A Corpus-Informed Text Reconstruction Resource for Learning About the Language of Scientific Abstracts

Laura M. Hartwell[a*] and Marie-Paule Jacques[b]

a. Grenoble 1 - LIDILEM, UJF, Valence, France
b. Grenoble 1 - LIDILEM and IUFM, Grenoble, France

Abstract. Both reading and writing abstracts require specific language skills and conceptual capacities, which may challenge advanced learners. This paper draws explicitly upon the *Emergence* and *Scientext* research projects which focused on the lexis of scientific texts in French and English. The teaching objective of the project described here was to create a collection of text reconstruction tasks targeting the patterns of English that are uncommon in French. These tasks are to be integrated within the platform *Enigma Plus* (http://elang.ujf-grenoble.fr/enigma/). The current project is the conception of a new module based on data-driven materials collected from Scientext, a corpus of medical and biology abstracts in English (http://scientext.msh-alpes.fr/scientext-site-en/spip.php?article9). This paper discusses the task focusing on the word *hypothesis,* the first of a dozen tasks based on authentic examples and designed to help learners of English as a foreign language to better read and write science abstracts. The results revealed several similarities and contrasts with the French findings. These results were integrated into the text reconstruction task. Findings of user practices reported in previous studies were taken into account to optimize completion of the task by the widest range of user practices and errors.

Keywords: corpora, abstracts, on-line text reconstruction, English for specific purposes, English as a foreign language.

1. Introduction

The reading and writing of abstracts requires specific language and conceptual capacities that may challenge even language skills of advanced learners. These ubiquitous, dense, and brief texts are a key element of written academic discourse as they serve to publicly announce one's work thereby enabling other researchers to identify it among the thousands of other published articles. Scientific abstracts contain

* Contact author: hartwell@ujf-grenoble.fr

In L. Bradley & S. Thouësny (Eds.), *CALL: Using, Learning, Knowing, EUROCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings* (pp. 117-123). © Research-publishing.net Dublin 2012

rhetorical and structural aspects which can be identified through a cluster of linguistic features (Cremmins, 1982; Pho, 2008; Swales & Feak, 2004).

An efficient comprehension of abstracts is essential to productive research by learners of English as a foreign language. In this context, descriptive grammar analyses are essential to language teaching (Oakey, 2002) and especially within contexts of language learning for specific purposes (Gledhill, 2000, 2011; Hartwell, 2011). Citing previous studies, McEnery and Wilson (1996) highlight the substantial differences between language use as empirically revealed through corpora study and the descriptions found in textbooks that may misleadingly offer less common language choices to the detriment of learning more frequent ones. Frequency is a condition for both *collocation*, referring to words that are frequently found together and lexico-grammatical patterns which Hunston and Francis (2000) define as "all of the words and structures which are regularly associated with the word and which contribute to its meaning" (p. 37).

This paper draws explicitly upon the *Emergence* and *Scientext* research projects which focused on the lexis of scientific texts in French and English (Cavalla & Grossmann, 2005; Tutin, 2010). One objective of the previous and current research is to identify collocations or patterns in French and English in order to help foreign language learning. The translation of a collocation does not necessarily employ the same structure as found in the original language. Tutin (2010) offers the example of *émettre une hypothèse* (emit a hypothesis), which can be translated by the English verb *hypothesize*, although no such verb exists in French (p. 136).

The teaching objective of the project described here is to create a collection of text reconstruction tasks targeting the patterns of English that are uncommon in French. These tasks are to be integrated within the platform *Enigma Plus,* which was initially designed to accompany the textbook *Minimum Competence in Scientific English* (Blattes, Jans, & Upjohn, 2003). The platform includes short unauthentic recordings accompanied by synchronized visual supports. After the presentation, a skeleton of the text is automatically displayed on the screen including the first two letters of each word to be identified. If the user types a correct word it appears throughout the skeleton, if not, the user is encouraged to enter a new word or listen to the text. This platform is an adaptation of John Higgins's Storyboard, which emanated from his program Rebuild, inspired by Tim John's Textbag in the early 1980s (Davies, 2007). This paper discusses the task focusing on the word *hypothesis,* the first of a dozen tasks based on authentic examples and designed to help learners of English as a foreign language to better read and write science abstracts.

2. Method

This section begins with a brief description of previous studies of the use of the French word *hypothèse.* Then, a comparison with English is formed by consulting the Scientext corpus. Scientext is a collection of academic works in both French and

English (Falaise, Tutin, & Kraif, 2011; Tutin, Grossmann, Falaise, & Kraif, 2009). The peer-reviewed articles in English, collected by the LiCorn team at the Université de Bretagne-Sud, were originally published by the editor BioMed Central and comprise sixty-two subthemes from the fields of biology and medicine. The corpus of abstracts counts 787,276 words from 3,381 research articles. From the results in both languages, exemplars of expressions were drawn to write a 300-word text for the text reconstruction task.

