New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material.

Fanny Meunier & Céline Gouverneur
To appear in Aijmer, K. (ed.) *Corpora and Language Teaching* . Benjamins.

Introduction

The present paper addresses the relationship between corpora and one commonly used type of pedagogical material in instructed English as a Foreign Language (EFL) settings, namely the textbook. Major English Language Teaching (ELT) publishers increasingly use native and learner corpora as input material on which to build, at least in part, new series of reference and pedagogical material such as dictionaries, grammar or vocabulary books. Surprisingly, however, English for General Purposes (EGP) textbooks are exceptions to the rule and still shy away from corpora. After a first section on the corpus tradition in ELT, we provide a survey of textbook research in section 2. The third section presents a new type of pedagogically annotated textbook corpus: the TeMa corpus (corpus of **Te**xtbook **Ma**terial). From a methodological viewpoint, we will show that the collection and annotation of pedagogical corpora allow for new automated search options which are not available in the manual 'page by page' approach often adopted in traditional textbook analyses. Concrete examples of search options and their preliminary results will be presented in section 4, together with suggestions on how pedagogically annotated corpora can help learners, teachers, and material designers meet new educational challenges.

1. ELT material and the corpus tradition

Corpora are no longer absent from the learning and teaching scene. Their legitimacy as useful pedagogical aids has now been established, and publications addressing the use of corpora in second/foreign language teaching abound: Burnard and McEnery (2000), Sinclair (2004) or Connor and Upton (2004) provide comprehensive edited volumes on the use of corpora in language teaching and learning; Botley, McEnery and Wilson (2000) focus on the use of multilingual corpora in teaching and research; Granger, Hung and Tyson (2002) address the links between computer learner corpora, second language acquisition and foreign language teaching; Mukherjee and Rohrbach (2006) and O'Keeffe, McCarthy, and Carter (2007) deal with the use of native and learner corpora in the classroom and the necessary mediation between research findings in corpus linguistics and classroom pedagogy.

Corpora have not only been valued in applied linguistics circles, they have found their way to the offices of major ELT publishers. The latter increasingly use native and learner corpora as a source of authentic data on the basis of which they build, at least in part, new series of reference and pedagogical material such as dictionaries, grammar or vocabulary books. The use of corpora has even become a selling point. Cambridge University Press (CUP) uses a special logo (viz. ◉) to advertise corpus-based publications and presents the various types of publications in tables[1] classified according to levels[2] (beginner, elementary, pre-intermediate,

[1] See http://www.cambridge.org/elt/corpus/corpus_based_books.htm for a presentation of the tables.

[2] The six levels usually correspond to the A1, A2, B1, B2, C1 and C2 levels of the Common European Framework for languages (CEF, see http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf)

intermediate, upper-intermediate and advanced) and type of publication. (including a mixture of target audience and contents specification: i.e. 'adult', 'exam', 'professional English', 'Cambridge copy collection', 'grammar', 'vocabulary', 'dictionaries', 'methodology and linguistics'). Other publishers focus on the link between their in-house corpus and their dictionaries (see for instance Longman and the Longman Corpus Network[3]).

As for the type of corpus used, ELT publishers have a marked preference for native corpora as a reference resource to inform their material. This preference for native corpora is justified by the need to present real, authentic English. In that respect, whilst CUP offers a 'real English guarantee' to the buyers and users of their material (http://www.cambridge.org/elt/corpus/corpus_based_books.htm), Longman assures its readership that [they] 'only see real English, as it is really used'. http://www.longman.com/dictionaries/corpus/index.html. As for MacMillan, the use of their World English Corpus is described as 'a unique modern database of over 200 million words revealing fresh information on how words are used and natural examples of English as it is written and spoken now!' (http://www.macmillandictionary.com/aboutcorpus.htm). Learner corpora are also being used by ELT publishers, but to a much smaller extent. Here again, when publishers refer to learner corpora, they seem to privilege in-house learner corpora (see for instance CUP and the Cambridge Learner Corpus at http://www.cambridge.org/elt/corpus/learner_corpus2.htm or Longman and the Longman Learner Corpus at http://www.longman.com/dictionaries/corpus/learners.html).

One of the first points that will be made in the present paper is that one extensively used type of ELT material which has, to date, not yet benefited from the corpus revolution is the English for General Purposes (EGP) textbook. Indeed, whilst publishers tend to acknowledge some sort of connections between EGP textbooks and corpora in terms of vocabulary selection or grammar syllabus, they admit that current textbooks are not corpus-based. To a query on the possible corpus-based nature of various EGP textbooks (Merlevede, 2006) Longman stated that "although [name of a textbook] uses the corpus-based Longman Grammar of Spoken and Written English for both the grammar and vocabulary syllabuses, the course itself is not corpus-based" (idem, 2006:94), McMillan stated that "the choice of vocabulary in [name of a textbook] is heavily influenced by the MacMillan English Dictionary which is based on the World English Corpus" (idem, 2006: 94), and CUP explained that they "have tried to tie [name of a textbook] into the Common European Framework as far as possible" […], that they "have [their] own Cambridge Corpus of over 700 million words" […] but that "[h]owever, [they] would not say that [name of a textbook] is corpus based as it began life before the Cambridge Corpus was really in full swing across the Press". CUP however states that they "hope to include [their] Corpus in all Secondary ELT titles in the future". (idem, 2006: 95)

Publishers seem to acknowledge the importance of corpora in ELT but fail to give information on how exactly the corpus is used (or could be used) to flesh out the linguistic contents of their textbooks.

2. A survey of textbook studies

2.1. General overview

[3] Visit http://www.longman.com/dictionaries/corpus/index.html for more details on the Longman Corpus Network

Increased interest in textbook analysis goes back to the early 1980s and several lines of research can be distinguished: development of general criteria for textbook analysis (Williams 1983, Cunningsworth 1984 and 1985, Chambers 1997, Sheldon 1988), assessment of textbooks as a useful/useless type of pedagogical material (Swales 1995, O’Neill 1993, Ranalli 2003, Harwood 2005), focus on specific classroom activities (Jacobs and Ball 1996), and focus on specific lexical or grammatical contents (Gabrielatos 1994, Biber et al 2004, Römer 2004a and b, Koprowski 2005, Meunier and Gouverneur 2007) Table 1 provides a non-exhaustive survey of textbook studies carried out over the last two decades[4] in terms of learning context (e.g. EFL, ESL, EGP, etc), type of textbook analysed (e.g. international, national etc), method of textbook analysis adopted (i.e. manual page-by-page analysis or automated corpus-based approach), level of proficiency addressed and number of volumes examined.

