3% Italian]
  (21) a. There exist positive means of earning money.
      b. Then came psychoanalysis.
      c. In the main plot appear the main characters: Volpone and Mosca.
      d. So arised the Saint Inquisition.

(22) a. "...it has appeared some cases of women that have killed their husbands.
      b. "...and from this moment begins the avarice.
      c. "...and appeared a lot more theatres.

Results: Unaccusative: grammatical vs. ungrammatical VS

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

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>V type</th>
<th># postverbal S</th>
<th># usable concordances</th>
<th>Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>Unergative</td>
<td>0</td>
<td>153</td>
<td>0/153 (0%)</td>
</tr>
<tr>
<td></td>
<td>Unaccusative</td>
<td>52</td>
<td>640</td>
<td>52/640 (8.1%)</td>
</tr>
<tr>
<td>Italian</td>
<td>Unergative</td>
<td>0</td>
<td>143</td>
<td>0/143 (0%)</td>
</tr>
<tr>
<td></td>
<td>Unaccusative</td>
<td>15</td>
<td>574</td>
<td>15/574 (2.6%)</td>
</tr>
</tbody>
</table>

Subcorpus: Spanish ICLE | Group: VS Spanish ICLE
Subcorpus: Italian ICLE | Group: VS Italian ICLE


**Examples: syntax-phonology**

**SV:** typically LIGHT

(23) a. …but they may appear everywhere.

b. …since the day eventually came…

c. …these people should exist, …

**VS:** typically HEAVY

(24) a. Against this society drama emerged an opposition headed by Oscar Wilde and Bernard Shaw.

b. …exists yet in Spain a group of people who are supposed to be professional soldiers.

c. It is almost disappearing the use of writing nice letters to friends.

---

**Result 3: syntax-discourse**

**EXAMPLES: syntax-discourse**

**VS:** FOCUS

(25) a. …there also exists a wide variety of optional channels which have to be paid.

b. So arised the Saint Inquisition.

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

**SV:** typically TOPIC

(26) a. 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…

b. …the world of drugs: mafias … problems with mafias finished … dangerous people making money … no reason why these people should exist.
Our results and CIA (1)

- **Contrastive Interlanguage Analysis** (Granger 1996, Gilquin 2001)

(a) 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.

- No significant differences between Italian and Spanish learners, as expected, except for frequency of inversion [8.1 (Sp) vs. 2.6 (It)].
  - Possible explanation: lexical bias.

Our results and CIA (2)

(b) 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 (Granger 2002: 12-13).

<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>UAM</td>
<td>85</td>
<td>63,836</td>
</tr>
<tr>
<td>TOTAL</td>
<td>336</td>
<td>264,212</td>
</tr>
<tr>
<td>vs. LOCNESS</td>
<td>436</td>
<td>324,304</td>
</tr>
</tbody>
</table>
Syntax-Lexicon

NS vs. NNS

Examples of VS in LOCNESS

- AdvP insertion  (3 tokens)
  (27) Thus began the campaign to educate the public on how one contracts AIDS.

- PP insertion   (7 tokens)
  (28) …along with this development has come opposition from both the medical field and a proportion of the general public, who…

- There-insertion (6 tokens)
  (29) Certainly there exists a demand for this work to be done.

Syntax-Phonology: NS vs. NNS

Heavy-Light subjects with Unacc Vs

- Subjects in SV structures with unacc Vs tend to be ‘light’ in NS (67.7% in ICLE+WriCLE and 68.1% in LOCNESS)
- Subjects in VS structures with unacc Vs are overwhelmingly ‘heavy’ in NS (81.0% in ICLE+WriCLE and 81.3% in LOCNESS)

Syntax-Discourse: NS vs. NNS

Focus and Topic subjects with Unacc Vs

- Most subjects in SV structures with unacc Vs in NS are topic (89.9% in ICLE+WriCLE and 83.5% in LOCNESS); just a few are focus (10.5% in ICLE+WriCLE and 16.5% in LOCNESS; p=0.223).
- Subjects in VS structures with unacc Vs are overwhelmingly focus in NS (98.3% in ICLE+WriCLE 100% in LOCNESS; p=0.784).
Some preliminary conclusions

- These results confirm that Spanish (and, presumably, Italian) learners of English produce postverbal subjects under exactly the same interface conditions as in L1 English (unaccusativity being a necessary but not a sufficient condition).

- Spanish and Italian learners show persistent problems in the syntactic encoding of the construction, producing mostly ungrammatical examples (it-insertion, 0-insertion, wrong XP).

- Spanish learners overuse the construction and show a lexical bias for the V exist.

