

TOCFL 作文語料庫的建置與應用*

張莉萍

台灣師範大學國語教學中心

提要

本文介紹台灣師範大學所新建置的學習者語料庫，內容取自以 CEFR 為架構的電腦華語寫作能力考試的作文，語料仍在增加中，現階段文本涵蓋來自 39 種不同母語背景的學習者，4,567 篇作文，71 個不同主題，約 150 萬字。語料經過前處理、自動斷詞與詞性標注外，半數語料並經由人工完成偏誤標記。初步以中介語對比分析方法，觀察分析學習者詞語、語法表現，目的是找出能區辨不同能力學習者的關鍵語言特徵，以提供教學者、測驗評量者對應 CEFR 能力等級的具體漢語語言描述。

關鍵詞：中介語、學習者語料庫、中介語對比分析、關鍵語言特徵

一、前言

近年來利用漢語中介語語料庫所做的分析研究不少，例如，張寶林（2010），肖奚強、張旺熹（2011），對於漢語習得、教學做出不少貢獻。然而大部分的研究分析限於語料庫的屬性，多是針對高等程度學習者產出的語料所做的觀察，屬於橫向的分析，較少看到縱向分析，也就是觀察不同能力學習者對於目標語言（漢語）的掌握。

本計劃刻正進行的語料庫研究正是為服務這樣的目的而建置的。期望從一個分級完善的學習者語料庫 (learner corpus)¹，找出不同階段學習者產出的語言特徵，提供教學者與測驗評量者具體的內容。一般學習者語料分級的方式，可以依據學習者程度分為初、中、高三級或是依據學習者在學時間分為大學一年級、二年級的語料等等，但這樣的分級很難達到結果可比的目标，因為初中高三級語言能力的界線模糊，不同使用者定義不尽相同；不同學校或學制的學生能力表現也不同，如果不在一開始語料的分級上統一標準，那麼不同語料庫觀察所得的結果可信度堪慮。因此，這個語料庫採用《歐洲語言共同參考框架：學習、教學、評估》（劉駿、傅榮主譯，2008，以下簡稱 CEFR）對語言能力的描述，作為語料分級的標準。就語言教學或測試的角度來看，如果有一個共同的框架為平台，那麼不同教學系統或測試所報告的分數或等級就可以在一共同標準上互相參照，近年來，不同类型的外語測試已經紛紛建立了與 CEFR 的對應關係，來減輕學習者、教學機構對於不同測試分數或成績理解解釋上的負擔。

基於此原因，本計劃選擇了華語文能力測驗 (Test of Chinese as a Foreign Language, 以下簡稱 TOCFL) 中的電腦寫作考試所產出的語料做為建置的來源。這個寫作測驗是以

*本研究得到國科會計劃(NSC 99-2631-S-003-012, 100-2631-S-003-004)以及教育部邁向頂尖大學計劃部分經費補助，特此感謝。

¹ 個人比較傾向使用「學習者語料庫」這個名稱，它與中介語語料庫意思相當。

CEFR 的三等六级 (A1/A2; B1/B2; C1/C2) 做为设计架构，题型内容以完成交际任务为导向，考生在一定的时间内以电脑输入完成写作内容²，语料样本涵盖基础级 (A2) 到流利级 (C1) 程度³，目前共搜集约 150 万字语料。本文分为两大部分，一是介绍 TOCFL 语料库的建置工作与内容；二是基于此语料库所观察得到的一些初步分析。

二、TOCFL 语料库的建置与内容

本计划最终的目标是希望寻找出可以区辨不同能力学习者的语言特征，反馈给教学单位与考试单位，以进一步修订词汇大纲、句法大纲。为了能观察不同能力学习者语言，并且确定这些学习者的能力等级能对应到 CEFR 的能力描述，选择了 TOCFL 电脑写作测验所产出的语料作为建置的基础。

