It is more complex to read letters than news (really?): on linguistic complexity and text-type variation in the recent history of English

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

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

PLM2007, Gniezno, September 2007

Linguistic complexity

Initial hypothesis

- intra-linguistic hypothesis (central assumption of this paper): Aspects within a language (genres or text types, historical stages, etc.) can be graded according to linguistic complexity.

Theoretical assumptions

(i) text types encode linguistic features and differ in complexity (Taavitsainen 2001:141).

(ii) complexity is influenced by linguistic 'circumstances' and is not inherent to the clauses (Crain & Shankweiler's 1988 Processing Deficit Hypothesis)

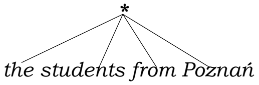
(iii) complexity as a relational (*than*-) notion

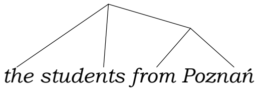
(iv) complexity as a relative notion: Frazier (1988:204): "there is no general unit of complexity (...) which would predict in 'absolute' terms the complexity of a sentence"

=> several metrics

Theoretical assumptions

(v) need for connectivity in syntax:





(Hawkins 2006:208, 'Minimize Domains' and 'Combination', 2006:211 'Phrasal Combination Domain)

Assumptions and 'hallmarks'

(vi) importance of the subject (external argument) as far as the determination of complexity is concerned.

· Davison & Lutz (1985:60): "the high load of processing would occur in subject position"

· Gibson (1998:27): "modifying the subject should cause an increase in the memory cost for predicting the matrix verb"

Goal and methodology

- **Working hypothesis**: arguments (external *ie* subjects, and internal *ie* objects) behave differently from adverbials.
- **Methodology**:
  - exploration of linguistic complexity in two text types in the recent history of English
  - analysis of the unmarked (preverbal) subjects (external arguments), unmarked (postverbal) objects (internal arguments) and adverbials (non-subcategorised components) of declarative sentences in a corpus 1750-1990

The corpus

corpus: *ARCHER* (British component)
periods: 1750-1799, 1850-1899, 1950-1990
text types:

- news: formal, written, public
- letters: more informal, written~speech-based, public~private

text type \ period	1750-1799	1850-1899	1950-1990	Total
news	26,138	23,213	24,235	73,586
letters	12,006	10,800	11,694	34,500
Total	38,144	34,013	35,929	108,086

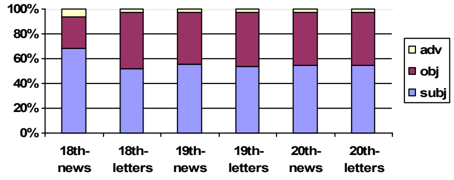
Table 1: The corpus (word totals)

The corpus

text type \ period		1750-1799	1850-1899	1950-1990	Total
news	subjects	1,474 nf=56.39	1,497 nf=64.48	1,676 nf=69.15	4,647
	objects	558 nf=21.34	1,134 nf=48.85	1,290 nf=53.22	2,982
	adverbials	132 nf=5.05	68 nf=2.92	89 nf=3.67	289
letters	subjects	1,040 nf=86.62	808 nf=74.81	930 nf=79.52	2,778
	objects	913 nf=76.04	671 nf=62.12	728 nf=62.25	2,312
	adverbials	53 nf=4.41	35 nf=3.24	39 nf=3.33	127
Total		4,170	4,213	4,752	13,135

Table 2: Distribution of subjects, objects and adverbials (nf = normalised frequency per 1,000 words)

The corpus



Graphic 1: Distribution of subjects, objects and adverbials

No significant distributional differences in the periods under investigation.

