

面向華語學習者的語料庫設計與建置

張莉萍*、周均亭

國立臺灣大學、國立台灣師範大學

lchang@ntu.edu.tw*

摘要

在外語教學領域中，各種教學法隨著時代興起或更迭，各種教學法都有不少的研究文獻，在華語教學界無論是為學習而建置的華語語料庫或是華語語料驅動學習方法（Data-Driven Learning, DDL）的相關研究則相對地少。為了突破語料庫在華語教學中的應用，本研究建置了專門針對華語學習或教學的語料庫——中文簡單語料庫（Chinese Easy Corpus, CEC）。CEC 有兩大功能，一是提供 200 多個經過人工編修的中高級語法點教學材料，總共 17 多萬個詞，可以視為一個小的學習型語料庫。二是為了剛接觸語料庫的初學者特別設計的簡化版語料庫，提供相對簡潔單純的操作介面，並簡化了檢索的方式。除了友善的使用方式外，本研究並參考了國外為語言學習而建置的語料庫如 SkELL、Learner's Sönaveeb 的作法，納入好句子指標參數（Good Dictionary Examples, GDEX），透過計算出的分數高低來呈現完整、易讀的句子，希望解決傳統大型語料庫應用於語言教學時經常受到的批評——過難、過長、上下文語意不清等不適合學習者閱讀的句子。本研究希望透過這個平台能增進語料庫直接應用於 DDL 華語教學中的潛力，以此提升使用者對語料庫的使用度、接受度，並降低那些可能由於語料庫的複雜介面或功能所引起對語料庫的恐懼或排斥；進而能更有效地透過語料庫觀察、分析真實語料，達到自主學習的最終目標。

關鍵詞：語料庫、語料驅動學習、中文簡單語料庫、好句子指標參數

一、前言

隨著科技的發展、電腦設備的普及，以語料庫為本（corpus-based）或語料庫輔助（corpus-aided）的外語教學法逐漸受到重視，其中，將語料驅動學習（Data-Driven Learning, DDL）直接應用於外語課堂的研究不勝枚舉（Chambers & O'Sullivan, 2004; Boulton, 2010; Cresswell, 2007）。這種教學方法主張學習者透過觀察、分析真實語料的方式，並經由教師從中引導，學習者最終能發現、歸納出語言的規則（Johns, 1991），經過這樣的一個學習過程，已經有許多研究證實其教學成效（見 Boulton & Cobb, 2017, Lee et al., 2019 的回顧）。雖然如此，語料庫輔助語言學習或教學在不少人心裡仍是望之卻步，最大的阻礙不外乎應用語料庫的技術或門檻過高，使用者需要耗費大量的時間；或是語料庫中的真實語料對學習者而言過難或非學習目標，需要花費超乎預期的時間排除困難，不符經濟效益（Breyer, 2009; Boulton & Cobb, 2017; Kilgariff et al., 2015; Lin, 2019; Poole, 2020）。

即使如此，在外語教學領域中，各種教學法隨著時代興起或更迭，在國外各種教學法都有不少的研究文獻，在華語教學界無論是為學習而建置的華語語料庫或是華語語料

驅動學習方法的相關研究則相對地少。鑑於已有不少研究證明 DDL 能促進學習者長期記憶的保存與有利日後自主學習 (Aston, 2001)，本研究欲突破前述應用語料庫於教學的挑戰，研發專門針對華語教學的「中文簡單語料庫」(Chinese Easy Corpus) 平台，以下簡稱 CEC，以促使更多人運用 DDL 於華語學習或教學。