2.1. Corpus-based analysis of the French word 'hypothesis'

Tutin (2010) consulted the *Cultural Identities in Academic Prose* corpus (KIAP) for the productive relations of the French noun *hypothèse* (pp. 99-100). The most frequent collocation is as the subject of the copula verb *être* (to be), with 1,255 tokens. By order of frequency, the verb *être* was followed by six attributes (*autre* "other", *different* "different", *même,* "same"…) each with 78 to 195 tokens. After the nouns *travail* (work) and *capital* (capital) linked by *de* (of), is a second verb *faire* (to make) with 48 tokens.

Cavalla and Grossmann (2005) took a complementary approach by examining the lexical verbs found in collocation of the noun *hypothèse.* Their study confirms that the first lexical verb to be collocated with *hypothèse* is the French *faire* (to make). Furthermore, they separate the verbs into four categories: propose, elaborate, verify, and argue.

For the present study, these categories have been regrouped into two sets: propose or elaborate and verify or argue (Appendix 1). There are 182 tokens in the first category; the eleven entries include the verb *faire* (make), but also *avancer* (to advance) and *émettre* (to emit). There are fewer tokens (104) but more variety in the second category, in which *tester* (to test), *confirmer* (to confirm), and *défendre* (to defend) head the list of 20 verbs.

2.2. Scientext analysis of the English lemma 'hypothesize'

The Scientext English corpus of abstracts was consulted for the lemma *hypothesis.* A total of 163 occurrences were detected. Thirty-four subheadings found within the abstracts were removed as well as one occurrence inserted within parentheses, leaving 128 tokens (Appendix 2). The results revealed several contrasts with the French findings. The verb *hypothesize* was found 73 times, most often conjugated in the past tense. Contrary to the French results, there were few tokens (14) and a variety of lexical verbs (9) within the category "propose or elaborate".

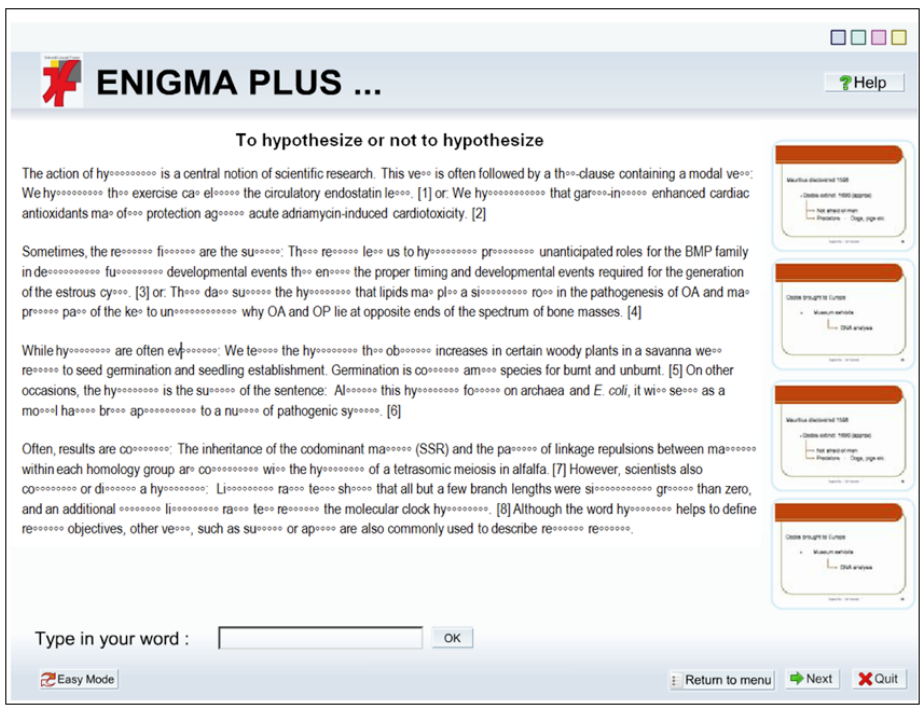
Within the category "verify and argue", there were a similar amount of tokens (88), verb variety (21) and use of the verb *test* in both languages. In English as in French, *hypothesis* was also the agent of several actions, including *involve, consider, focus on, imply, predict.* There were relatively few occurrences of the lemma *be* compared to the French. Furthermore, the adjectives *different* (2)*, other* (1), *first* (2), *same* (0) were rarer, however the expression *working hypothesis* (5) mirrored the use in French

(c.f. Tutin, 2010). *Hypothesis* was also found in ten prepositional phrases and within four compound nouns (e.g., *hypothesis tests*), a grammatical construction not found in French.