The corpus

Subjects

text type \ period	1750-1799		1850-1899		1950-1990		Totals
	non-pron	pron	non-pron	pron	non-pron	pron	
news	895	579	1,028	469	1,142	534	4,647
	60.71%	39.28%	68.67%	31.32%	68.13%	31.86%	
letters	262	778	208	600	218	712	2,778
	25.19%	74.8%	25.74%	74.25%	23.44%	76.55%	
Totals	1,157	1,357	1,236	1,069	1,360	1,246	7,425

Table 3: Pronominal and non-pronominal subjects (percentages per text type and period)

The corpus

Subjects

- No significant diachronic change.
- Differences in the ratios of pronominal subjects: whereas in the news approx. 65% of the subjects are non-pronominal, in the letters the percentage is the opposite (approx. 75% of the subjects are pronominal), which accords with the subjective style of the latter text type. (The text type of letters is stylistically marked and includes many personal pronouns fulfilling argument functions -- subject and object.)

The corpus

Objects

text type \ period	1750-1799		1850-1899		1950-1990		Totals
	non-pron	pron	non-pron	pron	non-pron	pron	
news	448	110	1,014	120	1,213	77	2,982
	80.28%	19.71%	89.41%	10.58%	94.03%	5.96%	
letters	646	267	463	208	562	166	2,312
	70.75%	29.24%	69%	30.99%	77.19%	22.8%	
Totals	1,094	377	1,477	328	1,775	243	5,294

Table 4: Pronominal and non-pronominal objects (percentages per text type and period)

The corpus
Objects

- Pronominal objects display a much lower percentage than pronominal subjects:
  - •approx. 10% in the news (*vs* 35% with the subjects), and
  - •approx. 25% in the letters (*vs* 75% with the subjects).
- In the case of news, the number of non-pronominal objects is significantly higher than the number of non-pronominal subjects.
- Increase of non-pronominal objects (from approx. 80% to 94%, especially in the news) => (i) end-weight and information as major principles of the thematic design of the object constituents.

The corpus
Adverbials

text type \ period	1750-1799	1850-1899	1950-1990	Totals
news	132	68	89	289
letters	53	35	39	127
Totals	185	103	128	416

Table 5: Adverbials (per text type and period; pronominality issues not considered)

According to Graphic 1, no diachronic change.

The metrics

- **size/length**:
  - · Wasow (1997:81): grammatical weight implies "size of complexity"
  - · Yaruss (1999:330): "attempts to separate length and complexity are somewhat artificial"
    - – **metric1**: # of words of the subjects, objects and adverbials
    - – **metric2**: # of words up to the 'marker' of the rightmost immediate constituent

The metrics
'Markers'

- – assumption: concept of 'incrementality': "the language processing system must very rapidly construct a syntactic analysis for a sentence fragment, assign it a semantic interpretation" (Pickering *et al* 2000:5)
- – concept: markers alone can characterise the syntactic status of the constituents to which they belong (~ Chomsky's syntactic heads; Hawkins' 2006:209 'Dependency'). The identification of the markers also relies on statistical information (Corley & Crocker 2000:137).
- – Kimball's (1973) 'New Nodes' principle: "grammatical words (e.g. complementizers, conjunctions, articles, etc.) signal the parser to open a new phrase" (reported by Frazier 1979:43)

The metrics
'Markers': examples

- – *Your Ladyship* dares me to stop in my new work! (1751Richardson.X3) [determiner as the marker of the noun phrase]
- – *Demands (...) without which I can no longer answer the Occasions of my Family* (1751Smollett.X3) [preposition as the marker of the prepositional phrase]
- – *Helen & Bill*, by the way, send their fondest regards to you both. (1950Thomas.X9) [conjunction as the marker of the coordinating construction]
- – *the humility which you laud in a character such as that of Macready* has always to me a certain falseness about it – (1876Trollope.X6) [*wh*-proform as the marker of the *wh*-clause]

The metrics
'Markers': examples (cont.)