Thank you!
VS in native Spanish

**Inergativos:** SV

A: ¿Qué pasó anoche en la reunión?
B: Un hombre gritó.

* Gritó un hombre.

**Inacusativos:** VS

A: ¿Qué pasó anoche en la reunión?
B: # Un hombre llegó.

Llegó un hombre.

**VS in native Spanish (2)**

**Inergativos:** VS

A: ¿Quién gritó anoche en la reunión?
B: Gritó un hombre.

**Inacusativos:** VS

A: ¿Quién llegó anoche a la reunión?
B: # Un hombre llegó.

**Data analysis (cont’d)-------**

- CONCORDANCES: 6 BASIC FILTERING CRITERIA:
  1. The verb must be intransitive (unergative or unaccusative).
     - ✓ In the screen of the television one or two “rombos” should appear. [unac]
     - ✓ Leontes cries and the statue talks. [unerg]
     - ✓ This government’s movement has created several opinions. [trans]
  2. The verb must be finite, with(out) aux.
     - ✓ …also it exists the psychological agressions… [finite no aux]
     - ✓ …the cases of men mistreated do not appear in the media. [finite aux]
     - ✓ This contradiction could disappear [finite modal]
     - ✓ There’s no reason for it to exist. [for clause + to inf]
     - ✓ Poor people cross borders to escape from poverty. [to-inf clause]
     - ✓ …let time pass… ['let' constructions]
     - ✓ …make everyone’s life go ahead [causative + infinitive]
     - ✓ Returning to the title of this paper,… [gerundive clauses]
     - ✓ …they go away in order to escape to France. ['in order to' clauses]
     - ✓ …women have to live with the aggressor [have to/ought to/able to]
     - ✓ …prudence was beginning to disappear. [verbal/aspectual periphrases]
     - ✓ Before entering the argumentation,… [small clauses]
     - ✓ …instead of following… [complement of P]
     - ✓ …likely to happen… [complement of A]
     - ✓ The tests to enter the army are quite difficult now. [complement of N]
  3. The verb must be in the active voice.
     - ✓ This contradiction could disappear. [active unaccusative]
     - ✓ This situation has already been happened. [passivised unaccusative]
  4. The subject must be an NP.
     - ✓ …it arose [diverse social ranks, the rich and the poor that depended on the property they had]. [inverted NP subject]
     - ✓ …it only remains [to add that nowadays we live in a world…] [extraposition]
     - ✓ It happened [that the countries which make the weapons are…] [extraposition]
  5. The sentence can be either grammatical or ungrammatical in native English.
     - ✓ This contradiction could disappear. [gram]
     - ✓ …it won’t exist nothing of what people don’t get bored or tired. [ungram]
  6. The subject can appear either postverbally (VS) or preverbally (SV).
     - ✓ …the real problem appears when they have to look for their first job. [SV]
     - ✓ So arose the Saint Inquisition. [VS]
Data analysis (cont’d) ---------

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

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...]

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, aapers, apering, apeared
  - All other possible misspelled forms:
    - appeard, apeard

Result: Unaccusative: Type of VS structures

- [Graph showing production rates for different types of VS structures in Spanish and Italian ICLE groups.]
Result: VS and (in)definiteness

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

- **Definite**: 41.4%
- **Indefinite**: 58.6%

**INDEFINITE**
...some decades ago, it appeared a new invention: 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>
<tbody>
<tr>
<td>LIGHT</td>
<td>0</td>
<td>(D) PRN</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>(D) ADJ N</td>
</tr>
<tr>
<td>HEAVY</td>
<td>2</td>
<td>(D) ADJ* N</td>
</tr>
<tr>
<td></td>
<td>(D) (ADJ) N*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(D) (ADJ) N*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(D) ADJ N* PP*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(D) ADJ N* PP*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(D) ADJ N* (PP*)</td>
<td></td>
</tr>
</tbody>
</table>

**Topic vs. Focus: Retrievability scale**

<table>
<thead>
<tr>
<th>Discourse old (TOPIC)</th>
<th>Discourse new (FOCUS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrievable + recent</td>
<td>retrievable + recent</td>
</tr>
<tr>
<td>recent</td>
<td>present</td>
</tr>
<tr>
<td>recent</td>
<td>recent</td>
</tr>
</tbody>
</table>

1. Evoked (text - old) Evoked (allo) Informative (ali) h-old d-old
2. Evoked (text - old) h-old d-old
3. New-such h-new d-new
4. Brand new Information focus Present focus intro h-new d-new
5. Unused h-old d-new