TOCFL 测验采用分等分级的考试方式，运用 CEFR 的能力指标来命题，现阶段研发了 A2-C1 等级的考试。参加某一等级考试的考生需要完成两个不同文体的作文：A2 是完成实用性的便条和看图说故事的文体；B1 是书信和一般记叙文；B2 则是应用文（特定功能的书信）和论说文（表达对特定事件的看法）；C1 是报告类（图表说明）和论说文（提出理由支撑观点）。每篇字数规范从 70 至 800 不等。该测验的评分方式为级分制，分为 0-5 级分，3 分为通过门槛，每份作文都经过两位以上阅卷老师评分。

2.1 语料库内容属性

语料库搜集自 2006 年以来至今（2012 年 5 月）TOCFL 写作文本，语料数量仍在持续增加中。现阶段语料库文本涵盖来自 39 种不同母语背景的学习者，4,567 篇作文，71 个不同主题，总计 1,561,942 字，927,051 词（不计标点符号）。搜集的语料包括 A2、B1、B2、C1 各等级，每一级语料篇数及字数统计请见表 1。以字数分布而言，B1 程度的语料占全部语料的近一半，约 74 万字，其次是 B2、A2 和 C1 等级。以作者母语分布来看，母语为日语的学习者最多，占全部学习者的百分之二十四，其次是英语、韩语等，表 2 罗列前十名学习者母语信息。

表 1 TOCFL 语料库语料分布一览

级别	主题数	篇数	词数	字数	百分比
A2	23	1,366	120,754	203,217	13.01%
B1	26	1,961	445,997	737,018	47.19%
B2	17	1,033	306,303	522,848	33.47%

² 关于此测验的题型范例请参考 <http://www.sc-top.org.tw/chinese/WT/test1.php> 网站。

³ 为利于本文阅读，将 CEFR 与 ACTFL、新 HSK、TOCFL 等级的对应关系整理如下表：

CEFR	ACTFL	新 HSK	TOCFL
C2	Distinguished	6 级	精通级
C1	Superior	5 级	流利级
B2	Advanced-mid	4 级	高阶级
B1	Intermediate-high	3 级	进阶级
A2	Novice-high,	2 级	基础级
A1	Novice-mid, Novice-low	1 级	入门级

C1	5	207	53,997	98,859	6.33%
总计	71	4,567	927,051	1,561,942	100.00%

表 2 TOCFL 语料库学习者母语分布

	母语	篇数	字数	百分比
01	日语	1187	381,848	24.45%
02	英语	727	253,096	16.20%
03	韩语	497	184,269	11.80%
04	越南语	486	180,260	11.54%
05	印度尼西亚语	444	146,117	9.35%
06	泰语	213	66,559	4.26%
07	西班牙语	162	50,254	3.22%
08	法语	134	44,562	2.85%
09	德语	95	33,864	2.17%
10	俄语	70	23,979	1.54%

在搜集语料的同时，完整注记了每篇作文考生（写作者）的母语、文类、题目、写作功能、总字数、能力等级、所得级分等讯息。

2.2 语料库建置过程

为了能利用工具来精确地搜寻语料库中的语言内容，中文语料需要经过断词处理。我们采用中央研究院的自动断词 (auto-segmentation) 与词性标注 (syntactic category tagging) 程序 (Chen & Liu, 1996) 进行自动处理，之后，笔者实际观察语料断词结果发现，考生因为输入法或声调问题，产生许多偏误（错字）现象，例如，在 A2 词表中出现「误会」这个词两次，并不是学习者在这个阶段会使用「误会」这个词，而是他们在输入「舞会」的「舞」时，因为声调问题，而输入「误会」这样的词。还有更多的断词或标记问题，是因为学习者声调或发音问题而造成不成词的结果，例如：

- (1) 你请我去你的无会以前 （考生将「舞会」输入为「无会」）
- (2) 我原来要去跟你们静祝静竹 （考生将「庆祝」输入为「静祝」、「静竹」）
- (3) 希望你使我的元望成真 （考生将「愿望」输入成「元望」）