- – *Nato's first mission* was now complete (1989TIM1.N9) ['s as the marker of the possessive phrase]
- – *the apotheosis of Scobie – culminating for me in the shower of rockets from H.M.'s Navy* – is sublimity. (1960Aldington.X9) [*ing*-form as the marker of the nexusless nonfinite clause]
- – *The declaration of neutrality demanded by the Minister of France*, might have been considered as superfluous [*ed*-form as the marker of the nexusless nonfinite clause]
- – *pleasure-seekers* are notoriously the most aggrieved and howling inhabitants of the universe, (1869Eliot.X6) [noun as the only element in the subject noun phrase]

The metrics

- **density**:
  - **metric3**: number of immediate constituents
  - **metric4**: ratio of (all the) words per immediate constituent

The metrics

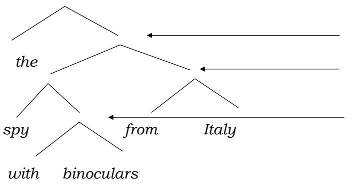
- **depth**:
  - **metric5**: non-terminal-to-terminal ratio, in a 'simple' (non-derivational) syntactic analysis
    - assumption: few non-terminal nodes implies weak complexity
    - phrases (1) and (2) differ as far as complexity is concerned:

(1) the spy with binaculars from Italy ('the spy is from Italy')

(2) the spy with binaculars from Italy ('the binaculars were made in Italy')

The metrics

(1) the spy with binoculars from Italy



the

spy

from

Italy

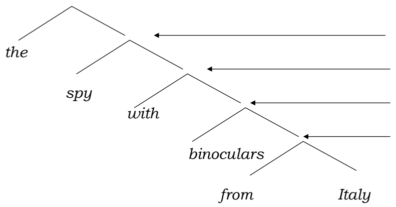
with

binoculars

3 non-terminal levels (Minimal Attachment, Frazier 1979; favoured by Clifton *et al* 1991:266 if the PP is not incoherent as a modifier of *spy* - if it is, then reanalysis takes place)

The metrics

(2) the spy with binoculars from Italy



the

spy

with

binoculars

from

Italy

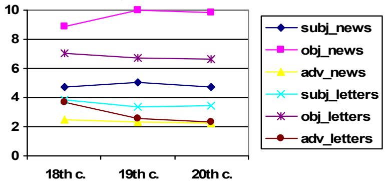
4 non-terminal levels (Late Closure in Frazier 1979 or Recency in Gibson *et al* 1996)

The metrics

- **(lack of) efficiency**:
  - **metric6**: ratio words-up-to-the marker / immediate constituents, inspired by Hawkins' (1994) IC-to-word ratio
  - **metric7**: on-line IC-to-word ratio, based on Hawkins' (1994)
    (aggregate of the partial divisions of the # of immediate constituents by the # of words of such a constituent (up to the marker))

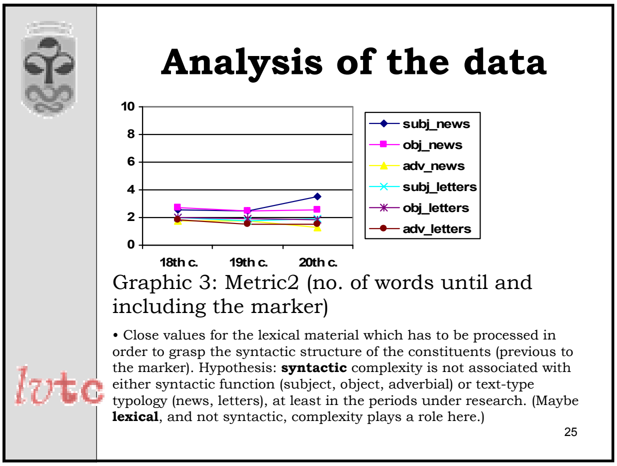
[*The first detachment*] Immediate Constituent 1 3 words	[*of the Austrian reinforcement,*] Immediate Constituent 2 4 words => 7 words up to here	[*amounting to 24,000 men*] Immediate Constituent 3 1 word up to and including the marker => 8 words up to the marker	
1/3 = 33.33%	2/7 = 28.57%	3/8 = 37.5%	aggregate 33.13%

Analysis of the data



Graphic 2: Metric1 (no. of words)

