

Stanisław Goźdz-Roszkowski (ed.)

**Explorations across
Languages and Corpora**

PALC 2009



PETER LANG

Internationaler Verlag der Wissenschaften

**Bibliographic Information published by the Deutsche
Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the
Deutsche Nationalbibliografie; detailed bibliographic data is
available in the internet at <http://dnb.d-nb.de>.

ISSN 1437-5281
ISBN 978-3-631-61677-2

© Peter Lang GmbH
Internationaler Verlag der Wissenschaften
Frankfurt am Main 2011
All rights reserved.

All parts of this publication are protected by copyright. Any
utilisation outside the strict limits of the copyright law, without
the permission of the publisher, is forbidden and liable to
prosecution. This applies in particular to reproductions,
translations, microfilming, and storage and processing in
electronic retrieval systems.

www.peterlang.de

Table of Contents

<i>Stanisław Goźdz-Roszkowski:</i> Introduction	5
--	---

Part One – National Corpora

<i>Mark Davies:</i> Semantically-Based, Learner-Oriented Queries with the 400+ Million Word Corpus of Contemporary American English.....	13
--	----

<i>František Čermák:</i> The Case of the Czech National Corpus: its Design and History	29
---	----

<i>Rafał L. Górski:</i> The Design of the National Corpus of Polish	45
--	----

<i>Adam Przepiórkowski and Piotr Bański:</i> XML Text Interchange Format in the National Corpus of Polish	55
--	----

Part Two – Corpus Tools, Information and Terminology Extraction

<i>Natalia Kotsyba, Andriy Mykulyakand and Igor V. Shevchenko:</i> UGTag: Morphological Analyzer and Tagger for the Ukrainian Language.....	69
--	----

<i>Michal Křen and Martina Waclawičová:</i> Database Framework for a Distributed Spoken Data Collection Project	83
--	----

<i>Adam Przepiórkowski and Grzegorz Murzynowski:</i> Manual Annotation of the National Corpus of Polish with Anotatornia	95
---	----

<i>Danuta Karwańska and Adam Przepiórkowski:</i> On the Evaluation of Two Polish Taggers	105
---	-----

<i>Tomas By:</i> Additional Comments on the Prolog Version of the Tiger Dependency Bank.....	115
--	-----

<i>Marcin Miłkowski:</i> Automating Rule Generation for Grammar Checkers	123
---	-----

<i>Piotr Pęzik:</i> Providing Corpus Feedback for Translators with the PELCRA Search Engine for NKJP	135
<i>Ewelina Kwiatek and Pius ten Hacken:</i> Evaluating the Efficiency of Multiterm Extract for Extraction of English and Polish Terms	145
Part Three – Corpus-based Language Studies	
<i>Janusz Badio:</i> What’s an a “DUCK”? A Corpus Study of Salience and Attention Within Animal-Related, Denominal Verb.....	163
<i>Lukasz Grabowski:</i> Application of Parallel Corpora in Typological Investigations: The Case of Using English-Russian Parallel Subcorpus of the National Russian Corpus in Typological Study of Motion Verbs	179
<i>Milena Herbal-Jeziarska:</i> Corpus-Based Morphology in the Czech Language	195
<i>Maria Perez Blanco:</i> The Language of Evaluation in English and Spanish Editorials: a Corpus-Based Study	207
Part Four – Cognitive Linguistics	
<i>Barbara Lewandowska-Tomaszczyk and Paul Wilson:</i> Culture-Based Conceptions of Emotion in Polish and English	229
Part Five – A Corpus-Assisted Perspective on Discourse Analysis and Ideology	
<i>Cinzia Bevitori:</i> The Meanings of <i>Responsibility</i> in the British and American Press on Climate Change: a Corpus-Assisted Discourse Analysis Perspective.....	243
<i>Mikołaj Deckert:</i> Towards an Axiological Picture of the EU – Evidence from a Polish TV News Corpus	261

Katarzyna Fronczak:
Keywords and the Discourse of the Northern Ireland Peace Process, 1997–
2007. A Case Study in the Election Manifestoes of the Democratic
Unionist Party and Sinn Féin 275

Joanna Kaim-Kerth:
Correspondence Analysis in Discourse Studies..... 295

Inga Massalina:
Selected Cognitive and Discoursal Aspects of the LSP of the Navy..... 303

Part Six – Academic Discourse

Ken Hyland:
Corpora and EAP: Specificity in Disciplinary Discourses 317

Silvia Cacchiani:
Keywords and Key Lexical Bundles as Cues to Knowledge Construction in
RAs in Economics 335

Christoph Haase and Josef Schmied:
Conceptualising Spatial Relationships in Academic Discourse: a Corpus-
Cognitive Account of Locative-Spatial and Abstract-Spatial Prepositions..... 351

Part Seven – Translation

Margaret Rogers:
Translation Memory and Textuality: Some Implications 371

Alejandro Curado Fuentes and Martin Garay Serrano:
Corpus Encoding and Integration for English-Spanish MT..... 389

Dimitra Anastasiou:
Localisation, Centre for Next Generation Localisation and Standards..... 401

Hanne Eckhoff, Dag Haug and Marek Majer:
Making the Most of the Data: Old Church Slavic and the PROIEL Corpus
of Old Indo-European Bible Translations 411

Cécile Frérot:
Parallel Corpora for Translation Teaching and Translator Training
Purposes..... 433