对于这些问题，本研究视为输入错误，并不影响整体语言表达。因此决定以人工更正这些因为输入法或学习者发音（声调）而产生的错误，再重新进行自动断词与标记工作以提高自动断词的正确率。也就是说，在断词之前必须先进行语料前处理 (pre-processing)——人工修正错字、溢出空格、标点符号等（中研院断词系统需要以标点符号为断句依据），再重新进行一次断词。因此，语料库建置过程与步骤包括：搜集->整理属性数据 -> 语料前处理-> 自动断词、词类标记。

为利日后研究所需，每篇作文皆保留了原始语料、人工更正错字后语料与经过自动断词后的语料三种版本。

2.3 語料庫偏誤標記工作

語料庫除了為每個詞標記詞類外，為利於研究分析學習者語言，建置學習者語料庫的工作几乎都包括了偏誤標記，而這項工作目前還是必須仰賴人工，一筆一筆的確認偏誤處並標注。一般而言，愈精細的分類，愈有助於二語習得研究或學習者語言的偏誤分析，英國劍橋學習者語料庫 (Cambridge Learner Corpus, CLC) 的標記符號就多达 88 種 (Nicholls, 2003)。目前，在漢語中介語語料庫中，應該屬 HSK 動態作文語料庫的偏誤標記系統最為周詳，從字、詞、句、篇、標點符號等角度分類，約 50 種偏誤標記⁴。

然而，越精細的標記系統也有不利之處，第一、標記愈多，人為判斷的一致性愈難達成，第二、標記愈多也讓人愈難掌握標記的意義，以及造成閱讀上的困難。偏誤標記的不一致性不只在人為判斷標記類別時產生，在改成正確表達方式時，修改者間的意見更是分歧，因為同一個語義的表達方式，不只一種，很難有「最好」的改法。例如下面這個例子，至少有兩種改法，那麼究竟應該歸為偏誤分析中的誤用 (replacement)? 應該使用「想」而學習者用了「要」；還是缺漏 (missing) 呢？應該使用「想要」而學習者只用了「要」。

(寫作任務：朋友邀請參加慶生會，說明自己不能去的理由)

原文：对不起，我要参加，可是没有空。

改 1：对不起，我想参加，可是没有空。

改 2：对不起，我想要参加，可是没有空。

因此，經過初期觀察學習者語料後，同時也考慮人力時間的因素，現階段這個語料庫僅針對八大類偏誤進行標記（不做更正），包括詞匯[L]、語法[G]、形式[F]、語序[W]、語義[S]、冗詞[R]、缺詞[M]、話題[T]。次類可視研究者需要，再訂定⁵。目前完成第一階段 3 分以上（通過門檻）語料的偏誤標記工作，約 80 萬字。初步統計顯示，A2 共標記了 3991 處偏誤，B1 標記了 16122 處，B2 標記了 4984 處，C1 標記了 321 處。分別占各級字數的 3.41%、3.48%、2.75%、0.85%，可以大致看出 A2 和 B1 學習者的偏誤率⁶，與 B2、C1 學習者存在差異。

三、研究方法與初步分析

利用中介語語料庫與電腦工具除了可以快速有效地得到學習者對特定語言點的使用頻率或正誤頻率的数据外，Leech (1998: xiv-xv) 舉了几个具研究價值的問題，例如，學習者顯著地過度使用(overuse)或少用(underuse)目標語中的哪些語言特徵；在哪些方面，學習者表現像目標語的語言特徵或非目標語特徵；學習者語言行為在多大程度上受到其母語的影響等等。這些問題也正是我們建立這個語料庫所欲探索的，我們希望能從