二、相關文獻回顧

在語言教學的課室中，DDL 有兩種實施的方式——直接法與間接法 (Boulton, 2010)。直接法，顧名思義，就是直接在課堂上直接帶領學生使用語料庫，透過檢索、分析語料來學習；間接法則是透過教師預先準備好語料庫驅動得來的學習材料 (紙本)，供學生在課堂上觀察和分析。Lee, et al. (2019) 採後設分析的方法，分析運用語料庫進行詞彙教學的 29 篇研究，這 29 篇研究的樣本都是採組間設計，即，有實驗組及對照組；他們的研究顯示，不論語料庫是以何種形式出現在課程中，效果都優於傳統教學法。具體而言，兼採間接、直接的綜合教學法教學成效最佳，效果量 (effect size) g 值達到「大效果」的程度；直接教學法次之，具有中等效果量；採紙本的間接教學法，雖僅達到小效果量的門檻，但成效也同樣優於傳統教學法。這三種方式中，綜合教學法的成效又顯著優於只採紙本的間接教學方式。如此可見，動手操作語料庫似乎對學習者的語言習得有實際的助益。

DDL 最早應用於英語做為外語 (ESL) 的課程，之後應用至不同學習者背景及目的語的課室環境中 (Mizumoto & Chujo, 2015; Boulton & Cobb, 2017; Lee et al., 2019)，張莉萍、曾鈺婷 (2023) 的前導研究則是將 DDL 應用於華語易混淆詞的教學中，招募了 5 位中高級程度的華語學習者為實驗對象，課程內容主要是針對 10 組經過驗證學習者易混淆的詞進行教學 (Chou & Chang 2021) —— (1) 結果、後果；(2) 表達、表示；(3) 滿足、滿意；(4) 適合、適當、合適；(5) 發展、發達；(6) 行業、職業；(7) 成立、建立；(8) 出現、發現；(9) 發明、發現；(10) 到達、達到。實驗課程為期 5 週，每週兩小時，共 10 個小時；前 5 個課時使用 DDL 間接教學法；後 5 個課時使用 DDL 直接教學法，使用 Sketch Engine 系統，讓學生直接操作電腦，教師在旁引導。研究結果顯示後測通過率較前測平均提升了 24%，且具有統計顯著性，顯示將 DDL 應用於易混淆詞的學習具有良好的成效。透過問卷與訪談，得知學習者對 DDL 教學法都抱持正面、積極的態度，無論直接法或間接法，學習者都偏好觀察經過篩選、安排的語料，可以從下面的訪談記錄 (摘錄自該研究第 13 頁)：

老師在的時候，我喜歡看整理過後的語料；但沒有老師的時候，我還是想看整理過後的語料來學習。(學習者 B)

比較喜歡看老師整理過語料的是因為比較快。(學習者 C)

我喜歡自學，但學習的材料還是要是老師篩選過的語料或是教材。因為不然可能會花太多時間在錯的句子或是不典型的句子上面。老師給我們已經篩選過的句子，這樣學習比較快。(學習者 E)

這個研究是第一個兼採 DDL 直接法與間接法於華語詞語教學的課堂中，而學生偏好間接 DDL 的原因幾乎都是為了省時。因此如何提升 DDL 的課堂效率，節省教師與學習者的時間應該是持續要努力的方向。

一般傳統的大型語料庫基本上不是針對教學所設計，容易引發的問題包括，語料庫有時檢索出詞彙特殊、非典型用例，這些非典型的用法反而成為干擾，降低學習的效率；再者，檢索結果對學習者過難，不易理解，可能導致學習者從此對語料庫感到卻步，造成反效果（Kilgarrieff et al., 2015）。這使得語料庫應用於 DDL 時面臨著一定的挑戰，因此，針對語言教師、學習者設計的語料庫便應運而生，其中以 Sketch Engine 所開發的 SkELL 最為人所知，類似概念的語料庫還有以愛沙尼亞語為目的語的 Sõnaveeb 語料庫。為了自動將適合學習者的句子從鉅量的語料中篩選出來，Kilgarrieff et al. (2008) 依據 Atkins and Rundell (2008) 提出的詞典例句選擇原則¹，開發出 GDEX (Good Dictionary Examples) 的概念，透過計算機，計算句子中一些可量化的指標（如句長、詞彙頻率等），再透過公式將這些指標加權後得到總分，由總分的高低決定句子是否呈現給使用者以及呈現順序的依據。以下分別介紹兩個為了學習或教學所建置的語料庫：SkELL 及 Sõnaveeb。