<i>Marlén Izquierdo:</i> The Combination of Comparable and Translation Corpora and How Translators May Benefit from IT	451
<i>Marta Kajzer:</i> Community Interpreting in a Blended Environment – Student and Teacher Assessment	481
<i>Maria Tymczyńska:</i> New XML-encoded Swahili-Polish Dictionary: Micro- and Macrostructure ..	497
<i>Piotr Burmann:</i> Insights into Selected Scientific and Technical Dictionaries Currently Available on the Polish Publishing Market.....	515
<i>Mirosława Podhajecka:</i> Research in Historical Lexicography: Can Google Books Collection Complement Traditional Corpora?	529
Towards the Creation of a Belarusian Grammatical Dictionary.....	547
Part Nine – Language Teaching and Learning	
<i>Alex Boulton:</i> Data-Driven Learning: the Perpetual Enigma.....	563
<i>Carmen Dayrell</i> Anticipatory ‘It’ in English Abstracts: a Corpus-Based Study of Non- Native Student and Published Writing.....	581
<i>Stefano Federici:</i> Analogy-Based Generation of Questions for E-Learning Platforms	599
<i>Przemysław Krakowian:</i> WEBCEF – The Project, Deliverables and Status Quo.....	611

Keywords and Key Lexical Bundles as Cues to Knowledge Construction in RAs in Economics

Silvia Cacchiani

Abstract: Numerous studies have recently focussed on lexical bundles, or extended collocations which we encounter in specific genres more frequently than expected by chance. As such, they help characterize disciplinary discourses. Using the tools of corpus linguistics, we explore a 2.7 million word corpus, focussing on discourse signalling devices, and 5-word, 4-word, and 3-word bundles in particular. This enables us to identify the basic means of knowledge construction in research articles in economics as against research articles in history and business/marketing. Data analysis provides compelling evidence for seeing discourse signals in RAs in economics as realizing conditional prediction and empirical hypothesis within the macro-speech acts of hypothesis, analysis/interpretation/generalization, and prediction.

Keywords: analysis/interpretation/generalization, (conditional) prediction, (empirical) hypothesis, disciplinary variation, discourse signals/discourse signalling, key bundles, keywords, knowledge construction, (macro-)speech acts, RAs in economics.

1. Reasons, aims, methodology

Research articles (RAs) have long been a major concern for EAP researchers. Corpus compilation and query tools have increasingly enabled researchers to detail lexico-grammatical and organizational features of different genres, and shifting the focus on other genres and on cross-disciplinary, cross-generic, and cross-linguistic variation has brought new insights into the whys and wherefores of RAs.

While corpus research has gone a long way in recording and identifying points of overlap and dissimilarities across genres, languages and disciplines, one point that has not received much attention so far is providing quantitative evidence for discourse analytical investigation carried out manually within the framework of speech act theory. In the light of this, we shall use the tools of corpus linguistics in order to provide clear evidence for the claim that knowledge construction in RAs in economics amounts to the complex interplay of the macro-speech acts of *hypothesis*, *analysis/generalization*, and *prediction* (cf. Merlini Barbaresi 1983). Indeed, economists (Brandis 1968) see *prediction* as a distinctive feature of their disciplinary discourse, and *hypothesis*, *analysis*

and *generalization* as ancillary to it. Evidence for this comes from Merlini Barbaresi's (1983) manual investigation of RAs in economics, which enables her to assess the relevance of *prediction* within the genre, categorize more and less frequent and distinctive speech acts, and also highlight relevant lexico-grammatical and organizational devices.

We therefore address the issue of discourse signalling in English RAs in economics in order to concentrate on knowledge construction within the genre. To this purpose, Section 2 verifies our working hypothesis: if *prediction* is a distinctive feature of the discourse of economics and economics RAs in particular, then cross-disciplinary comparison within the humanities will bring to the fore the related linguistic devices variously realizing this macro-speech act. We use the HEM-corpus of research articles¹, a collection of around 2,700,000 tokens equally distributed among the HEM-Economics (HEM-E), HEM-History (HEM-H), and HEM-Marketing (HEM-M). As a first step, we download a keyword list (Mike Scott's 1997 and following) of individual words (§ 2) in the HEM-E as against the HEM-H and HEM-M. Specifically, banking on research on lexical bundles, genre- and disciplinary variation (Biber *et al.* 1999, Biber 2006, Cortes 2006, Hyland 2008), we assume that whereas keywords for individual words might give us some cues to the disciplinary-specific discourse practices of research articles in economics, these are closely mirrored by key 5-word, 4-word, and 3-word bundles. We therefore adapt to our purposes previous classifications of the discourse functions of lexical bundles (§ 3), and in § 3.1 we provide an overview and discussion of the main functions by key lexical bundles in the discourse in economics. Some conclusive remarks and suggestions for future research follow in § 4.

2. Keywords in economics

In this section we compare wordlists from the HEM-E and the HEM-H subcorpora to test whether discourse-signalling devices clearly marking prediction and related macro-speech acts are "key" (cf. Scott 1997) in the HEM-E, i.e. unusually frequent in comparison with the HEM-H. The HEM-M is also used as a reference corpus in order to validate the results. Table 1 and Table 2 give a breakdown of the most salient one-word keywords in the HEM-E,

1 The HEM corpus of research articles was built and is currently held at the University of Modena and Reggio Emilia. The texts were downloaded from a panorama of academic journals in order to cover a good range of subdisciplines within economics (*HEM-Economics*: HEM-E), history (*HEM-History*: HEM-H), and business/marketing/management (*HEM-Marketing*: HEM-M) in the years 1999–2000. For more on this point: <http://www.cla.unimo.it/cofin/>.

respectively contrasted with the HEM-H and the HEM-M. We use italics for the words present in both frequency lists.