⁴ 關於該語料庫偏誤標記，請參考 <http://202.112.195.192:8060/hsk/help2.asp> 網站上「語料標注與代碼說明」。

⁵ 現階段本計劃的偏誤標記含次類，共 25 類，標記手冊請參考張莉萍 (2012) 華語學習者中介語料庫之建構計劃—子計劃三：電腦寫作考試語料庫之建構，國科會期中成果報告 (NSC 100-2631-S-003-004)。

⁶ 這裡的偏誤率僅針對我們所標記的偏誤類而言，將標記次數除以各級總字數所得。

宏观的角度来分析学习者的语言发展，除了传统偏误分析的方法，还可以采用中介语对比分析方法（contrastive interlanguage analysis, 以下简称 CIA）来寻求答案（Granger, 1998:12）。

CIA 是语料库语言学与第二语言习得研究结合后的新领域，主要是针对两类语料进行对比，一是本族人语料库和中介语语料库的比较（NL vs. IL）；一是不同母语背景中介语语料库之间的比较（IL vs. IL）（Granger, 1998:12）。应用前者对比分析可以找出学习者少用或过度使用的语言特征；应用后者对比分析，则可能找出针对不同母语背景学习者的语言特征。Granger 等学者（1998, 2002）所主编的书中，收录了许多采用此法分析学习者语料库的研究成果。基本上，Hawkins and Buttery（2009）所进行的关键语言特征（criterial features）研究也是应用这样的方法，所谓关键特征的概念，就像是警察办案为寻找嫌疑犯所画制的素描像一般，不需要看到这个人所有的特征，只要有足以区辨的特征，就可以在众人中找出线索（Salamoura & Saville, 2010: 102）。Hawkins 等人利用 CLC（约 4000 万词），寻找英语的关键特征。例如：不同等级程度的学生使用词汇的特征、不同类型关系子句在不同等级中的分布特征、不同等级的成分搭配特征等等。该研究计划的主要目标是 (1) 建立一套能区分 CEFR 六个等级的英语关键特征，(2) 分析不同母语背景学习者在每个等级的语言表现以及和这些关键特征的互动关系。

下面几节的讨论，是结合这些研究方法下所做的尝试。

3.1 高频动词的使用情况

Hawkins and Buttery (2009: 164) 的研究中发现，初阶学习者有过度使用常用动词的倾向，他们统计分析前 10 个最常用的英语动词（know, see, think, want, mean, get, go, say, come, need），结果发现除了 mean 以外，其他九个动词，与本族人语料库（英国国家语料库，NBC）比起来，学习者都过度使用，而过度使用的趋势随着学习者语言水平增高而趋缓。图一则是我们尝试选取 TOCFL 语料库中前十高频的一般动词，各级学习者使用的情况，有同样的发现。