Table 1: Overview of textbook research over the last two decades

Research area	**Author**	**Focus**	**Learning context**	**Textbook type**	**Method adopted**	**Level**	**No of vol.**
Authenticity	Römer (2004a)	modal auxiliaries	EFL	local: German EFL textbook & grammar	manually	secondary school	6
	Römer (2004b)	if clauses – spoken language	EFL	local: German	corpus-based (GEFL TC)	secondary school	12
	Römer (2006)	progressives (spoken data)	EFL	local: German	corpus-based (GEFL TC)	secondary shool	12
	Gilmore (2004)	discourse features	EFL EGP	international	page by page / manual		7 + 3
	Anping (2005)	vocabulary grammar	EFL	international + local (China)	corpus-based	5 levels: beginner to university	50
	Hyland (1994)	Modals	EAP				22
	Gabrielatos (1994)	possessives demonstrative	EFL EGP	international	page by page	Beginner	1
Grammar	Nitta and Gardner (2005)	grammatical tasks	EFL EGP	international	unspecified (page by page)	Intermediate	9
	Boxer and Pickering (1995)	speech acts: complaints	?	?	?	?	7
	Vellenga (2004)	- metalang. - explicit treatment of speech acts - metapragm. information	ESL & EFL	EFL: Integrated skills ESL: grammar books	page by page		8
Pragmatics	Miura (1997)	oral communication	ELT	“Government-authorized”		senior high school	16
	Cane (1998)	conversation skills	ELT				
Speaking	Chujo (2004)	vocabulary levels	- EGP - ESP	local: Japanese	corpus-based (wordlists)	intermediate and advanced	7
	Ranalli (2003)	learning strategies: repetition	- EFL - EGP	international	page by page	upper-intermediate	3

[4] Most of these studies deal with more than one linguistic or language teaching aspect. For clarity’s sake, the main focus was retained as the criterion for classification.

Vocabulary and phraseology	Reda (2003)	resource use recording vocabulary norms	- EFL - EGP voc books	international	unspecified (page by page)	beginner to advanced	6
	Gabrielatos (1994)	collocations	EFL	international	unspecified (page by page)		3
	Hill (1996)	verb form clustering	EFL	coursebooks and grammars		beginner	?
	Biber et al. (2004)	lexical bundles	EAP	American	corpus-based (T2K-SWAL)	University	
	Koprowski (2005)	lexical phrases	EFL EGP	international	manual (list)	intermediate and upper-intermediate	3
	Meunier and Gouverneur (2007) EFL	phraseology	EFL EGP	international	corpus-based (TeMa corpus)	Advanced	5
	Gouverneur (in press) EFL	high-frequency verbs	EFL EGP	international	corpus-based (TeMa corpus)	intermediate and advanced	3
	Gabrielatos (1994)	pronunciation	EFL EGP	international	page by page	Beginner	1
	Swales (1995; 2002)		EAP				
Other	Jacobs & Ball (1996)	group activities writing – grammar books	EFL EGP				
	Biber et al. (2002)		EAP				
	Paltridge (2002)	dissertation writing	EAP				
	Moreno (2003)	language of cause and effect					11

As can be seen from Table 1, studies have addressed a wide range of linguistic aspects with grammar and vocabulary taking centre stage. The focus on the authenticity of input is equally important and reflects the heated debate over authentic versus non-authentic (or adapted/simplified) language, i.e. whether pedagogical materials should be authentic or not, and to what extent they should be adapted to specific learners or classroom contexts (for an in-depth discussion on the issue, see Widdowson’s (2003) review on the topic). The table also shows that several studies deal with specific registers such as English for Academic Purposes (Swales 2002; Paltridge 2002; Biber et al. 2002) and that phraseology constitutes a recent research interest (Biber et al. 2004; Koprowski 2005, Meunier and Gouverneur 2007, Gouverneur in press).

2.2. Analysing textbooks: page-by-page vs. corpus approach?

The methodological approach adopted in the studies listed in Table 1 deserves further attention. Most studies were carried out using a manual, page-by-page approach. Only six recent studies have been conducted using more automatic methods: Biber et al (2004) Römer (2004b; 2006); Chujo (2004), Anping (2005), Meunier & Gouverneur (2007) and Gouverneur (in press and forthcoming). To our knowledge, the corpora exploited in the aforementioned studies are the only published cases of textbook corpora. The first textbook corpus is one component of the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL Corpus), designed by Biber et al. (2002). The corpus is a collection of academic language which students are exposed to in American universities. It contains 2,7 million words of spoken and written American English, among which 760,619 are written texts taken from academic textbooks[5]. The corpus, originally compiled with the aim of suggesting design principles for the new TOEFL, also serves other purposes. Biber et al. (2004) investigated what they call lexical bundles in EAP classroom use and textbooks. The study reveals that classroom academic discourse and EAP textbook discourse display specific language features.

The second example of a textbook corpus is the German English as a Foreign Language Textbook Corpus (GEFL TC). It was compiled by Römer (2004a) and consists of texts taken from two series of general textbooks intended for German learners of English. The texts included in the corpus are supposed to represent spoken language (e.g. dialogues). GEFL TC consists of about 100,000 words[6]. In her two corpus-driven studies, Römer (2004a; 2006) compared what she calls 'school English' with authentic English by examining two aspects of grammar (modal auxiliaries and progressives). The results she obtained revealed striking discrepancies between the spoken language included in the textbooks and real spoken data.

As for Chujo's work (2004), it aimed to measure the gradations of vocabulary across levels in EGP and ESP textbooks The texts included in the textbooks were scanned, proofread and part-of-speech tagged. Lemmatised wordlists were then computed and compared to a lemmatised and frequency wordlist from the British National Corpus. Chujo also compiled specialised vocabulary wordlists from the coursebooks and from tests he had selected. Although he does not mention the term explicitly, Chujo (2004) uses a corpus of texts taken from textbooks and tests.

A fourth example of textbook corpus has been compiled by Anping (2005) and consists of over one million words of text taken from international EFL textbooks and from textbooks made in China for Chinese EFL learners[7]. Most parts of the corpus are lexically and semantically tagged. Anping's (2005) corpus-driven study consisted in finding out whether the design of EFL textbooks used in China reflected recent learning theories and teaching approaches.