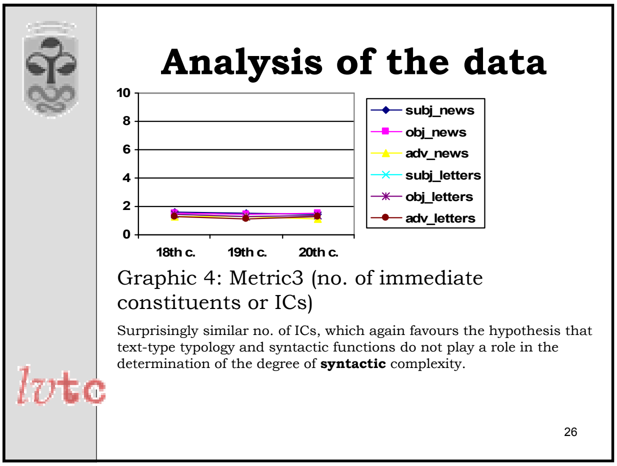
- Objects considerably longer than subjects in the two text types.
- Adverbials shorter than objects and subjects, even though most of them are (either absolutely or relatively) clause-final (further research).
- Objects longer in the news (O'Donnell's 1974: average length of syntactic units in written language is greater than in spoken language)
- Length of (non-pronominal) subjects is similar in the two text types.
- No statistical difference about the length of adverbials (practically identical in the 20th c).



Analysis of the data

Graphic 3: Metric2 (no. of words until and including the marker)

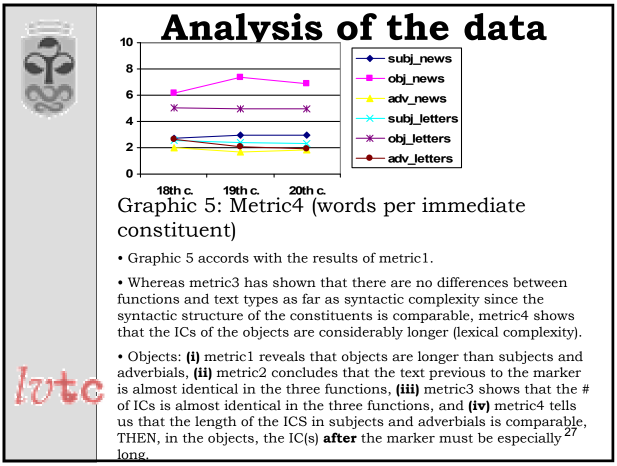
- Close values for the lexical material which has to be processed in order to grasp the syntactic structure of the constituents (previous to the marker). Hypothesis: **syntactic** complexity is not associated with either syntactic function (subject, object, adverbial) or text-type typology (news, letters), at least in the periods under research. (Maybe **lexical**, and not syntactic, complexity plays a role here.)



Analysis of the data

Graphic 4: Metric3 (no. of immediate constituents or ICs)

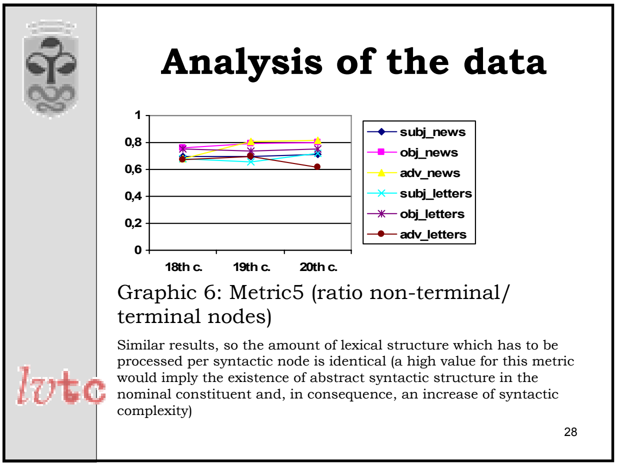
Surprisingly similar no. of ICs, which again favours the hypothesis that text-type typology and syntactic functions do not play a role in the determination of the degree of **syntactic** complexity.