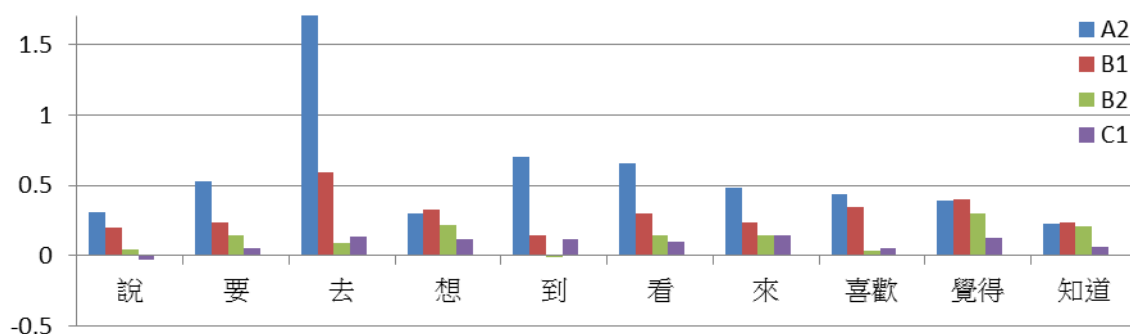


图 1 十个高频动词在 TOCFL 语料库和平衡语料库的频率比较

图中左边数值代表相对频率，0 基线表示本族人语料库（以下简称 NC）标准⁷，低于 0，表示少用；高于 0，表示过度使用。从图一也可以看到一个趋势，随着学习者语

⁷ 本研究采用的是中央研究院现代汉语平衡语料库，请参考 <http://db1x.sinica.edu.tw/kiwi/mkiwi/>。

言水平的提高，使用高频动词的频率越趋近本族人。「过度使用高频动词」也许是一个足以区别学习者程度的关键特征。

3.2 连接词语的使用情况

我们从初步的词表统计资料中发现（请见表 3），「因为、所以」这组连接词，在 A2、B1 语料库中，「所以」使用频率较高；在 B2、C1 语料库中，则「因为」使用频率较高。而 NC 呈现的则是「因为」使用频率高于「所以」，与 B2、C1 语料的结果一致。我们另外观察「可是、但是、不过」，也呈现类似的倾向，学习者从 B2 开始，三者使用频率高低与 NC 一致。从这初步的观察，似乎也印证了越高阶学习者的语言能力越趋近目标语人士。不过，这些词语的分布差异也可能跟语体有关，例如，书面语或口语语体的因素。于是，我们另外对比了一个本族人口语语料库（以下简称 SNC）⁸，发现「因为」的频率仍然较「所以」高；然而「可是」这组的频率与 NC 表现不同，「但是」的频率没有「可是」高。由此大致可见，「因为、所以」的使用频率与语体没有太大关系，但「可是」这组的使用频率和口语/书面语语体有关。我们推测，B2 以上学习者文本性质较接近书面语体，以 B2 考生所要达成的写作任务来看，这应该是合理的推测。

表 3 学习者连接词语使用频率与本族人语料库 (NC) 比较

	因为	所以	可是	但是	不过
A2 (%)	0.790036	1.175944	0.689832	0.091922	0.029813
B1 (%)	0.627358	0.765252	0.433411	0.186997	0.091256
B2 (%)	0.507341	0.441719	0.212861	0.238979	0.135813
C1 (%)	0.376188	0.324562	0.215236	0.213717	0.103632
NC (%)	0.155	0.133	0.051	0.102	0.053
SNC (%)	0.847238	0.561372	0.475045	0.156129	0.08534

从上述学习者使用「因为、所以」这组连接词语的趋势，我们不禁好奇学习者在初级阶段，使用「所以」的频率远高于「因为」的原因是什么。不同母语背景学习者是否具有相同的倾向？虽然影响语言学习/习得的因素很复杂，除了母语迁移外，还有语内迁移、学习策略、教学（教材）输入等等，不过，由于语料库中绝大多数初阶学习者是在台湾接受华语课程，可以排除教学、教材这些变量对不同母语背景学习者的影响，优先选择母语这个变因来观察。从表 4 中，可以明显看出母语为日语学习者使用「所以」频率普遍较母语为英语学习者高；英语学习者则是使用「因为」的频率较日语学习者高。从表中，也可以看出英语学习者使用「因为、所以」这组连接词的倾向和目标语人士较接近；日语学习者则到 B2、C1 等级才有这个倾向。

表 4 「所以、因为」英日语学习者使用情况

⁸ 本研究采用的是中研院对话语料库，该语料库由 30 个自由对话，26 个地图导引对话与 29 个主题对话的内容整理而成。对话总长度约 42 小时。经中央研究院词库小组自动断词与词类标记系统处理后，约计 40 万词。该语料库词表于 2012 年 4 月公开，可在 http://mmc.sinica.edu.tw/home_c.htm 下载取得。

		A2 (%)	B1 (%)	B2 (%)	C1 (%)
所以	母语为日语	1.555 ⁹	1.025	0.536	0.276
	母语为英语	0.984	0.587	0.282	0.171
因为	母语为日语	0.654	0.563	0.500	0.265
	母语为英语	0.813	0.626	0.585	0.372

