

LANGUAGE PRODUCTION: A WINDOW TO THE MIND?

Gaëtanelle Gilquin
F.N.R.S. – Université catholique de Louvain

1. Introduction

The human mind is a black box. We know what goes into the box and what goes out, but we do not know what happens inside. However, it is generally accepted that the manifestations of the mind, in particular language, may be used as a reasonable basis to extrapolate about the mind itself, the idea being to study something unobservable by means of something observable. In this article, we will try to learn more about the phenomenon of cognitive salience by analysing two types of language production which are expected to give access to it, namely corpus data and elicitation data. Using the highly polysemous verbs *give* and *take* as a case study, we will see that spontaneous production and elicited production may point in different directions, a discrepancy which requires explanation if language production is a window to the mind.

After briefly discussing the so-called black-box problem, we will show why corpus data and elicitation data are good candidates for the investigation of cognitive salience and we will describe the method of collection of the data for *give* and *take*. The results of the corpus analysis and the elicitation test will then be compared with each other and two factors will be suggested to account for the differences thus revealed. It will be concluded that spontaneous production and elicited production offer two different, but complementary views of the mind.

2. The black-box problem

When investigating cognitive phenomena, we are faced with a major problem, referred to as the black-box problem of the human mind (Nuyts 1992). The human mind is a black box and cannot be accessed directly.¹ As Pederson & Nuyts (1997: 3) put it,

[Conceptualization] never reveals itself directly at the observable surface of human behaviour; it only ‘appears’ indirectly, in disguise, coded in or filtered through the ‘structural principles’ of the many different types of human behavioural systems, linguistic and otherwise.

In other words, the human mind can only be accessed indirectly, via some of its manifestations. Language, as one type of systematic behaviour of humans, can be considered as a “relatively privileged (...) ‘source of information’ on conceptualization”, as Pederson & Nuyts (1997: 4) note, because it “explicitly encodes and transmits conceptual information”. If we agree that linguistic behaviour somehow reflects conceptualisation, we may then consider language production as a window to the mind, making it possible to shed some light into the black box and to catch a glimpse of something that would otherwise remain in the dark.

Using language as a basis to extrapolate about the human mind, however, would normally imply that language production is uniform for, if different types of language production offer different pictures, what can be said about conceptualisation? In what follows, we will compare two types of language production, namely spontaneous production (corpus data) and elicited production (sentence production data), and see what they tell us about the human mind, and more precisely about the phenomenon of cognitive salience.

3. Two types of language production

3.1. Corpus data and elicitation data as a means of accessing cognitive salience

The question of what is cognitively salient has most often been discussed in relation to the notion of prototypicality, that is the idea that categories are organised around a member which is maximally representative of the category and maximally salient in the mind. After originating in the field of psychology, most notably

¹ Admittedly, the brain can be observed (cf. magnetic resonance imaging) and the way it reacts to certain stimuli can give us indications about the way it works, but we have no way of directly accessing more abstract cognitive phenomena such as human conceptualisation.

with Rosch's work (e.g. 1973, 1975), the notion was taken over by linguists and applied to linguistic phenomena (see e.g. Taylor 1989).

In the field of linguistics, corpus data and elicitation data have each been argued to give access to prototypicality, or cognitive salience. Schmid (2000: 30) proposes the "From-Corpus-to-Cognition Principle", according to which textual frequency as attested in corpora instantiates entrenchment in the cognitive system. The more frequent a word or word sense is in language use, the more "entrenched" (salient) it is in the cognitive system (see also Langacker 1987). With respect to elicitation data, Geeraerts (1988: 218) claims the following:

[G]iven the presupposition that introspection yields only a partial insight into the semantic structure of the words that are investigated, we can also presuppose that it will be exactly the prototypical kinds of usage of those words, that reach the introspective consciousness of the language user.

The uses that are cognitively most salient are more likely to "pass the threshold of conscious attention" (ibid.), and hence to surface in elicitation tests, than the less salient uses.

If it is true that both corpus data and elicitation data give access to cognitive salience, one would expect them to produce the same results and point to the same items as cognitively most salient. However, an analysis of the semantic distribution of the verbs *give* and *take* in the two types of data shows that this is not necessarily the case, as we will see below.

3.2. Data collection

The corpus data used in this study of the cognitive salience of *give* and *take* come from Switchboard-1 (1993), a corpus of spontaneous telephone conversations in American English.² For each verb, 1,000 occurrences were randomly extracted, from which phrasal verbs were excluded. Each occurrence was then assigned a sense, on the basis of a semantic classification including 14 senses for *give* and 17 senses for *take* (cf. Tables 1 and 2).³