Table 1. Keywords (keyness) [HEM-E; HEM-H]

N	WORD	HEM-E. LST		HEM-H. LST		KEYNESS
		FREQ.	%	FREQ.	%	
1	<i>is</i>	44.683	1,71	18.537	0,76	9.526,10
4	<i>are</i>	19.918	0,76	6.697	0,27	6.015,50
7	model	6.575	0,25	594	0,02	5.443,30
11	<i>table</i>	4.405	0,17	329	0,01	3.900,20
15	<i>results</i>	4.242	0,16	341	0,01	3.663,7
19	we	11.833	0,45	4.293	0,18	3.173,80
26	variables	2.640	0,1	94		2.800,10
46	<i>variable</i>	1.821	0,07	83		1.839,70
47	<i>if</i>	7.728	0,3	3.037	0,12	1.806,3
48	data	2.749	0,11	417	0,02	1.765,30
53	sample	1.997	0,08	162		1.718,60
59	will	5.794	0,22	2.051	0,08	1.613,40
65	can	6.663	0,26	2.756	0,11	1.414,00
73	<i>probability</i>	1.467	0,06	107		1.307,30
92	<i>estimated</i>	1.227	0,05	77		1.115,10
103	estimates	1.157	0,04	75		1.069,5
111	models	1.527	0,06	219		1.013,70
135	<i>assume</i>	1.225	0,05	156		868,3
141	hypothesis	1.146	0,04	138		835,7
146	may	4.838	0,19	2.228	0,09	816,2
171	be	16.541	0,63	0,5243	0,46	740,4
179	<i>implies</i>	948	0,04	105		719,5
203	<i>case</i>	3.341	0,13	1.453	0,06	639,7
218	<i>follows</i>	1.012	0,04	171		605,8
246	analysis	2.109	0,08	784	0,03	542,1
252	hence	1.180	0,05	283	0,01	530,8
263	<i>assumption</i>	993	0,04	209		503
330	due	1.014	0,04	277	0,01	397,9
353	therefore	1.854	0,07	793	0,03	367,4

N	WORD	FREQ.	HEM-E. LST		HEM-H. LST		KEYNESS
			%	FREQ.	%	FREQ.	
360	<i>assumed</i>	905	0,03	243			362
377	<i>findings</i>	503	0,02	70			339,1
383	<i>denote</i>	379	0,01	28			336,3
384	<i>suppose</i>	492	0,02	67			335,7
384	<i>mean</i>	879	0,03	246	0,01		335,5
402	<i>shows</i>	1.091	0,04	374	0,02		318,7
409	<i>then</i>	3.252	0,12	1.828	0,07		312,4
423	<i>expectations</i>	593	0,02	123			304,1
437	<i>conditional</i>	313	0,01	19			294,7
439	<i>denotes</i>	259		6			292,9
449	<i>likely</i>	1.245	0,05	494	0,02		268,8
454	<i>find</i>	1.320	0,05	543	0,02		282,7
460	<i>analyse</i>	340	0,01	33			273,3

Table 2. Keywords (keyness) [HEM-E; HEM-M]

N	WORD	FREQ.	HEM-E. LST		HEM-M. LST		KEYNESS
			%	FREQ.	%	FREQ.	
3	<i>the</i>	198.074	7,58	334.918	6,04		6.723,90
5	<i>is</i>	44.683	1,71	59.130	1,07		5.574,10
11	<i>fig</i>	1.912	0,07	42			3.980,70
32	<i>if</i>	7.728	0,3	8.888	0,16		1.515,50
37	<i>table</i>	4.405	0,17	4.114	0,07		1.407,90
75	<i>case</i>	3.341	0,13	3.364	0,06		910,3
95	<i>since</i>	2.369	0,09	2.190	0,04		772,3
131	<i>then</i>	3.252	0,12	3.836	0,07		584,5
136	<i>suppose</i>	492	0,02	132			578,2
150	<i>assume</i>	1.225	0,05	966	0,02		526,7
167	<i>hence</i>	1.180	0,05	948	0,02		494,1
168	<i>implies</i>	948	0,04	662	0,01		488,8
172	<i>probability</i>	1.467	0,06	1364	0,02		472,7
201	<i>follows</i>	1.012	0,04	411	0,01		298,7
220	<i>assumed</i>	905	0,03	724	0,01		381,5

N	WORD	FREQ.	HEM-E. LST %	FREQ.	HEM-M. LST %	KEYNESS
233	<i>expected</i>	2.340	0,09	2.904	0,05	361,6
236	<i>estimated</i>	1,227	0,05	1.250	0,02	361,8
247	<i>assumption</i>	993	0,04	886	0,02	346,4
272	<i>denote</i>	379	0,01	170		314,7
349	<i>there</i>	4.710	0,18	7.432	0,13	246,7
379	<i>are</i>	19.918	0,76	36.996	0,67	227,03
493	<i>shall</i>	206		88		178,1
495	<i>denotes</i>	259		142		178,1

The keywords in Table 1 and Table 2 can be further grouped on the basis of their meaning and function. Specifically, the following categories can be identified:

1. Keywords used to instantiate or name the relevant (macro-)speech act:
 - i. hypothesis (HEM-E; HEM-H: *assume* [863,3], *assumption* [503], *suppose* [335,7]; HEM-E; HEM-M: *suppose* [578,2], *assume* [526,7], *assumed* [381,5], *assumption* [346,4]);
 - ii. analysis/interpretation/generalization (HEM-E; HEM-H: *estimated* [1.115,10], *estimates* [1.069,5], *analysis* [542,1], *find* [282,7], *analyse* [273,3]; HEM-E; HEM-M: *estimated* [361,4]);
 - iii. prediction (HEM-E; HEM-H: *expectations* [304,1]; HEM-E; HEM-M: *expected* [361,6]).
2. Keywords which point to different degrees of certainty, possibility, and probability in hypothesis, analysis and prediction:
 - i. modals (HEM-E; HEM-H: *will* [1.613,40], *can* [1.414], *may* [816,2], *likely* [268,8]; HEM-E; HEM-M: *x*);
 - ii. other verbs (HEM-E; HEM-H: *denote* [336,3], *shows* [318,7], *implies* [719,5]; HEM-E; HEM-M: *implies* [488,8], *denote* [314,7]);
3. Markers of hypothesis or cause and reason relations (cf. Siepmann 2005) between elements within the clause or sentence (HEM-E; HEM-H: *if* [1.806,3], *case* [639,7], *hence* [530,8], *due* [397,9], *therefore* [367,4], *then* [312,4]; HEM-E; HEM-M: *if* [1.515,50], *case* [910,3], *since* [772], *then* [584,5], *hence* [494,1], *follows* [298,7]);

4. Self-reflexive markers pointing to the disciplinary-specific reliance on visuals (HEM-E; HEM-H: *table* [3.900,20]; HEM-E; HEM-M: *fig* [3.980,70]), *table* [1.407,90]);
5. Keywords used to describe and present data (HEM-E; HEM-H: *results* [3.663,7], *data* [1.765,30], *findings* [339,1]; HEM-E; HEM-M: *x*), and statistics (HEM-E; HEM-H: *model* [5.443,30], *variables* [2.800,10], *variable* [1.839,70], *sample* [1.718,60], *probability* [1.307,30], *models* [1.013,70], *hypothesis* [835,7], *mean* [335,5], *conditional* [294,7]; HEM-E; HEM-M: *probability* [472,7]).

What these preliminary keyword lists seem to suggest is a high saliency of items related to hypothesis (*if*, *assum**, *suppose*, *case*) and data analysis/interpretation/generalization in the discourse of economics.

3. Classifying lexical bundles

Lexical bundles (Scott 1997: clusters) are word strings that (a) appear in a genre more frequently than expected by chance, and (b) occur in multiple texts in that genre (Biber *et al.* 1999). Given corpus size and research purpose, we set our frequency cut-off point at 2 hits per million word.

If scientific *prediction* in the discourse of economics is *conditional*, or, in other words, bears on a set of factors or variables and on their interaction, then *prediction* follows from and overlaps with the speech acts of hypothesis and data analysis/generalization, based on the application of a given model within a main argumentative text type (Werlich 1976). Most importantly, there must be some lexical bundles that are key in this respect, also pointing to the writer's different degrees of commitment to the truth of his proposition (in the sense of Searle and Vandervecken 1985; see Hyland's 2008 *stance signals*).

Although lexical bundles are identified on the basis of frequency across texts, they can serve different functions and subfunctions. Slightly adapting Hyland (2008: 13–19), a development from Biber (2006) and Biber *et al.* (2004), we set forth a restricted number of categories and subcategories which bring together insights from research in speech-act theory, genre analysis and discourse studies:

1. *Research-oriented* bundles, which comprise markers of *location*, *procedure*, *quantification* and *identification*, *description*, and *focus*:
 - a) *Location* bundles indicate time and place also in analyses and process descriptions (*in the first period*, *in the long-run*). *Text deixis* (in Table 2),

- a subset of *referential expressions* (term from Biber *et al.* 2004), can be also seen as belonging here;
- b) *Procedure* bundles point to research methodology and purpose. We include here also preliminary definitions, assumptions, and parameters and variables setting, e.g. *is defined as the, it is assumed that*;
 - c) *Quantification* bundles comprise Hyland's (2008) quantification bundles which specify quantity, identifying a position on a scale of quantity (cf. Biber *et al.*'s 2004) or pointing to a movement along the scale (*is the number of, an increase in the, more than # of the*). They comprise Biber *et al.*'s (2004) *referential expressions* which serve as *quantity specification*, but are extended in this study to include classificatory adjectives clearly pointing to an objective measure of frequency (*not significantly different, negative and significant*), and existential *there* and other forms exclusively used to introduce data description or exposition (cf. Biber *et al.*'s 2004 *identification bundles*), e.g. *that there is no, there exists a, there are a number of*;
 - d) *Description* bundles cover a wide range of vocabulary used to present data, usually for the possible outcomes of the application of a model. They make reference to physical or abstract entities. Some clusters here are *the first-order condition for, or the dependent variable is*. Inclusion into the class is not always straightforward, and marginal members of the category clearly point to quantification. One such example is *the probability of*: while it is used to describe the data presented in detailed statistics tables, it provides a quantitative description of the likely occurrence of a particular event. We also group here bundles which specify more general tangible and intangible attributes and features which do not point to a scale of quantity (*the nature of the*);
 - e) *Focus* bundles (in the sense of Hyland 2008)²: they provide information on the subject under investigation within a given research field (*of central bank independence, the inflation tax, Morck et al.*), which are not relevant to our investigation, are not listed in the tables;
2. *Text-oriented* bundles help organize the text and structure its meaning as a message or argument:
- a) *Transition signals*: they explicitate additive and contrastive relations between elements (*on the other hand*), and, we may want to add, coherence relations that do not seem to fit under the following subsets;
 - b) *Resultative signals*, or markers of inferential or causative relations (*this is due to the, these results suggest that*);

² The term *focus* is not taken here to cover Biber *et al.*'s (2004) *focus* bundles, which comprise *discourse organizers* which serve a *topic introduction* function (*if you look at, in this chapter we*) and *referential expressions* used to signal *identification* (*is one of the*).