Meunier & Gouverneur (2007) and Gouverneur (in press and forthcoming) are corpus-driven studies focusing on the treatment of phraseology in ELT textbooks. The data analysed come from a corpus of textbook material, the TeMa corpus (see section 3 for a detailed description) which contains over seven hundred thousand words. Before describing the features of the TeMa corpus, some terminological issues will be addressed. Two adjectives are usually used to refer to textbook corpora or 'corpora of coursebooks' as Gabrielatos calls them (2005: 5): 'pedagogic' and 'pedagogical'. According to dictionaries, they both mean relating to teaching methods or to the practice of teaching. In the literature however, the adjective 'pedagogic' is the preferred label and the expression 'pedagogic corpus' was coined by Willis (1993) and defined by Hunston (2002:16) as "a corpus consisting of all the language a learner has been exposed to. []. It can consist of all the coursebooks, readers etc a learner has used, plus any tapes etc they have heard." If one sticks to Hunston's definition, collecting a pedagogic corpus seems rather utopian. No learner, let alone his/her teachers, is in a position to provide an exhaustive list of all the language input he/she has been submitted to, be it inside or outside the classroom. Given what precedes, we suggest a more realistic definition of a pedagogic corpus as being a large enough and representative sample of the language, spoken and written,

[5] For a detailed description of the T2K-SWAL Corpus, see Biber et al. (2002: 19)

[6] For a detailed description of the GEFL TC, see Römer (2004a)

[7] For a detailed description of the corpus, see He Anping (2005)

a learner has been or is likely to be exposed to via teaching material, either in the classroom or during self-study activities. Typical teaching material includes texts, tapes and exercises. Taking into account the new definition provided, the examples of textbook corpora described above can be referred to as pedagogic corpora as they all consist of representative samples of textbook data intended for the teaching of EFL.

Hunston (2002:16) also suggests a number of possible exploitations of pedagogic corpora. First, a pedagogic corpus can be used for awareness-raising purposes by providing the learner with all the instances of a word or phrase he/she has encountered in various contexts (see for instance Biber et al. (2004) and their study of EAP vocabulary in textbooks). Secondly, the data included in a pedagogic corpus can be compared with a corpus of authentic English to check the authenticity of the language presented to the learners (as is the case for the two textbook analyses carried out by Römer in 2004a and 2006). As will be shown in sections 3 and 4, the TeMa corpus allows for a number of additional exploitations thanks to the annotation that has been inserted.

3. A new type of pedagogically annotated corpus for textbook research

The TeMa corpus has been collected in the framework of a research project on phraseology in language learning and teaching. A review of textbook studies convinced us that one way of facilitating an in-depth analysis of textbook material was have a computer readable version of the material. Once collected, the newly created pedagogic corpus could then be analysed with the help of typical corpus linguistics tools such as text retrieval software.

3.1. The textbooks in TeMa

The textbooks used for the compilation of the TeMa corpus were selected among recent[8] best sellers on the international ELT market, in similar proportion among the most renowned publishers. Thirty-two volumes of English for General Purposes (EGP) coursebooks were chosen for inclusion in TeMa. They include the student's book and workbook of the textbook series at the advanced and/or intermediate levels.

Table 2 presents the title, level, authors, date of publication and types of volumes available (i.e. Student's Book 'SB' and/or Workbook 'WB').

Table 2: Textbooks included in the TeMa corpus

Title	Level	Authors	Year	SB	WB	Publisher
Accelerate	Intermediate	Lodge, P. & B. Wright-Watson	1995	X	/	MacMillan Heinemann
	Advanced	Scott-Malden, S. & J. Wilson	1997	X	/	
Advance your English	Intermediate	/	/	/	/	CUP
	Advanced	Broadhead, A.	2003	X	X	
Clockwise	Intermediate	Jeffries, A.	2001	X	/	OUP
	Advanced	Forsyth, W.	2003	X	/	

[8] Only one textbook series was published in the 90s. The nine other series were published after 2000.

Cutting Edge	Intermediate	Cunningham, S. & P. Moor	2005	X	X	Longman
	Advanced	Cunningham, S. & P. Moor	2003	X	X	
English Panorama	Intermediate	/	/	/	/	CUP
	Advanced	O'Dell, F.	2003	X	/	
Initiative	Intermediate	/	/	/	/	CUP
	Advanced	Walton, R. & M. Bartram	2000	X	X	
Inside Out	Intermediate	Kay, S. & V. Jones	2000	X	X	MacMillan
	Advanced	Jones, C. & T. Bastow	2001	X	X	
Matters	Intermediate	Bell, J & R. Gower	2003	X	X	Longman
	Advanced	Bell, J & R. Gower	2001	X	X	
New Cambridge	Intermediate	Swan, M. & C. Walter	2003	X	X	CUP
	Advanced	Jones, L.	2002	X	/	
New Headway	Intermediate	Soars, L. & J.	2003	X	X	OUP
	Advanced	Soars, L. & J.	2003	X	X	

3.2. Corpus markup

The TeMa corpus went first through a markup stage in order to identify each section of the corpus. Figure 1 illustrates the incremental mark-up stage of the corpus. A first subdivision is based on the textbook series, the levels, and the type of book (i.e. SB or WB). Each textbook series is first given a code number. New Headway, for instance, has been assigned number *6*. An extra digit is then added to represent the levels (1 for advanced and 2 for intermediate). New Headway advanced is thus 6*1* and New Headway intermediate 6*2*. Each level is then further divided into student's book (1) and workbook (2).

The last subdivision represents the types of input provided: the texts (1), the transcription of the tapescripts (2), the vocabulary exercises (3) and the guidelines to these exercises (4). Each coursebook series is divided in sixteen potential subcorpora, each identified by a 4-digit markup.