SkELL 是專門為了學習者及教師設計，因此其操作介面及功能經過了簡化，檢索的結果與排序亦經過 GDEX 的分數決定，而 GDEX 的分數最主要決定的指標為下列 6 項（Kilgarrieff et al., 2008）：

1. 句長：句子所含的詞數應在 10 至 25 詞之間，避免過長或過短的句子；
2. 句子中詞語的頻率：句子中的詞語應由高頻詞組成，避免句子出現排序在 17,000 名之外的詞及罕用詞；
3. 句子中是否包含代名詞、指示詞：若句子中包含了代名詞或指示詞，代表該句需要在有上文語境的情況下才能完全理解，將降低「可讀性」；
4. 目標詞出現的位置：以目標詞出現在主句的句子，或目標詞出現在句子結尾位置為優先；
5. 句子是否完整：句子的完整度以該句是否由首字大寫的詞開頭，以句號、驚嘆號或問號結束為判斷的根據；
6. 是否與強搭配詞共現：目標詞與其高頻搭配詞同時出現的句子優先。

除了上述 6 項重要的指標外，GDEX 也納入其他的指標，像是黑名單、灰名單。黑名單是用以排除包含來自政治、宗教等議題的敏感詞（PARSNIP, Politics Alcohol Religion Sex Narcotics Isms Pork）的句子；灰名單的目的則是為了降低使用罕用詞或非罕用標點符號句子的分數²。

透過應用 GDEX 改善檢索結果及簡化介面便於使用者操作，SkELL 在推出後，成為 DDL 教學法的新寵兒，如 Hirata and Hirata (2019) 使用 SkELL 做為檢索工具的閱讀課程，學生的外語程度介於初中級而已，這與前人對於 DDL 僅適用於中高級程度以上學習者的認知不同。Troy (2022) 則是以 SkELL 做為 DDL 詞彙教學活動的工具，在課

¹ Atkins and Rundell (2008)認為好的例句需要能夠：1. 展現該詞典型、高頻的面向，且用法廣泛分布在真實的語言使用中（well-dispersed patterns of usage）；2. 提供足夠的資訊，有助於闡明其意；3. 有足夠的「可讀性」，而這所謂的可讀性指例句易使學習者理解。要使學習者易於理解，則例句不應包含不必要的困難詞彙與結構。同時，例句應在沒有上下文的情況下也能理解，因此例句中不應包含指代不明的代名詞、指示詞，因為這些語言成分需要依靠之前的語境來建立指涉。

² 由於 SkELL 似乎沒有公開其完整的 GDEX 計算公式，我們僅能大致瞭解其參考的指標項目，無法確切得知他們如何使用這些指標配分。在 GDEX 的計算上，另一個針對第二語言學習者的語料庫 Sõnaveeb 倒是有相當完整且詳細的討論。

堂中，英語程度介於初中級至中高級的學生，在不查閱課本或詞典的情況下，透過檢索 SkELL 並與同學討論生詞的用法與詞義。課後調查的結果顯示學生認為 SkELL 能夠有效幫助他們學習詞彙，他們也享受使用 SkELL 來學習詞彙的過程，超過半數的學生認為 SkELL 相較一般詞典是更好的學習工具，且在課程結束後，仍有 61.5% 的學生希望繼續使用 SkELL 學習詞彙。