3. Results

Drawing upon the comparison of the corpora results, eight complete sentences containing frequent uses of the lemma *hypothesize* were chosen as exemplars. Since the verb *hypothesize* is not found in French, it was put forth in the incipit and in the title *To hypothesize or not to hypothesize.* Research has shown that the first part of the reconstruction activity receives more attention from users (Hartwell, 2010a). The next section highlights the notion of research data as the sentence subject and contains the frequent collocation "supports" (Appendix 3). The third paragraph introduces the transparent lemma "test", which was the most frequent lexical verbal collocation of *hypothesis* in English. The last section focuses on the common expression containing a preposition (*are consistent with the hypothesis*), before finishing with the notion of contradiction (Figure 1).

Figure 1. Slide of text before user begins reconstruction



Previous studies have shown that two-thirds (65.5%) of the users will enter 100 entries or more, but only 22.5% will enter more than 150 entries (Hartwell, 2010b). This task includes 92 missing words, which represents 68 different words as several are repeated. The user only enters each individual word once; hence *hypothesis* will appear eight times when entered the first time by the user. These quantities were calculated to optimize completion of the task given a range of user practices and error as noted by the previous studies.

4. Discussion

This task is the first of a dozen to be created for the platform *Enigma Plus.* The lemma *hypothesize* was chosen as previous studies had evaluated the French use of this term within scientific discourse, in which it is most frequently collocated with the verb *faire* (make). However, among the 542 verbs found within the English abstracts of Scientext, the 50 most frequently occurring verbs constituted approximately ninety percent of all the verbs, but *make* was only 38th on the list and was not found to collocate with *hypothesis,* thereby confirming non-transparent differences across the two languages. This task targets discourse features that are unfamiliar to French speakers as they do not mirror practices of the first language.

Within the list of most frequent verbs related to describing the processes of scientific research, we find *show, compare, suggest, report, determine, examine,* and *appear* (Hartwell, forthcoming). For this reason, the reconstruction text ends with a note about two of these more common verbs: *suggest* and *appear.* This comment is also intended to encourage users to complete further reconstruction tasks.

Acknowledgements. The comparative research aspects of this study were financed by the Grenoble 1 Pôle SHS. The on-line computer assisted language learning aspects were funded by Pedagotice of Grenoble 1 and PRES of the Universities of Grenoble.

References

Blattes, S., Jans, V., & Upjohn, J. (2003). *Minimum Competence in Scientific English – Supplementary Materials.* Les Ulis : EDP Sciences. Retrieved from http://grenoble-sciences.ujf-grenoble.fr/pap-ebooks/upjohn/unit9_1

Cavalla, C., & Grossmann, F. (2005). Caractéristiques sémantiques de quelques « Noms scientifiques » dans l'article de recherche en français. *Akademisk Prosa, 3,* 47-59.

Cremmins, E. T. (1982). *The Art of Abstracting.* Philadelphia: ISI Press.

Davies, G. (2007). *Total Cloze Text Reconstruction Programs: A Brief History.* Retrieved from http://www.ict4lt.org/en/FWTHistory.doc

Falaise, A., Tutin, A., & Kraif, O. (2011). Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. *Proceedings from TALN, Montpellier 2011.* Retrieved from http://pro.aiakide.net/publis/2011TALNPaper-Falaise-Tutin-Kraif.pdf

Gledhill, C. J. (2000). *Collocations in Science Writing.* Tübingen: Gunter Narr Verlag.

Gledhill, C. J. (2011). The 'Lexicogrammar' Approach to Analysing Phraseology and Collocation in ESP Texts. *La Revue du GERAS*, *59*, 5-23.

Hartwell, L. (2010a). Impact of software design on on-line text reconstruction. *SYSTEM: An International Journal of Educational Technology and Applied Linguistics, 38*(3), 370-378. doi: 10.1016/j.system.2010.06.009

Hartwell, L. (2010b). Pratiques de reconstruction de texte en autoformation. *Les Cahiers de l'APLIUT, 29*(2), 81-96.

Hartwell, L. (2011). Learning On-Line about Modality in Written and Oral English. *Proceedings from ICT for Language Learning.* Florence, Italy, 2011.

Hartwell, L. (forthcoming). Corpus-informed descriptions: English verbs and their collocates in science abstracts. *Études en didactique des langues*.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins Publishing Company.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Pho, P. D. (2008). Research Article Abstracts in Applied Linguistics and Educational Technology. *Discourse Studies, 10*(2), 231-250.

Oakey, D. (2002). Formulaic Language in English Academic Writing: A Corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using Corpora to Explore Linguistic Variation* (pp. 111-129). Amsterdam: John Benjamins Publishing Company.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed.). Ann Arbor: University of Michigan Press.