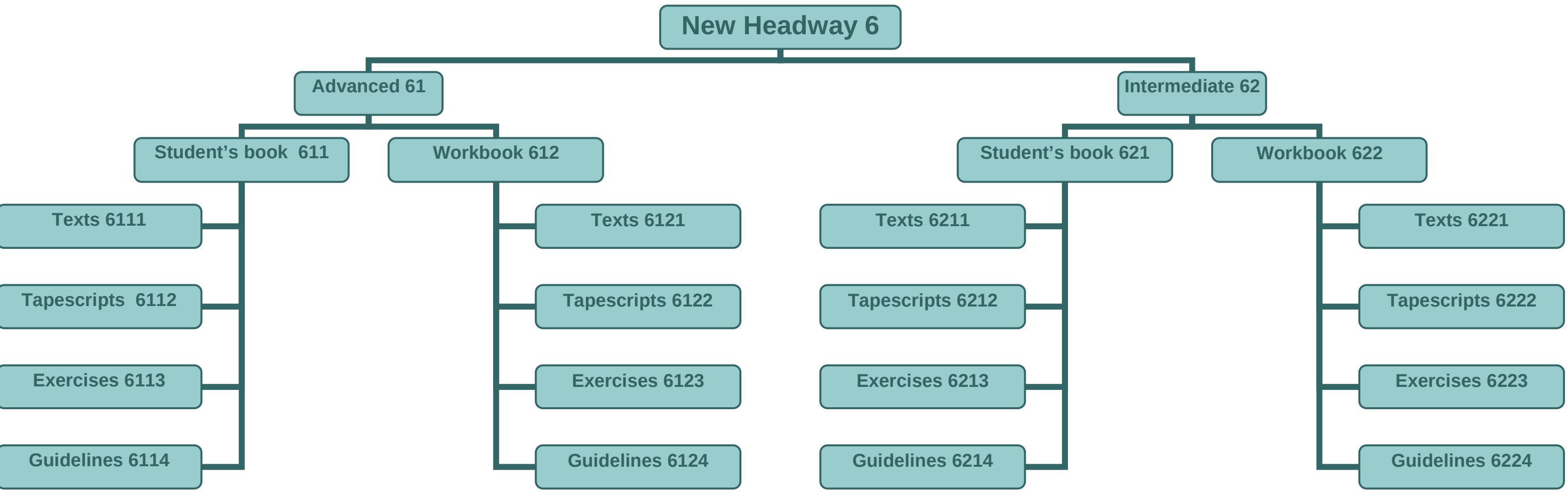
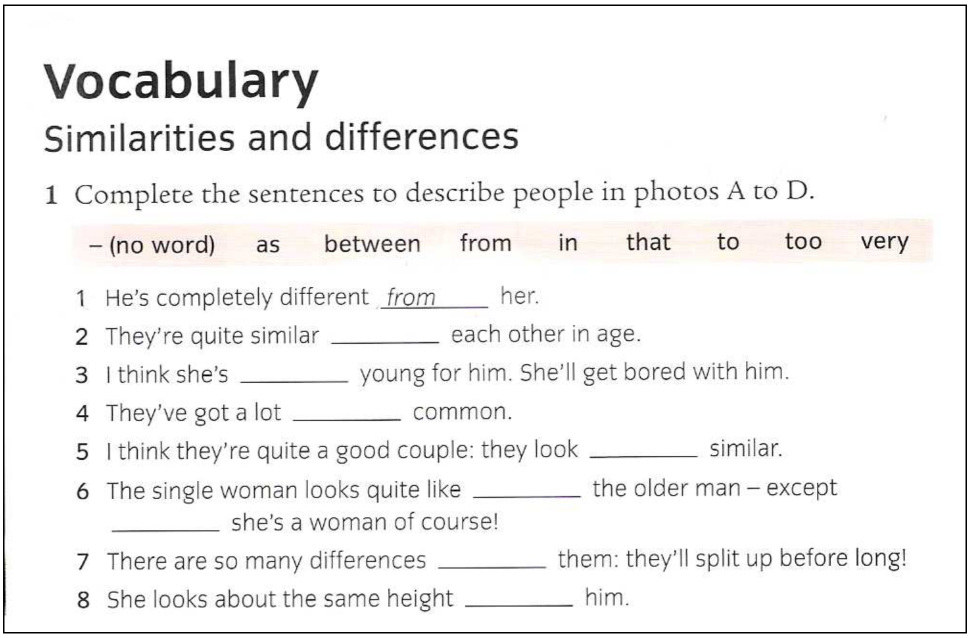


Figure 1: Markup stage of the TeMa corpus

The TeMa corpus is innovative in a number of ways. First, it is rather large. With over 700,000 words of textbook material, it is one of the largest pedagogic corpora available. A second aspect is the richness of the pedagogic input, i.e. not only texts (as was the case for T2K-SWAL and for Anping's corpus), not only spoken data (as was the case for GEFL TC) but both, plus vocabulary exercises and guidelines to the exercises. Thirdly, the language collected in the TeMa corpus comes from international EGP textbooks intended for any learners of English as a Foreign Language, with no mother tongue background restriction. The type of target audience of the textbooks collected in the four other pedagogic corpora was more specific: T2K-SWAL Corpus only contains EAP books for American students, GEFL TC is based on textbooks designed specially for German EFL learners, and a large part of Anping's corpus consists of texts taken from EFL textbooks made in China for Chinese learners.

3.3. Pedagogical tagging

Another aspect that singles out TeMa from other types of textbook corpora is the pedagogical tagging that has been applied to the vocabulary subcorpus. Corpus annotation is very common in corpus linguistics and many corpora can be part-of-speech tagged, syntactically parsed, or even error-tagged (as is the case for many learner corpora). The type of tagging that has been used for TeMa belongs to what is commonly called 'problem-oriented tagging' and which is defined by de Haan (1984: 125) as the phenomenon whereby users will take a corpus and add to it their own form of annotation, oriented particularly towards their own research goal. The subcorpus of vocabulary exercises in TeMa has been tagged according to the pedagogical tasks the learners have to perform when doing the exercises. This pedagogical tagging has been applied to all the vocabulary exercises on the basis of both the learning activities learners have to engage in (e.g. match words and their definitions, choose an item from a multiple-choice option, etc.) and on the pedagogical status of the lexical items in the exercises (e.g. a pre-selected list of words in a box). The following excerpt from the student's book of *Clockwise Intermediate* (Forsyth 2003) illustrates how vocabulary exercises have been coded. Figure 2 is a scan of the actual page of the book, whilst Figure 3 presents its annotated corpus version.



Vocabulary

Similarities and differences

1 Complete the sentences to describe people in photos A to D.

– (no word) as between from in that to too very

1 He's completely different *from* her.
2 They're quite similar ________ each other in age.
3 I think she's ________ young for him. She'll get bored with him.
4 They've got a lot ________ common.
5 I think they're quite a good couple: they look ________ similar.
6 The single woman looks quite like ________ the older man – except ________ she's a woman of course!
7 There are so many differences ________ them: they'll split up before long!
8 She looks about the same height ________ him.