	Sense	Example
1	Hand	<i>He pulled a handkerchief from his pocket and gave it to her.</i>
2	Provide	<i>Please give your seat to an elderly or disabled.</i>
3	Communicate	<i>Would you like to give me your name?</i>
4	Cause	<i>All that driving has given me a headache.</i>
5	Use	<i>Give your time to the community.</i>
6	Administer	<i>Can't you give her anything for the pain?</i>
7	Allow	<i>He was given the power to appoint the bank's chairman.</i>
8	Job	<i>My algebra teacher always gives us a lot of homework.</i>
9	Organise	<i>The ambassador is giving a banquet for the visiting president.</i>
10	No longer resist pressure	<i>The branch suddenly gave beneath him.</i>
11	Decide	<i>The judge gave him a nine-month suspended sentence.</i>
12	Do (delexical sense)	<i>He turned to us and gave a big smile.</i>
13	Idioms	<i>You have to give way to traffic coming from the right.</i>
14	Other	

Table 1. Semantic classification of *give*

	Sense	Example
1	Grab	<i>Let me take your coats.</i>
2	Move	<i>Our guide took us around the cathedral.</i>
3	Buy	<i>What newspaper do you take?</i>
4	Ingest	<i>Take two aspirins and go to bed.</i>
5	Require	<i>It takes strength and stamina to be a long-distance runner.</i>
6	Record	<i>A nurse took his temperature every hour.</i>

² Spoken data, rather than written data, were chosen on the assumption that speech is a more direct manifestation of conceptualisation than writing, which usually goes through stages of planning, monitoring and post-editing.

³ This classification is based on five learners' dictionaries: *Oxford Advanced Learner's Dictionary* (Wehmeier 2000), *Collins Cobuild English Dictionary for Advanced Learners* (Sinclair 2001), *Longman Dictionary of Contemporary English* (Summers 2001), *Macmillan English Dictionary* (Rundell 2002) and *Cambridge Advanced Learner's Dictionary* (Gillard 2003).

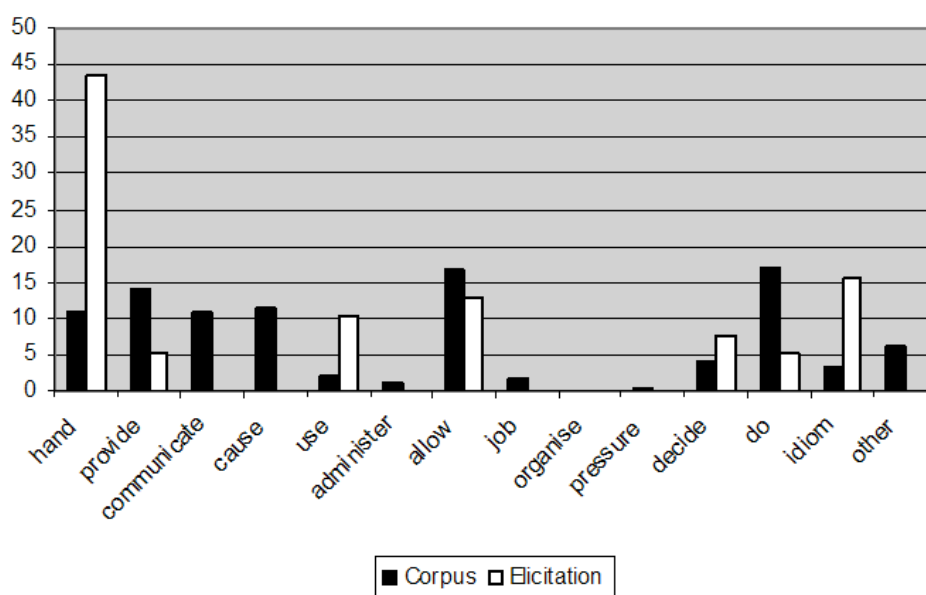
7	Engage in	<i>Shelley is taking economics at university.</i>
8	Consider	<i>She took his remarks as a compliment.</i>
9	Accept	<i>Do they take credit cards in this shop?</i>
10	Assume	<i>I did all the work, but Gill took all the credit.</i>
11	Experience	<i>The school took the full force of the explosion.</i>
12	Use	<i>It's more interesting to take the coast road.</i>
13	Capture	<i>The rebels succeeded in taking the town.</i>
14	Work	<i>If the cortisone doesn't take, I may have to have surgery.</i>
15	Do (delexical sense)	<i>Let's take a walk down the river.</i>
16	Idioms	<i>The Olympics take place every four years.</i>
17	Other	

Table 2. Semantic classification of *take*

In order to collect the elicitation data, an online sentence production experiment was carried out, where subjects were given 20 word stimuli (including *give* and *take*), one after the other and in randomised order, and were asked to type in the very first sentence they could think of including the word. The test was designed by means of the software WebExp1 (Keller et al. 1998) and was available on the Internet. It was taken by 40 participants, all of them native speakers of American English. The data produced by the subjects for *give* and *take* were classified according to the sense they exemplified, using the same semantic classification as for the corpus data.

3.3. Analysis and comparison of the data

Corpus data and elicitation data approach the problem of cognitive salience from a different angle, answering the questions “what are the most frequent senses of *give* and *take* in language?” and “what are the senses of *give* and *take* that come first to mind?”, respectively. These two questions, however, turn out to provide different answers, as appears from Figures 1 and 2.