Sõnaveeb 是愛沙尼亞語言研究機構所開發的網路詞典³，也是第一個能讓使用者查找真實愛沙尼亞語語料的資料庫 (Koppel, 2017; Koppel et al., 2019)。除了愛沙尼亞語外，Sõnaveeb 也收錄部分俄語及英語詞條，其中愛沙尼亞語的詞條有 30 萬；俄語為 22 萬詞；英語為 11.6 萬詞。對於每個詞，Sõnaveeb 除了提供該詞在詞典上的定義及例句外，亦顯示由該詞組成的短語，及其可能對應到俄語的翻譯詞，此外，該系統使用愛沙尼亞國家語料庫 (Estonian National Corpus) 中的文本計算出關鍵詞與搭配詞的用例，可以顯示關鍵詞的近義詞與搭配詞，並提供該詞在真實語言的使用例句。

Sõnaveeb 針對愛沙尼亞語的學習者另外建置了一個專門的介面——Learner's Sõnaveeb⁴。Learner's Sõnaveeb 的目標族群是程度介於 A2-B1 (初級到接近中級的程度) 的愛沙尼亞語學習者，收錄 6000 個基本的詞彙。在檢索的功能上，相較 Sõnaveeb 提供的結果，Learner's Sõnaveeb 顯示的詞典定義較短，另外，為了減輕學習者的負擔，若近義詞或搭配詞超出基本 6000 詞的範圍則不予顯示。在 Sõnaveeb 中，各檢索結果提供的例句皆經過 GDEX 的計算，分數在 0-1 之間，計算的參數由硬性參數 (hard classifier) 及軟性參數 (soft classifier) 組成，限於文長，此處省略其參數內容。為了瞭解 GDEX 分數的效果與經過人工編纂的詞典例句兩者的優劣，Koppel (2019) 整理了 160 個句子，並請 5 位詞典編纂者及 9 位學習者評測句子的合適性，其中 40 句是 GDEX 分數高的語料庫句子；40 句是 GDEX 分數低的語料庫句子；40 句是隨機從語料庫挑選、未經 GDEX 評分的句子；40 句來自詞典中以人工編纂的例句。這個研究的結果顯示，最好的句子為來自詞典經人工編纂的句子，這些句子中有 96% 成功獲得受測者的認同；次佳的是經 GDEX 評為高分的語料庫句子，其中有 85% 得到受測者的認同；隨機從語料庫挑選而未經 GDEX 篩選的句子中，只有 60% 得到受試者的支持；而經 GDEX 評為低分的語料庫句子中，只有 6% 的句子受測者認為合適。這說明雖然 GDEX 挑選出來的句子仍比不上經過人工編纂的句子，但仍較隨機從語料庫檢索出的句子佳。

三、 中文簡單語料庫

為了提供更便利的資源給有意願使用 DDL 的教學者與學習者，本研究嘗試建置一個較為簡單的資源平台，暫稱為「中文簡單語料庫」(簡稱 CEC)。CEC 有兩個主要的功能，一是為教學者提供了已經製作好的語法點語料單；二是為使用者簡化了語料庫介面以及提供可讀性高的檢索結果。以下分兩小節說明這兩大部分。

3.1 語料單資源

語料單是間接 DDL 成敗的關鍵，因為語料單是學習者觀察、分析、演繹或歸納的基本來源。在製作語料單之前，教學者通常要做的工作包括瀏覽前人研究的成果或是進

³ Sõnaveeb，網址 <https://sonaveeb.ee/>

⁴ Learner's Sõnaveeb，網址 <https://sonaveeb.ee/lite>

入學習者語料庫看看學習者的使用情況，以做為準備教學材料的重點(張莉萍、曾鈺婷，2023)。對於使用間接 DDL 的教學者而言，製作語料單是費時費力的工程。而這個平台現階段已針對 208 個中高級華語語法點，提供教授這些語法點所需的語料單，多數語法點選擇自臺灣常用教材《當代中文課程》(鄧守信主編)第三、四、五冊，針對語法點給的語料(例句)，大致有三種不同的呈現方式，以下以具體舉例方式說明。