Figure 2: Vocabulary exercise as it appears in the textbook

```
<CLISB-U6-P24-E1>

1213(BC)–#$
1213(BC)as#$
1213(BC)between#$
1213(BC)from#$
1213(BC)in#$
1213(BC)that #$
1213(BC)to#$
1213(BC)too #$
1213(BC)very#$


1213(CB)He's completely different 1213(AB)from# her$
1213(CB)They're quite similar 1213(AB)to# each other in age$
1213(CB)I think she's 1213(AB)too# young for him. She'll get bored with him$
1213(CB)They've got a lot 1213(AB)in# common$
1213(CB)I think they're quite a good couple : they look 1213(AB)very# similar$
1213(CB)the single woman looks quite like 1213(AB)no word# the older man-except
1213(AB)that# she's a woman of course !$
1213(CB)there are so many differences 1213(AB)between# them ; they'll split up before
long !$
1213(CB)She looks about the same height 1213(AB)as# him$
```

Figure 3: Pedagogically annotated vocabulary exercise as it appears in the corpus

As can be seen from Figure 3, each exercise is given a unique reference. The example presented in Figure 3, i.e. <CLISB-U6-P24-E1>, is taken from **CL**ockwise **I**ntermediate **S**tudent's **B**ook **–** **U**nit **6** **–** **P**age **24** **–** **E**xercise **1**. The four-digit tag before each word or sentence (1213 in this case) refers to the origin of the exercise (see Figure 1 for an illustration). The two-letter tags between brackets (BC) indicate the pedagogical status of the lexical items presented. BC is used when words are presented in a box (hence B) from which items are to be selected to complete (hence C) the sentences in the exercise. As for the introductory tag in front of each exercise line, it gives information on the pedagogical task that has to be performed. CB in Figure 3 means "complete the sentence with words from a box". The tags within the exercise sentences (e.g. AB) refer to the status of the lexical items within the exercise sentences: "from", "to", "too" are all preceded by AB as they are all answers taken from a box, hence (AB). Each sentence ends with a dollar sign ($) and within the sentences, the answers are followed by a hash (#). These additional signs make it easier to spot the beginning and the end of sentences as well as to extract and contextualise the exact lexical items being practised in the exercises. To tag our corpus pedagogically, a list of over 80 tags has been drawn. Seven main types of pedagogical tasks were identified during the compilation of the vocabulary exercises subcorpus: complete, define, match, replace, understand, correct, (re)write. Each main type was in turn divided into subcategories. Two tasks will be illustrated and explained hereafter: complete and match[9].

3.3.1. The *COMPLETE* tag

The "complete" category, represented by the generic capital letter C*, can be subdivided in seven more specific tasks.

(C) alone is used when learners have to complete a sentence, fill in a blank in a sentence without any prompt (such as a pre-selected list of words, multiple choice options, first letter of the words, etc). Learners have to retrieve the words or expressions from their mental lexicon on the sole basis of the context provided.

2113 (C) Can I get past please? oh I'm sorry, are my bags 2113 (A) *in the way*#? I'll put them up in the locker.$

The answer, which has to be provided by the learner, is preceded by the status tag **(A)**.

(CB) stands for "complete from a box" and refers to tasks where learners have to complete sentences choosing from a pre-selected set of words provided in a box. The lexical items which are presented in the box, and which should be used to complete the sentence, are given the tag **(BC)** for "box to complete". The answers get the AB tag **(**"answer from a box"). Here follows an example of (CB):

5113(BC)dubbed#$
5113(BC)subtitles#$

5113(CB)Foreign-language films can be shown with 5113(AB)*subtitles*# or they may be 5113(AB)*dubbed*#.$

In **(CE)**, "complete-exercise", learners have to complete sentences using lexical items that have to be chosen from an exercise done previously. In such cases, the answers are referred to as **(AE)** (answer from exercise), such as in:

[9] The complete list of tags will be available in Gouverneur (forthcoming b)

3213 (CE) All I did was ask him to smile for the camera and he 3213 (AE) gave me a black eye#.$

3213 (CE) His aggressive behaviour is unacceptable. He should be 3213 (AE)*charged with assault*#.$

The expressions *give a black eye* and *to be charged with assault* had been practised in an exercise done previously.

The **(CT)** tag, “complete-text”, is used in front of sentences which should be completed with lexical items taken from a text seen previously in the unit, as in:

0213(CT)Products with too much 0213(AT)*packaging*# use a lot of energy to produce and distribute.$

0213(CT)Find out if there is anything harmful in a product by writing to the 0213(AT)*manufacturer*#.$

The words *packaging* and *manufacturer* are words included in a text read previously in the unit. The answers are preceded by **(AT),** which stands for “answer from a text”.

The **(CZ)** tag was chosen to refer to sentences which learners have to complete with a lexical item they have to pick from a multiple choice included in the sentence. The options provided are referred to with the tag **(BZ)** and the correct answer is given the tag **(AZ)**, as illustrated below:

2113(CZ)A person who resembles a famous person can be called 2113(BZ)a lookalike/ a lookout/ an onlooker# 2113(AZ)*a lookalike*#.$

2113(CZ)The proverb “Look before you 2113(BZ)jump/ leap/ strike# 2113(AZ)*leap*# means you should think about the possible dangers before you do something.$

(CZX) tasks are similar to the previous ones but instead of choosing the correct answer from a list of possible options, learners have to cross out the wrong answer.

2213(CZX)remind 2213(BZX)someone to do something/ someone about an appointment/ someone of another person/ to phone someone# 2213(AZX)~~to phone someone~~#$

2213(CZX)forget 2213(BZX)to do something/ someone’s birthday/ of something/ about something# 2213(AZX)~~of something~~#$

The tag **(CW)** is used when learners have to provide a morphologically derived form for a given word, for instance the adjective derived from a particular noun, as in:

2213(CW)industry# 2213(AW)*industrial*#$
2213(CW)history# 2213(AW)*historical*#$
2213(CW)crowd# 2213(AW)*crowded*#$

The tag to refer to the answer in those cases is **(AW)**

3.3.2. The *MATCH* tag

Matching exercises are extremely common in textbooks. To annotate such exercises, the tags are placed in front of the lexical items to be matched. Given the fact that very often, the two parts of a matching exercise have the same “weight”, the left- and right-hand parts of the exercises were arbitrarily assigned the (MQ) for “match-question” and (MA) for “match-answer” tags respectively.

6213(MQ)strong# 6213(MA)coffee#$
6213(MQ)full-time# 6213(MA)job#$
6213(MQ)film# 6213(MA)star#$

9113(MQ)I don’t believe a word of it!# 9113(MA)I don’t believe it at all#$
9113(MQ)To eat your words# 9113(MA)To admit being wrong#$
9113(MQ)By word of mouth# 9113(MA)By speaking and not by writing#$

As illustrated in the examples above, the two elements to be matched can be the two parts of multi-word units (collocations or compound nouns for instance) or paraphrases and synonyms. It must be noted that in the case of synonyms, paraphrases and meanings, the exercises have not been classified in the ‘definition’ category as the primary pedagogical technique underlying the task is to make the connections (matching) between the two elements and not to define a given item.

Here again, both parts of the matching exercise can come from various types of input : a text **(MQT/MAT)**, a previous exercise **(MQE/MAE)** or a box **(MQB/MAB)**. Some of the possible options are presented in the following box.