Analysis of the data

Graphic 5: Metric4 (words per immediate constituent)

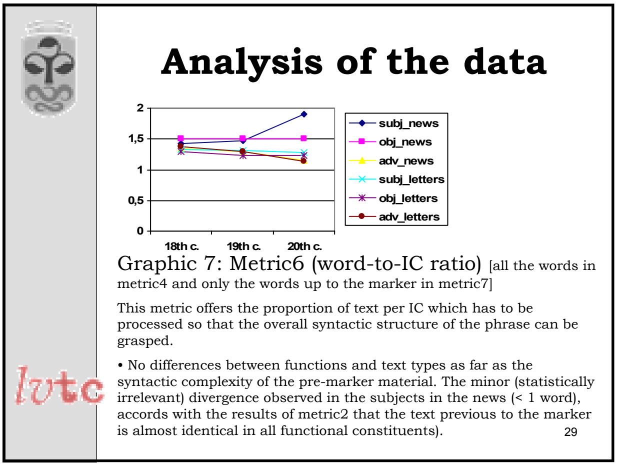
- Graphic 5 accords with the results of metric1.
- Whereas metric3 has shown that there are no differences between functions and text types as far as syntactic complexity since the syntactic structure of the constituents is comparable, metric4 shows that the ICs of the objects are considerably longer (lexical complexity).
- Objects: **(i)** metric1 reveals that objects are longer than subjects and adverbials, **(ii)** metric2 concludes that the text previous to the marker is almost identical in the three functions, **(iii)** metric3 shows that the # of ICs is almost identical in the three functions, and **(iv)** metric4 tells us that the length of the ICS in subjects and adverbials is comparable, THEN, in the objects, the IC(s) **after** the marker must be especially long.



Analysis of the data

Graphic 6: Metric5 (ratio non-terminal/ terminal nodes)

Similar results, so the amount of lexical structure which has to be processed per syntactic node is identical (a high value for this metric would imply the existence of abstract syntactic structure in the nominal constituent and, in consequence, an increase of syntactic complexity)

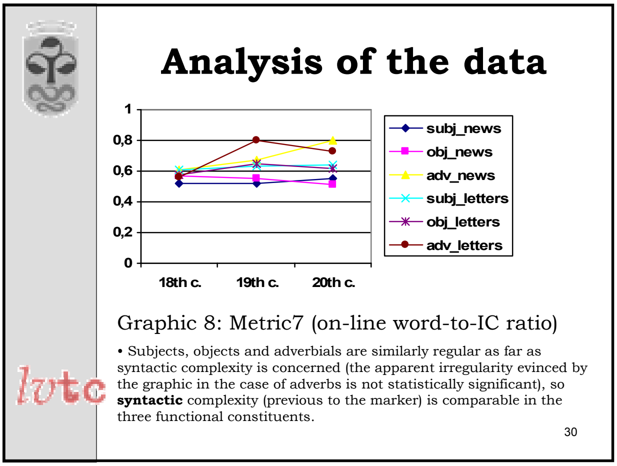


Analysis of the data

Graphic 7: Metric6 (word-to-IC ratio) [all the words in metric4 and only the words up to the marker in metric7]

This metric offers the proportion of text per IC which has to be processed so that the overall syntactic structure of the phrase can be grasped.

- No differences between functions and text types as far as the syntactic complexity of the pre-marker material. The minor (statistically irrelevant) divergence observed in the subjects in the news (< 1 word), accords with the results of metric2 that the text previous to the marker is almost identical in all functional constituents).



Analysis of the data

Graphic 8: Metric7 (on-line word-to-IC ratio)

- Subjects, objects and adverbials are similarly regular as far as syntactic complexity is concerned (the apparent irregularity evinced by the graphic in the case of adverbs is not statistically significant), so **syntactic** complexity (previous to the marker) is comparable in the three functional constituents.

Results of the analysis

- general statistical remark:

  same (relative) proportion of subjects, objects and adverbials in Late Modern and Contemporary English

  (agreeing with Graphic 1)

Results of the analysis

- **pronominal status of the constituents**:
  - subjects:
    - diachronic variation:
      - similar proportions of pronominal and non-pronominal subjects in the three periods
    - text-type variation:
      - news: 65% are non-pronominal
      - letters: 25% are non-pronominal (subjective style)
  - objects:
    - diachronic variation:
      - progressive increase of non-pronominal objects
    - text-type variation: (weaker differences)
      - news: 90% are non-pronominal
      - letters: 75% are non-pronominal