推测日语学习者的这个使用现象受母语影响的可能性不低。因為在日語中，だから（所以）的頻率遠高於ので（因為）。以上是从词语使用上的表现来讨论母语迁移这个因素，这个因素是否可以成为区辨 B1 和 B2 学习者的一个特征，也就是说，B1 和 B2 学习者之间重要的区别是，B1 学习者的语言表现在极大部分仍受到母语的干扰，当然，这个预测要成立的话，还必须从更多方面来验证。至于是不是可以说具备 B2 能力的学习者在连接词语方面的表现与母语人士表现趋于一致，也有待日后从更多面向来考察。

3.3 把字句的使用情况

「把字句」一直以来被视为华语学习的难点(吕文华, 2008: 339; 邓守信, 2009: 110)，一般学者所持的看法是学习者对把字句采取回避策略，然而近年来，一些以语料库为本的研究显示，学习者不如我们预期中的少用把字句，有些研究甚至指出把字句过度使用的情况不少，例如，刘颂浩（2003）、黄月圆、杨素英（2004）；另外，张宝林（2010）根据 HSK 动态作文语料库统计出来数据，针对把字句，学习者使用率为 0.092%（仅次于「是…的」句、是字句和有字句），文中并举把字句在人民日报（本族人语料）的使用率介于 0.0754%-0.0767%之间，显示学习者使用把字句的频率不比目标语人士低。另外，我们从 NC 和 TOCFL 语料库中统计出来的使用率分别是：0.144%、0.138%。虽然在台学习者使用率没有较本族人高，但相距不大，综合两地学习者与本族人语料库数据显示，学习者使用把字句的频率与母语者并没有显著差异。

虽然学习者整体使用把字句的频率与母语者没有显著差异，然而我们观察不同能力学习者之间的使用情况发现，是有显著差异的。如表 5 所示，使用率随着语言能力的提高而增加，而且这差异达到显著标准。如果从 A2、B1 学习者使用率低于平均来看，一般教师或专家学者认为学习者回避使用把字句的「感觉」也就没错，加上前述 HSK 语料库所收集的是高程度考生作文语料，更加可以印证语言能力越高的学习者越趋近本族人使用词语的情况。

表 5 把字句在 TOCFL 语料库中分布情况¹⁰

级别	使用次数	使用率 (%)
A2	106	0.0878
B1	544	0.1220

⁹ 数据的算法是将日语者使用「所以」词数除以日语者总词数，表示母语为日语学习者，使用「所以」的频率。

¹⁰ 由於 C1 考試語料太少，這裡的 C 級指的是原 C1 考試語料加上台師大國語中心（MTC）C 級手寫作文語料，總詞數是 263,432。

B2	643	0.2099
C	353	0.134

四、结语与未来工作

本文利用所建置的语料库，观察前十高频一般动词、连接词「因为、所以」、把字句等三个学习者使用情况，可以看出一个共同现象，即随着学习者语言能力的提高，语言表现越趋近本族人语言表现。而且，这个现象是透过语言特征的具体数据所得到的实证结果。至于这些语言特征是不是可以用来作为区辨不同能力等级的关键特征，则需要更进一步的分析。现阶段的工作只是起步也是尝试，如果可行，未来期望能提供各等级能力描述更具体的内容。举例来说，全美外语教学学会 (American Council on the Teaching of Foreign Languages, ACTFL) 所制订的中文能力纲要 (Chinese Proficiency Guidelines) (ACTFL, 1987) 中针对每一个等级所举的代表性例子，可以视为关键特征，撰写者以列举语言特征的方式来说明不同能力程度学习者达到的语言能力内涵，这些语言特征是针对某一等级特别突出的特征，以下是书写能力中上程度 (intermediate-high level) 的中文纲要内容：