- c) *Structuring signals*, or text-reflexive cues that organize stretches of discourse or direct the reader elsewhere (*paper is organized as follows, table # and table #, as shown in table*);
 - d) *Framing signals*, which delimit an argument by specifying cases, circumstances of application of models and variables within the model. Some examples are *in the presence of the, if and only if, if the number of*³;
3. *Participant-oriented bundles*, which focus on personal and impersonal (terms from Biber *et al.* 2004) stance and engagement features:
- a) *Stance features*, i.e. writer's attitude and evaluation (*are more likely to be, may be due to*);
 - b) *Engagement features*, directly addressing the reader (*it can be seen that, it is straightforward to show*).

Since stance and engagement features may overlap to some extent, keeping in mind that the main purpose of this paper is to provide evidence for equating disciplinary knowledge constructions with the speech acts of hypothesis, analysis/generalization and prediction, no attempt is made at this stage to set them apart.

Strictly speaking, these categories are not mutually exclusive. Quite the contrary, bringing together features pertaining to various levels of analysis, they may overlap to different extents. To take one example, procedure bundles used for preliminary definitions, assumptions, and parameters and variables setting may also show engagement or stance features (*can be interpreted as, can be thought of as, I estimate the model using*). This, however, does not seem to detract from the ability of the above-mentioned categories to serve as a framework for distinguishing key discourse-signalling lexical bundles.

3.1. Key lexical bundles in economics

Tables 3 and 4 below provide a breakdown of 5-word, 4-word and 3-word lexical bundles. The following data are given for the top ten bundles in each category: ranking, lexical bundle, frequency in the HEM-E, keyness. 4-word and 3-word lexical bundles are not included when they can be subsumed under one or more 5-word lexical bundles with similar keyness value. Single underlining is used for the top 25 5-word, 4-word, and 3-word items in the respective frequency list. As will be seen, inclusion within categories is not always

³ Note that Hyland's (2008) *framing signals* represent a small fraction of Biber *et al.*'s (2004) *intangible framing attributes*, a subset or their *referential expressions* which, the other way round, cut across Hyland's (2008) transition, resultative and framing signals.

straightforward. Most importantly, some lexical bundles do not seem to serve one single function at a time (e.g. *when there are* [quantification; framing signal], *purpose of this paper is* [procedure signal; structuring signal], *results are consistent with the* [resultative signal; procedure]), might serve different functions within different contexts (*can be thought of as* [procedure signal or stance signal], *in the form of a* [description or structuring signal]), or can only be assigned a specific function on the basis of their collocational behaviour within larger bundles (*of this paper is in the purpose of this paper is* [procedure and structuring signal], or *is beyond the scope of*, but not *the scope of this paper*, in *is beyond the scope of this paper* [structuring signal]).

Table 3. Keywords (keyness) for 5-word, 4-word, and 3-word bundles [HEM-E; HEM-H]

RESEARCH-ORIENTED

LOCATION

5-word: 7 in table # and table # [58; 80,3]; 21 panel B of table # [38; 51,4]; 50 the left-hand side of [29; 38,2]; 74 in columns # and # [25; 33]; 88 on the right-hand side [24; 31,6]; 101 column # of table # [22; 30,3]

4-word: 42 in the second period [74; 97,6]; 208 in table # are [37; 48,8]; 498 a model in which [24; 31,6]; 499 in section # and [24; 31,6]

3-word: 40 in the model [207; 272,9]; 59 in our model [166; 218,8]

PROCEDURE

5-word: 10 we assume that the [133; 175,5]; 11 is assumed to be [130; 171,4]; 17 it is assumed that the [50; 52,7]; 37 the null hypothesis of no [32; 42,2]; 108 are summarized in table # [31; 28,7];

4-word: 11 is assumed to be [130; 171,4]; 24 it is assumed that [114; 116,1]; 35 is defined as the [78; 108,2]; 45 can be written as [70; 92,3]; 50 is a function of [94; 87,3]; 71 the change in the [90; 78,6]; 92 is the number of [65; 70,6]; 117 is the sum of [21; 27,7]; 140 can be used to [87; 59]; 219 the hypothesis that [131; 111,3] ...

3-word: 341 I assume that [67; 88,3]

QUANTIFICATION

5-word: 114 more than # of the [21; 27,7]

4-word: 1 an increase in the [326; 354,4]; 3 significant at the # level [102; 134,5]; 26 significantly different from zero [86; 113,4]; 30 the # level of significance [35; 46,1]; 62 of an increase in [61; 80,4]; 47 is equal to the [81; 90,8]; 89 to an increase in [74; 71,8]; 92 is the number of [65; 70,6]; 193 the total amount of [38; 50,1]

3-word: 208 is less than [118; 116,3]; 220 the higher the [97; 11,2]; 222 magnitude of the [106; 111,2]; 245 is higher than [81; 106,8]; 246 equal to # [98; 106,6]; 463 is lower than [58; 76,5]

DESCRIPTION

5-word: 70 the relative size of [26; 34,3]; 72 the standard deviation of the [25; 33]; 73 the present value of the [25; 33]; 79 the dependent variable is the [24; 31,6]; 100 the first-order condition for [23; 30,3]; 103 the short-run Phillips curve [22; 29]; 58 is a decreasing function of [28; 36,9]; 59 is a function of the [37; 36,9]; 69 is an increasing function of [26; 34,3]; 102 an increasing function of the [22; 29]; 124 there is a continuum of [20; 26,4]; ...

4-word: 7 as a function of [184; 192,7]; 33 the distribution of the [80; 105,5]; 61 the effect of a [73; 80,7]; 63 the standard deviation of [61; 80,4]; 71 the change in the [90; 78,6]; 73 the coefficient on the [59; 77,8]; 74 the first-order conditions [59; 77,8]; 92 is the number of [65; 70,6]; 94 the effect of the [56; 70,3]; 99 the probability that the [52; 68,6]; 104 as a percentage of [51; 67,2]; 112 the probability of a [50; 65,9]; ...