    Thus, the difference is sharper in the case of subjects, where it is conditioned by text-type idiosyncracies, whereas the pronominality of objects is conditioned by general informative and end-weight principles.
  - adverbials: pronominality issues not considered

Results of the analysis

- **Size of the constituents**:
  - similar size of subjects, objects and adverbials in the periods under research
  - subjects: shorter than objects
  - objects:
    - much longer than subjects (**lexical** complexity)
    - longer immediate constituents (**lexical** complexity)
- **syntactic complexity of the constituents:**
  - subjects, objects and adverbials display similar **syntactic** complexity:
    - similar ratio of non-terminal nodes per word
    - similar size of the textual material previous to the marker
    - similar number of immediate constituents
    - similar number of immediate constituents previous to the marker
  - Subjects, objects and adverbials display different **lexical** complexity:
    - Subjects and adverbials: similar size of immediate constituents
    - objects: longer immediate constituent(s) after the marker (**lexical** complexity)

Concluding remarks

- Text-types can be linguistically characterised and can be placed on a scale of complexity by investigating the (linguistic) complexity of the clausal constituents
- Minor diachronic differences between Late Modern and Present-Day English as far as complexity is concerned; only the objects evince a drift towards more **lexical** complexity

Concluding remarks

- **Syntactic** complexity does not play a role in the structural characterisation of non-pronominal subjects, objects and adverbials.
- **Lexical** complexity characterises objects as more complex than subjects and adverbials.
- Proportions of pronominal subjects and objects reveal differences of **lexical** complexity between:
  - news (formal written language)-more complex-
  - and letters (informal speech-like language)-less complex-.
- Non-pronominal constituents reveal that the **lexical** complexity of news and letters is similar in the case of subjects and adverbials, but different in the case of objects (which tend to be more complex in news than in letters).

Food for thought

- Beaman (1984:46): “spoken language is just as complex as written, if not so on some measures”
- Halliday (1985:62): “each [sub-language] is complex in its own way. Written language displays one kind of complexity, spoken language another (...) the complexity of written language is lexical, while that of spoken language is grammatical”

Further research

- more text types (Biber 1992:158: "[w]ritten registers differ widely among themselves in [...] complexity, whereas spoken registers follow a single pattern with respect to their kinds of complexity")
- also marked ('moved', non-preverbal) subjects, subjects in passive sentences and ('moved', non-postverbal) objects
- fine-grained syntactic analysis:
  - differences between adjuncts (modifiers) and adverbial complements (Hawkins 2006, 2007)
  - differences of right- and left-adjunction/branching: differences of positioning of complements and adjuncts relative to the heads
  - complexity according to the semantic typology of adjuncts (Ernst 2002)

References

Arnold, Jennifer E., Thomas Wasow, Anthony Losongco and Ryan Ginstrom (2000) "Heaviness: the effects of structural complexity and discourse status on constituent ordering". *Language* 76/1: 28-55.

Beaman, Karen (1984) "Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse". Ed. Deborah Tannen. *Coherence in spoken and written discourse.* Norwood: NK: Ablex (45-80).

Biber, Douglas (1992) "On the complexity of discourse complexity: a multidimensional analysis". *Discourse Processes* 15: 133-163.

Clifton, Charles Jr., Shari Speer and Steven P. Abney (1991) "Parsing arguments: phrase structure and argument structure as determinants of initial parsing decisions". *Journal of Memory and Language* 30: 251-271.

Corley, Steffan and Matthew W. Crocker (2000) "The modular statistical hypothesis: exploring lexical category ambiguity". Eds. Matthew W. Crocker, Martin Pickering and Charles Clifton Jr. *Architectures and mechanisms for language processing.* Cambridge: Cambridge University Press (135-60).

Crain, Stephen and Donald Shankweiler (1988) "Syntactic complexity and reading acquisition". Eds. Alice Davison and Georgia M. Green. *Linguistic complexity and text comprehension: readability issues reconsidered.* Hillsdale, NJ: Lawrence Erlbaum (167-192).

Davison, Alice and Richard Lutz (1985) "Measuring syntactic complexity relative to discourse context". Eds. David R. Dowty, Lauri Karttunen and Arnold M. Zwicky. *Natural language parsing. Psychological, computational, and theoretical perspectives.* Cambridge: Cambridge University Press (26-66).