1. 依據搭配詞呈現

例如，教授「照理說」這個語法點時(出自第 4 冊教材)，讓學習者觀察的語料單有「照理說+應該」「照理說+應該+轉折」等等，這是因為透過語料庫的搭配工具，我們知道「照理說」常常和「應該」還有一些轉折詞，如「不過、卻」等共同出現，教學者透過每張語料單的引導，最終讓學習者瞭解「照理說」的用法，而非僅僅學習「照理說」的字面意義—按照道理說。這個語法點，我們製作了 5 張語料單，每張語料單 10 個句子，共有 50 個例句。例句如下：

(1) 日前他向我求婚，照理說我應該很高興的答應才對，可是有件事梗在心中，使我一時拿不定主意。

2. 依據不同結構呈現

例如，教授「無論……都」這個語法點時(出自第 4 冊教材)，我們也是準備 5 張語料單，共 50 個句子，依據的是這個語法點的次結構來呈現，包括「無論 A 或 B」「無論 A 還是 B」「無論 V 不 V」「無論+疑問詞」「無論+正反義詞」這五個結構，每個結構使用一張語料單。茲舉「無論+正反義詞」這個結構的例句如下：

(2) 無論中外，研究少年犯罪問題的調查報告中，家庭原因都是佔了最大的比率。

3. 依據字數呈現

有些語法點不適用上述的方式呈現，這時只能用傳統的詞數或是字數的多寡來呈現。例如「一面 A，一面 B」這個語法點(出自第 5 冊教材)，語料由短到長的例句如下：

(3) 一面看，一面笑

(4) 一面加麵粉，一面加水

(5) 一面打工當家庭教師，一面上學

以上所述的語法點例句來源都是從 COCT2019 書面語語料庫中檢索而來⁵，然為了因應 DDL 教學所需，部分經過人工修改，修改的主要原因是低頻詞、專有名稱或是有文化意涵、晦澀難懂的部分，這個經過人工校正的語法點語料庫，總計 172,784 詞，也可以視為為教學或學習設計的一個小語料庫⁶，因此，除了可以根據語法點來檢索例句外，也可以根據學習者想查詢的關鍵詞來檢索，如圖 1 是使用者在語法點這個語料庫中，以「下來」做為關鍵詞檢索的結果畫面，總共檢索到 118 筆，而這些語料都經過人工校正，是可讀性高的句子。除了可以檢索到使用者想要的用例外，也可以從例句中看

⁵ 國教院 COCT 語料庫網站：<https://coct.naer.edu.tw/>

⁶ 中高級華語語法點例句查詢系統，網址：<http://grammarpoints.ddns.net/grammarpoints.php>

到語法點的分類資訊。如果對哪一個語法點有興趣，可進一步點選，觀察該語法點的語料。

Advanced Chinese Grammar Query
(中高級語法點查詢)

You are looking for sentences that contain '下來'

118 sentences are found

[Download](#)

[back to front](#) [new query](#)

第 1 頁 / 共 3 頁 [下一頁](#) [最後一頁](#)

No.	ID	分類觀察	例句	GDEX
1	4_06_06_25	(維持一個狀態)	我是一個僥倖在城市裡生存下來的鄉下小孩。	88.72
2	5_02_06_03	其實是+(一個詞)+而不是	當小販從架上取下來時，我才發現這其實是白熊而不是狗。	87.26
3	4_06_06_18	(持續一段時間)	讓我堅持下來、沒有放棄的最大原因有兩個：一個是丈夫的理解和支持，另一個就是我的信念—「經歷就是財富」。	85.19
4	3_02_05_12	尤其是	有的時候，科學需要停下來想一想，尤其是當它的目標與現在社會的道德觀念有衝突時。	84.07
5	4_12_02_22	誰知道+(預期差距大)	我只是從後面叫他，誰知道他就從椅子上摔下來，我真的不是故意的。	83.37
6	4_06_06_28	(維持一個狀態)	那一夜是怎麼活下來的，我已經不知道了。	83.28
7	4_06_06_07	V下來	我在她對面的椅子坐下來。	83.11
8	S1_02_09	被+(N)+保護	在一場意外中，這個嬰兒幸運地被保護了下來。	82.35
9	14_07_01_28	續+頁+不	我倒是希望自己作夢，把夢記下來就變成小說，不然我還真不知道要寫什麼才會比較有創意。	81.97