3-word: 375 probability of a [64; 84,4]

TEXT-ORIENTED

TRANSITION SIGNALS

4-word: 77 on the other hand [562; 281]

RESULTATIVE SIGNALS

5-word: 26 table # reports the results [36; 47,5]; 44 these results are consistent with [30; 39,5]; 49 table # presents the results [29; 39,5]; 77 results are presented in table [25; 33]; 78 table # shows the results [24; 33]; 84 this is due to the [24; 31,6]; 111 this result is consistent with [21; 27,7]; 128 does not depend on the [20; 26,4]; 129 leads to an increase in [20; 26,4]

4-word: 66 results are consistent with [61; 80,4]; 87 is determined by the [66; 71,9]; 110 the reason is that [66; 65,9]; 157 these results suggest that [42; 55,4]; 166 this implies that the [41; 54]; 170 the results of this [55; 53,3]; 179 is due to the [61; 52,3]; 458 may be due to [25; 33]

3-word: 38 depends on the [281; 282,9]; 89 implies that the [195; 184,5]; 232 so that the [196; 108,3]; 381 such that the [88; 84,3]

STRUCTURING SIGNALS

5-word: 1 table # and table # [184; 242,6]; 2 paper is organized as follows [102; 134,5]; 4 the paper is organized as [77; 101,5]; 5 are presented in table # [78; 101,5]; 8 are reported in table [56; 76,5]; 9 of this paper is to [60; 64,3]; 13 purpose of this paper is [42; 55,4]; 26 table # reports the results [36; 47,5]; 49 table # presents the results [29; 38,2]; 51 are given in table # [29; 38,2]; ...

4-word: 38 table # reports the [76; 100,2]; 422 as shown in the [26; 34,3]

FRAMING SIGNALS

5-word: 23 in the presence of the [112; 124,2]; 48 with respect to the [164; 90,6]; 67 if and only if the [26; 34,3]; 109 in the case of a [45; 28,7]

4-word: 5 if and only if [150; 197,7]; 155 the case in which [42; 55,4]; 157 for the case of [42; 55,4]; 165 the case where the [41; 54]; 457 if the number of [25; 33]; 478 other things being equal [24; 31,6]

PARTICIPANT-ORIENTED

STANCE AND ENGAGEMENT FEATURES

5-word: **12 can be thought of as [43; 59,3]; 20 are more likely to be [38; 50,1];** 31 it can be shown that [48; 46,1]; 40 is easy to see that [31; 40,9]; 89 it can be seen that [24; 31,6]; 139 it is easy to see that; 146 can be interpreted as a [19; 25]; 161 is straightforward to show that [19; 25]

4-word: 188 it can be shown [56; 50,3]; 458 may be due to [25; 33]

3-word: 186 may not be [199; 123,1]; 318 consider the case [83; 93,4]; 471 we consider the [88; 76,2]

Table 4. Keywords (keyness) for 5-word, 4-word, and 3-word bundles [HEM-E; HEM-M]

RESEARCH-ORIENTED

LOCATION

5-word: **4 at the end of the [85; 193,5]; 5 are presented in table # [78; 177,6]; 6 at the beginning of the [62; 141,2]; 11 in table # and table [58; 132]; 18 the right-hand side of [45; 104,5];** 40 panel B of table# [37; 84,2]; 65 the second half of the [31; 70,6]; 76 the left-hand side of [29; 66]; 106 in columns # and # [25; 56,9]; 134 in the rest of the [23; 52,4]; 184 in the second half of [20; 45,5]; ...

4-word: 343 in the present model [21; 30,4]; 348 the model presented in [21; 30,4]

3-word: 466 in our model [166; 54,3]; 494 in panel A [37; 53,5]

PROCEDURE

5-word: **14 it is assumed that the [50; 111,6]; 25 can be thought of as [43; 97,9];** 37 is a function of the [37; 84,2]; 159 is the sum of the [21; 47,8]; 203 can be interpreted as a [19; 43,3]; 220 as a measure of the [18; 41]; 225 I estimate the model using [18; 41]

4-word: 15 is assumed to be [130; 91,5]; 17 it is assumed that [114; 89,1]; 38 it is shown that [49; 68,8]; 79 we assume that the [133; 53,5]; 91 is the ratio of [36; 52,1]; 264 where T is the [23; 33,3]

3-word: 44 the null hypothesis [233; 150,6]; 153 suppose that the [84; 89,6]

QUANTIFICATION

5-word: 46 to an increase in the [36; 82]; 50 the # level of significance [35; 79,7]; 61 the rate of growth of the [31; 70,8]; 133 there are two types of [23; 52,4]; 147 there are a number of [22; 50,1]; 156 more than # of the [21; 47,8]; 159 is the sum of the [21; 47,8]; 198 to the size of the [19; 43,3]; 450 increase in the rate of [13; 29,6]; 452 at the # level in [13; 29,6]

4-word: **3 an increase in the [326; 251,4]; 5 the growth rate of [135; 195,2]; 8 significant at the # [145; 127,5]; 9 the size of the [322; 127,3]; 14 significantly different from zero [86; 97,0];** 250 is equal to # [24; 34,7]; 401 there is some evidence that [14; 31,9]

3-word: **21 is equal to [215; 208,1];** 34 the increase in [223; 166,8]; 251 rise in the [73; 71,1]

DESCRIPTION

5-word: **8 as a function of the** [77; 175,5]; 37 is a function of the [37; 84,2]; 52 effect of a change in [34; 77,4]; 74 percentage change relative to the [30; 68,3]; 81 the probability of success of [29; 66]; 86 is a decreasing function of [28; 66]; 100 is an increasing function of [26; 59,2]; 121 in the form of a [24; 54,6]; 113 the standard deviation of the [25; 56,9]; 122 the first-order conditions for [24; 54,6]; ...