Ernst, Thomas (2002) *The syntax of adjuncts.* Cambridge: Cambidge University Press.

Frazier, Lyn (1979) *On comprehending sentences: syntactic parsin strategies.* Blooomington, In.: Indiana University Linguistics Club.

Frazier, Lyn (1988) "The study of linguistic complexity". Eds. Alice Davison and Georgia M. Green. *Linguistic complexity and text comprehension. Readability issues reconsidered.* Hillsdale, NJ: Lawrence Erlbaum (193-221).

Gibson, Edward (1998) "Linguistic complexity: locality of syntactic dependencies". *Cognition* 68/1: 1-76.

Gibson, Edward (2000) "The dependency locality theory: a distance-based theory of linguistic complexity". Eds. Alec Marantz, Yasush Miyashita and Wayne O'Neil. *Image, language, brain. Papers fron the First Mind Articulation Symposium.* Cambridge, MA.: MIT (95-126).

Gibson, Edward, Neal J. Pearlmutter, Enriqueta Canseco-Gonzalez and Gregory Hickok (1996) "Recency preference in the human sentence processing mechanism". *Cognition* 59: 23-59.

Halliday, M.A.K. (1985 [1989]). *Speech and written language.* Oxford: Oxford University Press.

Hawkins, John A. (1994) *A performance theory of order and constituency.* Cambridge: Cambridge University Press.

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

Hawkins, John A. (2006) "Gradedness as relative efficiency in the processing of syntax and semantics". Eds. Gisbert Fanselow, Caroline Féry, Ralf Vogel and Matthias Scxhlesewsky. *Gradience in grammar. Generative perspectives.* Oxford: Oxford University Press (207-226).

Hawkins, John A. (2007) "Performance and grammatical variation in the ordering of verb, direct object and obliques". Plenary lecture read at DGFS, University of Siegen, Feb.

Lewis, P. Shapiro, Patric McNamara, Edgar Zurif, Susan Lanzoni and Laird Cermak (1992) "Processing complexity and sentence memory: evidence from amnesia". *Brain and Language* 42/4: 431-453.

McWhorter, John H. (2001) "The world's simplest grammars are creole grammars". *Linguistic Typology* 5: 125-166.

Miller, George A. and Noam Chomsky (1963) "Finitary models of language users". Eds. R. Duncan Luce, Robert R. Bush and Eugene Galanter. *Handbook of mathematical psychology. Vol. 2.* New York: Wiley (419-492).

Newmeyer, Frederick J. (1998) *Language form and language function.* Cambridge, Mass.: MIT.

O'Donnell, Roy (1974) "Syntactic differences between speech and writing". *American Speech* 49: 102-110.

Pérez-Guerra, Javier and Ana E. Martínez-Insua (2006) "Subjects and complexity in the recent history of English". Paper read at DELS, University of Manchester, April.

Pérez-Guerra, Javier and Ana E. Martínez-Insua (2007) "Do some genres become more 'complex' than others?". Paper read at DGFS (Syntactic Variation and Emerging Genres), University of Siegen, Feb.

Pickering, Martin J., Charles Clifton Jr. and Matthew W. Crocker (2000) "Architectures and mechanisms in sentence comprehension". Eds. Matthew W. Crocker, Martin Pickering and Charles Clifton Jr. *Architectures and mechanisms for language processing.* Cambridge: Cambridge University Press (1-28).

Rohdenburg, Günter (1996) "Cognitive complexity and increased grammatical explicitness in English". *Cognitive Linguistics* 1/2: 149-82.

Taavitsainen, Irma (2001) "Changing conventions of writing: the dynamics of genres, text types, and text traditions". *EJES* 5/2: 139-150.

Wasow, Thomas (1997) "Remarks on grammatical weight". *Language Variation and Change* 9/1: 81-105.

Yaruss, J. Scott (1999) "Utterance length, syntactic complexity, and childhood stuttering". *Journal of Speech, Language, and Hearing Research* 42/2: 329-344.