圖 1. 在語法點系統中檢索「下來」的部分結果

例如，點選「V 下來」這個語法點（出自教材第 4 冊），可以發現有 50 個句子，教學者可以將這 50 個句子分類下載下來，成為三張語料單。第一張是觀察「下來」的方向性語義，如「坐下來、飄落下來」；第二張是表達時間持續性的語義，如「堅持下來、沿襲下來」；第三張是表達狀態的持續，如「留下來、繼承下來」。由此可見，這個語料庫既可以做為直接 DDL 的工具，也可以做為間接 DDL（紙本語料單）的工具。

3.2 簡化的語料庫

在 CEC 的平台上除了上述的「中高級華語語法點例句查詢系統」外，另一個是沒有經過人工校正的真實語料，暫稱為「一般查詢系統」⁷，這個系統未來將以中央研究院平衡語料庫 4.0 版的 1000 萬詞為語料內容。現階段測試語料約 116 萬詞，與現有的大型、平衡語料庫不一樣的地方是，這個系統是特地為語言教學、學習而設計，它的特色為人性化的檢索介面和提供適合學習的語料。舉例來說，中文不像印歐語言，詞與詞之間有明顯的空格，為了避免斷詞對使用者造成的不便，在 CEC 中，允許使用者直接輸入字串。因為就算是相當熟悉語料庫的使用者，有時也會為斷詞所困擾，加上語料庫系統中的斷詞結果不一定一致。例如，在國教院 COCT 語料庫中，動補結構「想起來」「想出來」斷為一個詞，能式補語「想不到」也斷為一個詞；但和它結構相同的「看不到」卻又斷為三個詞。初階的語料庫使用者常因為斷詞問題，而檢索不到語料。

CEC 的檢索功能除了支援一般的詞彙、詞組搜尋外，也可使用正則表達式 (regular expression)，以利檢索不連續的固定表達形式，如「對……而言」，類似的定式常出現在華語語法點中，透過正則表達式可以輕鬆檢索出來。雖然入門學習者可能還未具備這種檢索技巧，但考量自主學習的學習者，系統仍然提供此功能。

除了友善的檢索方式，CEC 也參考 SkELL 及 Sönaveeb 的做法，依漢語的語言特性及語料性質調整，加入 GDEX 的功能以自動篩選出好例句。現階段的作法由四個指標

⁷ 一般查詢系統，網址：<http://grammarpoints.ddns.net/general.php>

的原始分數經過加權後加總計算，GDEX 分數越高，代表該句越可能是學習者容易閱讀的句子（可參考圖 1 的最右欄）。這四個指標分別是「句子是否完整」、「句長」、「詞彙難度」、「詞頻」，限於文長，分數計算等細節本文略過。

四、結語

語料庫對語言研究者而言，已經算是成熟、方便的工具了。惟在語言教學上，往往因為真實語料的不可控，以及操作語料庫的門檻、教學者所花費的時間成本等因素，而未能普遍應用於課堂。為了解決這些問題，本研究建置了中文簡單語料庫（CEC），期望這個資源平台可以做為想嘗試應用 DDL 教學或學習的有效輔助資源。CEC 的特點是：（1）提供簡潔的操作介面，並簡化了語料檢索的方式，減少使用者學習操作的時間成本、降低對語料庫的恐懼；（2）CEC 參考了國外針對學習者設計的語料庫，以句子的長度、句中詞頻、詞彙難易度等可量化指標，計算出分數，以決定句子在檢索系統中出現的順序，同時排除那些由於太難、太長、不完整等不適合用於教學的句子，減少教師花費時間篩選語料，學習者也能自行以語料庫學習語言；（3）在語法點查詢系統中，提供了經過人工設計與校正的 200 多個語法點教學材料（9000 多個句子），節省教學者教案設計的時間。本研究希望透過這個平台能增進語料庫直接應用於 DDL 華語教學中的潛力，以此提升使用者對語料庫的使用度、接受度，並降低那些可能由於語料庫的複雜介面或功能所引起對語料庫的恐懼、排斥；進而能更有效地透過語料庫觀察、分析真實語料，達到自主學習的最終目標。