4-word: 48 the present value of [65; 64,5]; 63 a positive relation between [40; 57,8]; 66 the distribution of the [80; 57,7]; 91 is the ratio of [36; 52,1]; 94 the opportunity cost of [77; 51,8]; 131 probability of success of [31; 44,8]; 132 the real value of [31; 44,8]; 139 the coefficient on the [59; 44,1]; 140 the ratio of the [72; 44,1]; 148 percentage change relative to [30; 43,4]; 157 the elasticity of substitution [29; 41,9]; ...

3-word: 131 long-run equilibrium [94; 94,4]; 163 present value of [163; 86,6]; 256 a positive relation [49; 70,6]; 273 the variance of [117; 68,1]; 276 the symmetric equilibrium [47; 68]; 500 the critical value [50; 53,2]

TEXT-ORIENTED

TRANSITION SIGNALS

5-word: 353 as is the case with [14; 31,9]; 354 as well as in the [14; 31,9]; 357 such a way that the [14; 31,9]; 427 in the same way as [13; 29,6]

4-word: **7 on the other hand** [152; 150,9]

RESULTATIVE SIGNALS

5-word: **10 as a result of the** [59; 134,3]; **23 due to the fact that** [44; 100,2]; 45 table # reports the results [36; 82]; 34 the results in table # [38; 86,5]; 45 table # reports the results [36; 82]; 48 on the basis of the [35; 79,7]; 70 these results are consistent with [30; 68,3]; 82 table # presents the results [29; 66]; 95 is due to the fact [26; 59,2]; 11 results are presented in table # [25; 56,9]; 116 this is due to the [24; 54,6]; ...

4-word: 84 in such a way that [29; 66]; 400 it follows that the [31; 28,5]

STRUCTURING SIGNALS

5-word: **1 table # and table #** [184; 418,9]; **2 is organized as follows** [102; 232,2]; **5 are presented in table #** [78; 177,6]; **12 are reported in table #** [56; 127,5]; 26 purpose of this paper is [42; 95,6]; 43 the remainder of the paper [36; 82]; 44 are shown in table # [36; 82]; 62 are summarized in table # [31; 70,6]; 64 as shown in table # [31; 70,6]; 170 table # shows that the [20; 45,5]; ...

4-word: 137 as shown in fig [31; 44,8]

3-word: 412 the basic model [77; 56,9]

FRAMING SIGNALS

5-word: **19 in the context of the** [45; 102,5]; **20 in the case of a** [45; 102,5]; **25 as in the case of** [44; 100,2]; 29 from the point of view [41; 93,3]; 55 in the context of a [33; 75,1]; 94 if and only if the [26; 59,2]; 210 in the sense that the [18; 41]; 221 the extent to which the [18; 41]; 278 in the absence of any [16; 36,4]; 298 in the presence of a [15; 34,1]; 343 if the probability of success [14; 31,9]; 345 if this is the case [14; 31,9]

4-word: **6 if and only if** [150; 151,3]; 152 in the first case [48; 42,9]

3-word: 98 with respect to [594; 108,3]

PARTICIPANT-ORIENTED

STANCE AND ENGAGEMENT FEATURES

5-word: **16 it can be shown that** [48; 109,3]; **21 it should be noted that** [45; 102,5]; **25 can be thought of as** [43; 97,9]; 33 it is important to note [39; 88,8]; 49 it is easy to see [35; 79,7]; 66 is easy to see that [30; 68,3]; 87 it is clear that the [28; 63,7]; 88 it is not possible to [28; 63,7]; 99 it is well known that [26; 59,2]; 108 it is interesting to note [25; 56,9]; 112 it is possible that the [25; 56,9]; ...

Setting aside text-oriented structuring signals which clearly mark organization or purpose of the paper (e.g. HEM-E; HEM-H: *paper is organized as follows, the paper is organized as, purpose of this paper is*), it is easy to see that the remaining discourse signals point to knowledge construction within RAs in economics as the complex interplay of the macro-speech acts of hypothesis, analysis/interpretation/generalization, and (though less represented) prediction. In other words, economic analysis is characterized by the use of models and recurrence to exogenous and endogenous variables as specified by the model. The model enables the analyst to make claims and predictions in the form of hypotheses based on facts. Specifically, the following features are brought to the fore:

1. Hypothesis is a distinctive feature of RA in economics (e.g. HEM-E; HEM-H: *we assume that the, it is assumed that the, is assumed to be, it is assumed that*; HEM-E; HEM-M: *it is assumed that the*).
2. Other procedural signals clearly mirror the reliance of RAs in economics on definitions and parameter settings, which are functional to data analysis and discussion (e.g. HEM-E; HEM-H: *is a function of the, is the sum of the*; HEM-E; HEM-M: *can be thought of as, is the ratio of*).
3. Frequent and highly salient items comprise quantification, description, location and structuring bundles. Quantification markers mirror the dependence of Economics rhetoric on statistics and numeric representation. Related to this, the density of location, description and structuring signals clearly mirrors the reliance of Economics rhetoric on visual representation of numeric data, which enables the writer to identify a reliable basis for his/her inferences and assertions. Some examples: quantification markers such as *an increase in the, significant at the # level, significantly different from zero, the # level of significance* (HEM-E; HEM-H), and *to an increase in the, the # level of significance, an increase in the, the size of the, significantly different from zero* (HEM-E; HEM-M); location, as in *in table # and table, panel B of table #, the left-*