9113(MQT)date back# 9113(MA)*be invented in*#$
9113(MQT)turn into# 9113(MA)*change in form or nature*#$

9113(MQE)ground plan# 9113(MA)*a drawn plan of a building at ground level 1*#$
9113(MQE)main artery# 9113(MA)*big or principal road*#$

6123(BMA)for#$
6123(BMA)from#$

6123(MQ)The company isn’t liable# 6123(MAB)*for*# 6123(MA)any damages caused to vehicles parked on the premises#.$
6123(MQ)Bill is emotionally detached# 6123(MAB)*from*# 6123(MA)his parents. He hardly ever speaks to them#.$

Answers may also be chosen from a multiple choice included in the sentence. The assigned tags are then **(MQZ)** for the question part, **(BZM)** for the box to choose from, and **(MAZ)** for the answer, as illustrated below:

1213(MQZ)my brother is married# 1213(BZM)a woman called Jenny/ to a woman called

Jenny# 1213(MAZ)*to a woman called Jenny*#$
1213(MQZ)my brother married1213(BZM)a woman called Jenny/ to a woman called Jenny#
1213(MAZ)*a woman called Jenny*#$

1213(MQZ)he met# 1213(BZM)her at a party/ to her at a party# 1213(MAZ)her at a party#$
1213(MQZ)he was introduced# 1213(BZM)her at a party/ to her at a party# 1213(MAZ)*to her at a party*#$
1213(MQZ)he was fascinated# 1213(BZM)by her/ in her# 1213(MAZ)*by her*#$
1213(MQZ)he was very interested# 1213(BZM)by her/ in her# 1213(MAZ) *in her*#$

Given the size of the pedagogical tagset created for the annotation of the corpus (about 80 in total), it must be acknowledged that the tagging stage was extremely time-consuming. The compilation of the vocabulary exercises required careful selection and analysis. Suggestions of tags were inserted in the paper copy of the textbook for each exercise selected and, only then, was the exercise ready for compilation. Our progression in analysing the exercises often forced us to come back to previously annotated exercises whenever we discovered subtleties of tasks that had not been encoded, which inevitably led to numerous checks, revisions and adaptations. However, once the annotation stage is completed, the corpus offers numerous paths for exploitations, as will be shown in section 4.

4. Meeting new pedagogical challenges

We will now examine in what way the collection and annotation of a textbook corpus can help meet the new pedagogical challenges mentioned in the title of the paper. A pedagogically annotated corpus makes it possible to explore the data from a variety of perspectives never addressed before or addressed on a much smaller scale given the manual analysis involved. On a descriptive level, using a pedagogically tagged textbook corpus makes it possible to provide a solid empirical description of the material under analysis. A comparison, of vocabulary selection across levels can be carried out and it becomes possible to determine what a specific level actually means in terms of vocabulary selection (e.g. by providing a list of all the words/expressions practised in the exercises of a level and by comparing it with a list of words form a lower/higher level). The relationship between the input provided in the texts and audio files and the words/expressions practised in the exercises can also be addressed. The various types of cognitive tasks that learners have to perform when doing the exercises can be analysed. The results of such studies can help raise the publishers' or textbook writers' awareness of the types of exercises they propose, in what proportion and to what target audience. Gouverneur (in press), in a pilot study on the aforementioned questions, reports on some preliminary results. A comparison of the collocations presented in the exercises reveals a total lack of consistency between the textbooks examined, i.e. very few were common to all textbooks. As to the weight of pedagogical tasks, the study shows that some tasks are common to all levels of proficiency, such as 'complete' tasks, for instance, whilst some others are more specific to one level, such as the 'replace' tasks which are very frequent in the advanced textbooks but can hardly be found at the intermediate level. The study also states that not enough tasks promote cognitive processes such as noticing or receptive and productive retrieval. Cognitively oriented SLA research has demonstrated the importance of noticing, extrapolation and rehearsal (see among others De Bot et al. 2005). A detailed analysis of a pedagogically tagged textbook corpus like TeMa helps researchers specify where and when exactly noticing, extrapolation and rehearsal of lexical items are practised in the textbook and, subsequently, propose possible improvements to current practice. One such improvement could be the addition of extra electronic input to the textbook. As is the case with learners' dictionaries, which now almost invariably include a CD-Rom version containing extra material such as concordance lines, thesaurus, extra examples, or exercises, it would be reasonable to expect a similar evolution for textbooks. The accompanying CD-Rom would not only include the transcript of the audio files and texts included in the paper version of the textbook but also more texts, more exercises and more authentic native corpus input. Such add-ons, combined with user-friendly search options, would make it possible, for teachers and learners alike, to access more contextualised instances of words and expressions and to practise independent or teacher-led data-driven learning activities.

Another issue worth addressing with the help of a textbook corpus such as TeMa is the metalanguage used in the textbooks to refer to vocabulary and phraseology. Analysing the TeMa subcorpora containing the guidelines to the exercises helps us identify (1) the type of metalanguage used by material designers (general terms such as words or expressions or specific terms such as collocations or idioms?); (2) the consistency in their terminology (does the term idiom refer to fixed idioms, to conversational routines, to pragmatic phrasemes or to other types of multiword units?). A pilot study carried out on a small proportion of the corpus (Meunier and Gouverneur 2007) has shown that the metalanguage used in textbooks, and more particularly in the guidelines to the exercises, is still far too general and indirect. Textbook designers tend for instance to make use of terms such as 'words' or 'expressions' instead of using more specific terms which, as some argue (Lewis 2001), are not more difficult to remember and understand. The use of a (limited) set of specific and pedagogically-oriented terms might even facilitate the understanding of important concepts.

Although the results reported here only deal with a limited number of research questions, the richness of the TeMa corpus allows for far more exploitations; the focus could be on one type of linguistic features (e.g. high frequency verbs); spoken and written input could also be compared in order to investigate language mode variations; and along the lines of what has already been done by Römer (2004a; 2004b) an analysis of the authenticity of textbook material could be carried out on larger corpora of textbook material. It should also be added that the pedagogical tagging of the TeMa corpus which was heavily oriented towards lexis could easily be extended to other aspects such as grammar exercises and metalanguage, or speech-act analysis.

5. Conclusion

The previous sections have introduced what we believe to be a rather innovative type of corpus, namely a pedagogically annotated corpus of textbook material. Its description, collection and annotation procedures have been outlined, and preliminary results of exploitations have also been provided.

We are aware that the methodology adopted in our research project has its limitations: the corpus collection and annotation is limited in size (focus on texts, audio transcripts, lexically-oriented exercises and guidelines to those exercises); no possible feedback on open questions types of exercises can be accessed (be it teacher-learner or peer interaction); no feedback on teacher input is available; and no general format-like information (pictures, maps, etc.) has been encoded.