在這個大數據、科技日新月異的時代，外語學習的途徑已經不只限於課室，學習者可以透過各式各樣的媒介來學習目標語。本研究預期合適的語料庫對於各種程度學習者想要提升自身語言能力的協助將超乎一般人的想像，期望透過這個研究可以讓更多人關注語料驅動學習的方法與應用。

五、參考文獻

- 張莉萍、曾鈺婷（2023）。語料驅動學習的華語實驗課程與成效分析，科技與中文教學，第14卷第1期，1-25。
- 鄧守信（主編）（2016）。當代中文課程第3冊、第4冊，台北：聯經出版社。
- 鄧守信（主編）（2018）。當代中文課程第5冊，台北：聯經出版社。
- Aston, Guy. (ed.) (2001). Learning with corpora: An overview. In *Learning with Corpora*, 6-45. Houston: Athelstan.
- Atkins, B. T. S., and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60(3), 534-572.
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348-393.
- Breyer, Y. (2009). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, 22(2), 153-172.
- Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics*, 16(2), 141-158.
- Chambers, A., & O'Sullivan, Íde. (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL*, 16(1), 158-172.

- Chou, C., & Chang, L. (2021). A Preliminary Study of Confusable Words in L2 Chinese Textbooks. In *Chinese Lexical Semantics*, ed. by Ming Liu, Chu-Ren Kit, and Qi Su, 796-809. Cham: Springer.
- Cresswell, A. (2007). Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. In *Corpora in the Foreign Language Classroom*, ed. by Encarnación Hidalgo, Luis Quereda, and Juan Santana, 267-287. Amsterdam: Editions Rodopi BV.
- Hirata, Yoko, & Hirata, Yoshihiro. (2019). Applying 'Sketch Engine for Language Learning' in the Japanese English classroom. *Journal of Computing in Higher Education*, 31(2), 233-248.
- Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal*, 4, 1-16.
- Kilgarriff, Adam, Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, ed. by Elisenda Bernal and Janet DeCesaris, 425-432. Barcelona: Documenta Universitaria.
- Kilgarriff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and language learning with the Sketch Engine and SKELL. *Revue française de linguistique appliquée* [French Journal of Applied Linguistics], 20(1), 61-80.
- Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks [Automatic detection of good dictionary examples in Estonian learner's dictionaries]. *Eesti Rakenduslingvistika Ühingu aastaraamat* [Estonian Papers in Applied Linguistics], 13, 53-71.
- Koppel, K. (2019). Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks [Suitability of automatically selected example sentences for learners' dictionaries as tested on lexicographers and language learners]. *Lähivõrdlusi. Lähivertailuja* [Close Comparison], 29, 84-112.
- Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V., & Michelfeit, J. (2019). SKELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. In *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 conference*, ed. by Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, Jelena Kalla, Miloš Jakubiček, Simon Krek, and Carole Tiberius, 763-782. Brno: Lexical Computing CZ s.r.o.
- Lee, H., Warschauer, M. & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, 40(5), 721-753.
- Lin, M. H. (2019). Becoming a DDL teacher in English grammar classes: A pilot study. *The Journal of Language Learning and Teaching*, 9(1), 70-82.
- Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, 22, 1-18.
- Poole, R. (2020). "Corpus can be tricky": revisiting teacher attitudes towards corpus-aided language learning and teaching. *Computer Assisted Language Learning*, 35(7), 1620-1641.