- hand side of (HEM-E; HEM-H), and at the end of the, at the beginning of the, in table # and table (HEM-E; HEM-M); description markers such as the standard deviation of the, the present value of the, the dependent variable is the (HEM-E; HEM-H), and percentage change relative to the, the probability of success of (HEM-E; HEM-M); structuring markers such as table # and table #, are presented in table #, are reported in table (HEM-E; HEM-H and HEM-E; HEM-M).*
4. Applying a model to data analysis also calls for setting and testing individual variables in turn. Specifying cases, particular instances and conditions under which a statement can be accepted and pointing to limitations within the argument is a function of framing devices, which score high among the top 25 key bundles in all frequency lists. Some examples: *in the presence of the, with respect to the, if and only if the, if and only if, the case in which, other things being equal (HEM-E; HEM-H), and in the context of the, in the case of a, as in the case of, if and only if (HEM-E; HEM-M).*
 5. Data interpretation and generalizations based on reliable inferences are introduced by resultative signals that help discuss research processes and outcomes: *table # reports the results, these results are consistent with, this is due to the, leads to an increase in, is determined by the, the reason is that, these results suggest that, this implies that the (HEM-E; HEM-H); as a result of the, due to the fact that, on the basis of the, these results are consistent with, explained by the fact that, these results suggest that the (HEM-E; HEM-M).* Different degrees of certainty can combine with generalizations and data interpretation. Salient stance and engagement bundles are: *it can be shown that (HEM-E; HEM-H), can be thought of as (HEM-E; HEM-H and HEM-E; HEM-M), it is easy to see (HEM-E; HEM-M).* Whereas prediction is not widely represented, stance bundles clearly marking this speech act are *are less likely to be, is more likely to be, can be shown that the, must be the case that (HEM-E; HEM-H), and, among the top 25 key bundles, are more likely to be (HEM-E; HEM-M).*

4. Conclusions

The results presented above provide compelling evidence for seeing knowledge construction in RAs in economics as resulting from the complex interplay of the macro-speech acts of *hypothesis, analysis/generalization, and prediction*. Specifically, such speech acts account for the presence of a number of keywords and key bundles which, the other way round, flesh out prediction in the relevant expository–argumentative text type. Whereas, however, the discourse signals we

identified are clearly amenable to hypothesis, analysis, generalization and interpretation, prediction seems to be relatively underrepresented.

Conditional prediction and empirical hypothesis are precisely a good place to start when looking at RAs in economics from the point of view of genre and (second language) writing⁴. Whereas EAP textbooks essentially cover the language of definitions, data analysis, description of visuals and data interpretation to different extents (see e.g. Swales and Feak 2004, a multidisciplinary course), prediction has not received much attention in the EAP textbooks, possibly because of its highly disciplinary-specific nature. One preliminary question for further consideration, therefore, is expansion of the present analysis to explore conditional prediction in RAs in economics, which amounts to an empirical hypothesis, or a claim, supported by data analysis and bearing on a set of controlled factors and variables. Once we posit a premise-conclusion coherence relation as variously marking prediction, investigation can cover a number of discourse signals, also labelling devices (in the sense of Francis 1994, Moreno 2004), comprising, but not restricted to, the ones identified here (e.g. *if, then, assum**, *therefore, thus, suggest**, *denote, impl**, *due, follow**, *hypothes**, *estimat**, *result**, *find**, *likely, in the context of the, if and only if, these results suggest that, the results show that, it is easy to show that, this implies that the*).

References

- Brandis, R. (1968). *Principles of Economics*. Homewood, Illinois: Irwin.
- Biber, D. (2006). *University Language. A Corpus-based Study of Spoken and Written Registers*. Amsterdam: Benjamins.
- Biber, D., Conrad, S. and V. Cortes (2004). "If you look at ...: lexical bundles in university teaching and textbooks". *Applied Linguistics*, 25 (3): 371–405.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson.
- Francis, G. (1994). "Labelling discourse: an aspect of nominal-group lexical cohesion". In: M. Coulthard (ed.). *Advances in Written Text Analysis*. London: Routledge: 83–101.
- Hyland, K. (2008). "As can be seen. Lexical bundles and disciplinary variation". *English for Specific Purposes*, 27: 4–21.
- Merlini Barbaresi, L. (1983). *Gli atti del discorso economico: la previsione. Status illocutorio e modelli linguistici nel testo inglese*. Parma: Edizioni Zara.

4 See, in this respect, Neugeboren's (2005) concise and accessible set of guidelines (not an EAP textbook) for students writing essays in economics. Note, however, that whereas it takes the student through the stages of planning, revising and editing essays and other term assignments, it does not provide extensive coverage of the language of economic analysis.

- Moreno, A.I. (2004). "Retrospective labelling in premise-conclusion metatext: an English-Spanish contrastive study of research articles on business and economics". *Journal of English for Academic Purposes*, 3: 321–339.
- Neugeboren, R. (2005). *The Student's Guide to Writing Economics*. London: Routledge.
- Scott, M. (1997). *Wordsmith Tools*. Oxford: Oxford University Press.
- Searle, J.R. and D. Vandervecken (1985). *Foundations of Illocutionary Logic*. Cambridge: Cambridge University Press.
- Siepmann, D. (2005). *Discourse Markers across Languages. A Contrastive Study of Second-level Discourse Markers in Native and Non-native Text with Implications for General and Pedagogical Lexicography*. London: Routledge.
- Swales, J.M. and C.B. Feak (2004) [2nd ed.]. *Academic Writing for Graduate Students. Essential Tasks and Skills*. Michigan: Michigan University Press.
- Werlich, E. (1976). *A Text Grammar of English*. Tübingen: Quelle und Meyer.