Despite these limitations however, we believe that the collection, annotation and exploitation of this new type of corpus makes it possible to access empirical evidence otherwise inaccessible. This type of empirical evidence includes access to frequency patterns of use not only in the input provided to the learners but also in the types of exercises suggested; access to the actual connections (made or not) between the input and its exploitation in the exercises; access to the metalanguage used to introduce formal aspects of language.

Access to such type of information helps foster a reflexive approach to textbook editing and provide evidence-based guidelines to improving textbooks. Analyses such as those presented in section 4 also helped us reveal what is good about textbooks. We have for instance demonstrated the growing awareness of the phraseological nature of the language and the presence of recycling and rehearsal exercises. As for potential areas of improvement, it should be made clear that they are not restricted to promoting the authenticity of textbook material. Although we believe it essential to offer learners input which is as authentic as possible, adapted or simplified input also has its relevance, especially perhaps at lower levels of proficiency. Other domains that could benefit from the analysis of pedagogically tagged corpora include: a reconsideration of the links between important issues revealed in the SLA literature and their possible inclusion and exploitation in textbooks; an improved awareness (on the part of the teachers) of what is contained in the material they use; and a possible revision of the grammatical and lexical metalanguage present in textbooks. We also believe that, thanks to a corpus approach[10], the inevitable initial limitations of the paper and ink format could disappear.

We can only but hope that other types of pedagogically annotated corpora addressing similar or different issues as the ones presented here will soon be created and exploited.

References

Anping, He (2005) Corpus-based Evaluation of ELT textbooks. Paper presented at the joint conference of the American Association of Applied Corpus Linguistics and the International Computer Archive of Modern and Medieval English, 12-15 May 2005, University of Michigan.

Biber, D., S. Conrad, R. Reppen, P. Byrd & M. Helt (2002) Speaking and Writing in the University: A Multidimensional Comparison. TESOL Quarterly 36 (1): 9-48.

Biber, D., S. Conrad & V. Cortes (2004) If you look at …: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics 25 (3):371-405.

Botley, S., McEnery, A. M. & Wilson, A. (eds) (2000) Multilingual Corpora in Teaching and Research. Amsterdam, Rodopi.

Boxer, D., & Pickering, L. (1995). Problems in the presentation of speech acts in ELT materials: The case of complaints. ELT Journal, *49*, 44-58.

Burnard, L. & McEnery A. (eds.) (2000). Rethinking Language Pedagogy from a Corpus Perspective. Papers from the Third International Conference on Teaching and Language Corpora. Frankfurt/M.: Peter Lang.

Cane, G. (1998). Teaching Conversation Skills More Effectively. The Korea TESOL Journal, *1*, 31-37.

Chambers, F. (1997) Seeking Consensus in Coursebook Evaluation. ELT Journal 51 (1): 29-35.

[10] The ‘corpus approach’ mentioned is seen as: including much larger quantities of similar input for learners, providing opportunities for data-driven learning, providing access to corpus searches and words in context, etc.

Chujo, K. (2004) Measuring Vocabulary Levels of English Textbooks and Tests. Using a BNC Lemmatised High Frequency Word List. In Nakamura, J., N. Inoue & T. Tabata (eds) English Corpora under Japanese Eyes. Amsterdam/ New York, 231-249.

Connor, U. & Upton, T. (2004) (eds.) Applied Corpus Linguistics: A Multidimensional Perspective. Amsterdam: Rodopi.

Cunningsworth, A. (1984) Evaluating and Selecting EFL Teaching Materials. Longman: Heinemann.

Cunningsworth, A. (1995) Choosing Your Coursebook. Oxford: Heinemann.

De Bot, K., Lowie, W. and Verspoor, M. (2005) Second Language Acquisition. An Advanced Resource Book. Routledge.

De Haan, P. (1984) Problem-oriented Tagging of English Corpus Data. In Aarts, J. & W. Meijs (eds.) Corpus Linguistics. 123-139. Amsterdam, Rodopi.

Gabrielatos, C. (1994) Collocations. Pedagogical Implications and Their Treatment in Pedagogical Materials. Unpublished essay, research centre for english and applied linguistics, University of Cambridge. Available at http://www.gabrielatos.com/Collocation.htm

Gilmore, A. (2004) A Comparison of Textbook and Authentic Interactions. ELT Journal 58 (4): 363-374.

Gabrielatos, C. (2006) Corpus-based Evaluation of Pedagogical Materials If-conditionals in ELT Coursebooks and the BNC. Paper presented at the 7th Teaching and Language Corpora (TALC) Conference, 1-4 July 2006, Paris, Université Paris 7 Diderot.

Gouverneur, C. (in press) The Phraseological Patterns of High-frequency Verbs in Advanced English for General Purposes: a corpus-driven approach to EFL textbook analysis. In Meunier F. and Granger S. (eds.) Phraseology in Foreign Language Learning and Teaching. Amsterdam/Philadelphia. Benjamins.

Gouverneur, C. (forthcoming) Phraseology in Foreign Language Learning and Teaching: a corpus-based study of EFL textbooks. PhD Dissertation. Université Catholique de Louvain, Belgium.

Granger, S., Hung, J. & Petch-Tyson, S. (2002) (eds.) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam: John Benjamins.

Harwood, N. (2005) What Do We Want EAP Teaching Materials For? In Journal of English for Academic Purposes 4 (2):149-161.

Hill, V. J. (1996) Verb-form Clustering and Syllabus Design. System 24 (4): 529-536.

Hunston, S. (2002) Corpora in Applied Linguistics. Cambridge: Cambrdige University Press.

Hyland, K. (1994) Hedging in Academic Writing and EAP Textbooks. English for Specific Purposes 13: 251-281.

Jacobs, G.M., & Ball, J. (1996). An Investigation of the Structure of Group Activities in ELT Coursebooks. ELT Journal, 50 : 99-107.

Koprowski M. (2005) Investigating the Usefulness of Lexical Phrases in Contemporary Coursebooks. ELT Journal 59(4): 322-332.

Merlevede, J. (2006) Corpus Linguistics Implications in English Language Teaching. An Analysis of Corpus-based Pedagogical Materials. MA dissertation, Université catholique de Louvain, Louvain-la-Neuve.

Meunier, F. & Gouverneur C. (2007) The treatment of Phraseology in ELT Textbooks. In E.Hidalgo, L.Querada and J.Santana (eds) Corpora in the Foreign Language Classroom. Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC), University of Granada, Spain, 4-7 July, 2004. Language and Computers series. Amsterdam/Atlanta: Rodopi, 119-139.