Tutin, A. (2010). *Sens et combinatoire lexicale : de la langue au discours* (Unpublished Dossier en vue de l'habilitation à dirigier de la recherche). Grenoble: Université de Stendhal.

Tutin, A., Grossmann, F., Falaise, A., & Kraif, O. (2009). Autour du projet Scientext: étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. *Linguistique de Corpus*. Retrieved from http://w3.u-grenoble3.fr/lidilem/labo/file/Lorient_vfinale.pdf

Appendix 1. Lexical verbs collocated with *hypothèse*

Action in relation to the hypothesis	Lexical verbs (number of tokens)
Propose / elaborate (182 tokens)	Faire (113), avancer (17), émettre (16), poser (9) formuler (9), proposer (6), effectuer (4), présenter (4), introduire (2), énoncer (1), former (1)
Verify / argue (104 tokens)	Tester (35), confirmer (12), défendre (9), valider (6), vérifier (6), justifier (4), renforcer (4), infirmer (3), corroborer (3), discuter (3), étayer (3), examiner (3), mettre à l'épreuve (3), conforter (2), privilégier (2), soutenir (2), appuyer (1), légitimer (1), opposer (1), récuser (1)

Appendix 2. Collocates of *hypothesis*

Action in relation to the hypothesis	Verbs (number of tokens) or head noun (number of tokens)
To hypothesize (73 tokens)	hypothesized (35), hypothesize (19), is/are hypothesized (8), has/have been hypothesized (6), have hypothesized (1), hypothesizing (1), may hypothesize (1), was hypothesized (1), hypothesized (1 – part participle as modifier)
Propose / elaborate (14 tokens)	lead to (3), present (3), discuss (2), propose (2), address (1), prompt (1), pursue (1), offer (1), illustrates (1)
Verify /argue (88 tokens)	Test (40), support (21 – including "gave support to"), confirm (2), involve (2), strengthen (2), affected by (1), appears to depend on (1), base on (1), consider (1), contradict (1), disprove (1), evaluate (1), examine (2), explore (1), focus on has (1), imply (1), investigate (1), predict (1), prove (1), reject (2), use (2)
To be (12 tokens)	was (5), is (4), if the … is true (3),
Other (10 tokens)	consistent with the (5), in agreement with (1), under the [noun phrase] hypothesis (1), in the hypothesis that (1), in accord with (1), compatible with the (1)
Modifier within a compound noun (4 tokens)	Tests (1), generating study (1), testing (1), null-hypothesis behavior (1)

Appendix 3. Reconstruction text

To hypothesize or not to hypothesize

The action of hypothesing is a central notion of scientific research. This verb is often followed by a that-clause containing a modal verb: **We hypothesize that exercise can elevate the circulatory endostatin level**. [1] or: **We hypothesized that garlic-induced enhanced cardiac antioxidants may offer protection against acute adriamycin-induced cardiotoxicity**. [2]

Sometimes, the research findings are the subject: **These results lead us to hypothesize previously unanticipated roles for the BMP family in determining fundamental developmental events that ensure the proper timing and developmental events required for the generation of the estrous cycle.** [3] or: **These data support the hypothesis that lipids may play a significant role in the pathogenesis of OA and may provide part of the key to understanding why OA and OP lie at opposite ends of the spectrum of bone masses.** [4]

On other occasions, the hypothesis is the subject of the sentence: **Although this hypothesis focuses on archaea and *E. coli*, it will serve as a model having broad applicability to a number of pathogenic systems.** [5] When being evaluated, it often becomes a direct object: **We tested the hypothesis that observed increases in certain woody plants in a savanna were related to seed germination and seedling establishment.** [6]

Results may confirm a hypothesis: **The inheritance of the codominant markers (SSR) and the pattern of linkage repulsions between markers within each homology group are consistent with the hypothesis of a tetrasomic meiosis in alfalfa.** [7] However, scientists also contradict or disprove a hypothesis: **Likelihood ratio tests showed that all but a few branch lengths were significantly greater than zero, and an additional likelihood ratio test rejected the molecular clock hypothesis.** [8] Although the word hypothesis helps to define research objectives, other verbs, such as suggest or appear are also commonly used to describe research results.

Published by Research-publishing.net
Dublin, Ireland; Voillans, France
info@research-publishing.net

CALL: Using, Learning, Knowing
EUROCALL Conference, Gothenburg, Sweden
22-25 August 2012, Proceedings
Edited by Linda Bradley and Sylvie Thouësny

Typeset by Research-publishing.net
Cover design: © Raphaël Savina (raphael@savina.net)
Aquarelle reproduced with kind permission from the illustrator: © Sylvi Vigmo (sylvi.vigmo@ped.gu.se)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-03-2 (paperback)
Print on demand (lulu.com)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Bibliothèque Nationale de France - Dépôt légal: décembre 2012.