Chinese. Grammatical features of writing style are still essentially reflective of speech. Can express time frames somewhat accurately through the use of time words rather than particles. However, rather consistent control of perfective aspect marker *le* where it parallels writer's native language past tense. Demonstrates a basic control of syntactic patterns, all typical of speech, beyond simply elaborated skeletal S-V-O sentences by using patterns such as preposed object for topic prominence (*Gōngzuò, tā hái méi zuòwán*), resultative/directional constructions (*bānjìnlái*), and relative clause modification (*wǒmen qùnián mǎi de + Noun*), but with persistent errors. Patterns such as the *bǎ-* construction and the *shì...de* construction used inconsistently and with error. Emerging clause linking is evident through the use of speech-based sentential adverbs and conjunctions, such as *yàoburán, suírán...kěshì*. Evidence of connected discourse is also emerging.

(ACTFL, 1987: 485)

此份能力纲要叙述中，借着学习者对「了、时间词、把字句、是…的、动补结构、宾语前提、关系子句、要不然、虽然、可是」等等的使用来说明语言学习者在这个阶段语言的表现。然而撰写者为什么挑选这些语言特征，传统上，多是以专家经验或直觉意见为主，缺乏实证或量化的依据。

当然，利用语料库研究所受的限制也不少，例如，现阶段 C 级程度语料不足，很难量化学习者在这个阶段的语言特征；不同母语背景或不同主题、语体的语料分布不均，无法从各面向来观察分析语料。除了持续有计划的补充不足语料，研究者现阶段仍可以先从较高频的语言形式观察分析学习者语言，以促进汉语作为第二语言习得的研究。

参考文献：

邓守信（2009）《对外汉语教学语法（修订二版）》。台北：文鹤出版社。

- 黃月圓、楊素英（2004）漢語作為第二語言的把字句習得研究，《世界漢語教學》2004年第1期。
- 劉駿、傅榮（主譯）（2008）《歐洲語言共同參考框架：學習、教學、評估》。北京：外研社。
- 劉頌浩（2003）論把字句運用中的回避現象及把字句的難點，《語言教育與研究》第2期。
- 呂文華（2008）《對外漢語教學語法探索（增訂本）》，北京：北京語言大學出版社。
- 肖奚強、張旺熹主編（2011）《首屆漢語中介語語料庫建設與應用國際學術討論會論文選集》，北京：世界圖書出版公司北京公司。
- 張寶林（2010）回避與泛化——基於HSK動態作文語料庫的把字句習得考察，《世界漢語》第24卷第2期，263-278。
- American Council on the Teaching of Foreign Languages [ACTFL]. (1987). ACTFL Chinese Proficiency Guidelines. *Foreign Language Annals*, 20(5), 471-487.
- Chen, K. J. & S.H. Liu (1992) Word identification for Mandarin Chinese sentences. *Proceedings of COLING 1992*, 101-107.
- Granger, Sylviane (ed.) (1998) *Learner English on computer*. London & New York: Longman.
- Granger, Sylviane, J. Hung & S. Petch-Tyson (eds.) (2002) *Computer learner corpora, second language acquisition and foreign teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Hawkins, John A. & P. Buxter (2009) Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Program. In L. Taylor and C. J. Weir (eds.), *Language testing matters: investigating the wider social and educational impact of assessment*, Studies in Language Testing volume 31, Cambridge: UCLES/Cambridge University Press, 158-175.
- Leech, Geoffrey (1998) Learner corpora: What they are and what can be done with them. In Sylviane Granger (ed.), *Learner English on computer*, xiv-xx. London & New York: Longman.
- Nicholls, Diane (2003) The Cambridge learner corpus: error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference*, 572-581. 28-31 March, Lancaster, United Kingdom.
- Salamoura, Angeliki & Nick Saville (2010). Exemplifying the CEFR: criterial features of written learner English from the English Profile Programme. In I. Bartning, M. Martin and I. Vedder (eds.) *Communicative proficiency and linguistic development: intersections between SLA and language testing research*, 101-132. Eurosla: Monographs Series, 1.