Miura, T. (1997) An analysis of “Aural/Oral Communications A”. English textbooks in Japanese Upper Secondary School. MA Dissertation, The University of Birmingham.

Moreno, A. I. (2003) Matching Theoretical Descriptions of Discourse and Practical Applications to Teaching: the Case of Causal Metatext. English for Specific Purposes 22: 265-295.

Mukherjee, J. & Rohrbach, J.M. (2006) Rethinking Applied Corpus Linguistics from a Language-pedagogical Perspective: New Departures in Learner Corpus Research. In Kettemann, B. & Marko, G: (eds) Planning, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop. Frankfurt am Main: Peter Lang, pp. 205-232.

Nitta, R. & S. Gardner (2005) Consciousness-raising and Practice in ELT Coursebooks. ELT Journal 59 (1): 3-13.

O'Keeffe, A., McCarthy, M. & Carter, R. (2007) From Corpus to Classroom: Language Use and Language Teaching. Cambridge: Cambridge University Press.

O'Neill , R. (1993) Are Textbooks Symptoms of Disease? Practical English Teaching 14 (1): 11-13.

Paltridge, B. (2002) Thesis and Dissertation Writing: an Examination of Published Advice and Actual Practice. English for Specific Purposes 21: 125-143.

Ranalli, J.M. (2003) The Treatment of Key Vocabulary Strategies in Current ELT Coursebooks: repetition, resource use, recording. Unpublished MA dissertation. Centre for English Language Studies, University of Birmingham.

Reda G. (2003) English Coursebooks: Prototype Texts and Basic Vocabulary Norms. ELT Journal 57(3): 260-268.

Römer, U. (2004a) Textbooks: a Corpus-driven Approach to Modal Auxiliaries and their Didactics. In Sinclair, J. (ed.) How to Use Corpora in Language Teaching. Amsterdam/Philadelphia: Benjamins.

Römer, U. (2004b) Comparing Real and Ideal Language Learner Input: The Use of an EFL Textbook Corpus in Corpus Linguistics and Language Teaching. In: Aston, Guy, Silvia Bernardini and Dominic Stewart (eds.). Corpora and Language Learners. Amsterdam: John Benjamins. 151-168.

Römer, U. (2006). Looking at *Looking*: Functions and Contexts of Progressives in Spoken English and 'School' English. In: Renouf, Antoinette & Andrew Kehoe (eds.). The Changing Face of Corpus Linguistics. Papers from the 24th International Conference on English Language Research on Computerized Corpora (ICAME 24). Amsterdam: Rodopi. 231-242.

Sheldon, L. (1988) Evaluating ELT Textbooks and Materials. ELT Journal 42(4): 237-246.

Sinclair, J. (ed.) (2004) How to Use Corpora in Language Teaching. Studies in Corpus Linguistics 12. Amsterdam/Philadelphia, Benjamins.

Swales, J. M. (1995) The role of the Textbook in EAP Writing Research. English for Specific Purposes 14(1): 3-18.

Swales, J. M. (2002) Integrated and Fragmented Worlds: EAP Materials and Corpus Linguistics. In Flowerdew, J. (ed.) Academic Discourse. Harlow: Longman, 150-164.

Vellenga, H. (2004) Learning Pragmatics from ESL & EFL Textbooks: How Likely? TESL-EJ, Vol.8, nr 2. Available from http://www.kyoto-su.ac.jp/information/tesl-ej/ej30/a3.html

Widdowson, H. (2003) Defining Issues in English Language Teaching. Oxford: Oxford University press.

Williams, D. (1983) Developing criteria for textbook evaluation. ELT Journal 37 (3): 251-255.

Textbooks

Bell, J & Gower, R. (2001a) Matters. Student's Book Advanced. London: Longman.

Bell, J & Gower, R. (2001b) Matters. Workbook Advanced. London: Longman.

Bell, J & Gower, R. (2003a) Matters. Student's Book Intermediate. London: Longman.

Bell, J & Gower, R. (2003b) Matters. Workbook Intermediate. London: Longman.

Broadhead, A. (2003a) Advance your English. Coursebook. Cambridge: Cambridge University Press.

Broadhead, A. (2003b) Advance your English. Workbook. Cambridge: Cambridge University Press.

Cunningham, S. & Moor, P. (2005a) Cutting Edge. Intermediate Student's book. London: Longman.

Cunningham, S. & Moor, P. (2005b) Cutting Edge. Intermediate Workbook. London: Longman.

Cunningham, S. & Moor, P. (2003a) Cutting Edge. Advanced Student's book. London: Longman.

Cunningham, S. & Moor, P. (2003b) Cutting Edge. Advanced Workbook. London: Longman.

Forsyth, W. (2003) Clockwise. Advanced Classbook. Oxford: Oxford Uninversity Press.

Jeffries, A. (2001) Clockwise. Intermediate Classbook. Oxford: Oxford Uninversity Press.

Jones, C. (2001) Inside Out. Workbook Advanced. London: MacMillan.

Jones, C. & T. Bastow (2001) Inside Out. Student's book Advanced. London: MacMillan.

Jones, L. (2002) New Cambridge Advanced English. Student's Book. Cambridge: Cambridge University Press.

Kay, S. & V. Jones (2000a) Inside Out. Student's Book Intermediate. London: MacMillan.

Kay, S. & V. Jones (2000b) Inside Out. Workbook Intermediate. London: MacMillan.

O'Dell, F. (2003a) English Panorama. Advanced Student's book1. Cambridge: Cambridge University Press.

O'Dell, F. (2003b) English Panorama. Advanced Student's book 2. Cambridge: Cambridge University Press.

Scott-Malden, S. & Wilson, J. (1997) Accelerate. Advanced Student's book. MacMillan Heinemann.

Soars, L. & J. (2003a) New Headway. Intermediate Student's Book. Oxford: Oxford Uninversity Press.

Soars, L. & J. (2003b) New Headway. Intermediate Workbook. Oxford: Oxford Uninversity Press.

Soars, L. & J. (2003c) New Headway. Advanced Student's Book. Oxford: Oxford Uninversity Press.

Soars, L. & J. (2003d) New Headway. Advanced Workbook. Oxford: Oxford Uninversity Press.

Swan, M. & Walter, C. (2003a) The New Cambridge English Course. Student. Intermediate Cambridge: Cambridge University Press.

Swan, M. & Walter, C. (2003b) The New Cambridge English Course. Practice. Intermediate. Cambridge: Cambridge University Press.

Walton, R. & Bartram, M. (2000) Initiative. Student's book. Cambridge: Cambridge University Press.