Figure 1. Distribution of the senses of *give* in corpus data and elicitation data (%)

With *give*, the most frequent sense is the delexical one, illustrated by (1) and (2), where the verb has little semantic content of its own and the meaning is mainly expressed by the object (*thought* and *poke* in the examples). While this sense represents 17% in the corpus data, it only represents some 5% in the elicitation data. Participants in the elicitation test sometimes produce such “light verb constructions”, e.g. (3), but much more

often, they produce sentences where *give* has the sense of “handing”, as in (4).⁴ This sense seems to be well entrenched, since it represents a percentage of 43.5% (to be compared with 11% in the corpus), which is almost three times as much as the next use in the elicitation data, that of idiom (15.5%). In the corpus data, by contrast, the delexical sense is closely followed by the sense of “allowing”, illustrated by (5) and representing a proportion of 16.5%.

- (1) Well, I never **gave** it any thought myself. (Switchboard)
- (2) The worst of the violence in the Flintstones was when somebody **gave** Fred a poke in the snoot. (Switchboard)
- (3) Can you **give** me a call next week? (Elicitation)
- (4) I **gave** him the book. (Elicitation)
- (5) He went and bought one of the low priced student tickets, and they, they **gave** him the seat that normally belongs to John Bel John. (Switchboard)

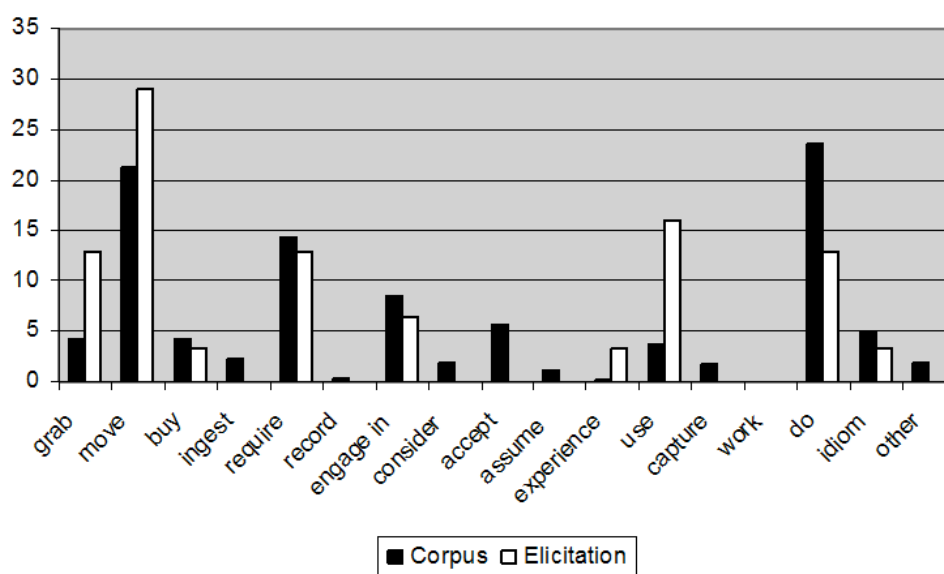


Figure 2. Distribution of the senses of *take* in corpus data and elicitation data (%)

The results for *take* present similarities with those of *give*. Here too, the most frequent sense in the corpus is the delexical one, as in (6). It represents a majority, with some 24%, but is closely followed by another sense, the sense of “moving”, illustrated by (7), which has a proportion of 21.5%. The delexical sense does occur in the elicitation data, e.g. (8), but it is not so common (13%). When presented with the stimulus *take*, subjects tend to produce sentences such as (9) and (10), where the verb has the sense of “moving” (29%).⁵ The next sense in the elicitation data, that of “using”, is comparatively much less common (16%).

- (6) And, uh, I think they ought [to] **take** a long, a good long hard look at, at who is allowed to carry a gun. (Switchboard)
- (7) Uh, it was about two years ago she got sick on me and I **took** her to the, uh, to the vet's because she wasn't eating. (Switchboard)
- (8) **Take** a shower, please. (Elicitation)
- (9) He wanted to **take** me to a baseball game. (Elicitation)
- (10) Can I **take** it with me? (Elicitation)

The comparison of the corpus data with the elicitation data reveals a discrepancy between the two types of language production. Points of similarity are not totally excluded (thus, the most often elicited sense of *take*, “move”, comes second in the corpus data), but on the whole, it appears that the most frequent sense in language

⁴ It is interesting to note that this sense is the sense that is usually described as prototypical in the cognitive literature, see e.g. Newman (1993, 2005).

⁵ Unlike the most often elicited sense of *give*, this sense of *take* does not correspond to the prototype of the verb as defined in the cognitive literature. In the literature, it is the sense of “grabbing” that is usually considered prototypical, cf. Norvig (1986) and Norvig & Lakoff (1987).

and the sense that comes first to mind do not overlap, which confirms the results obtained by similar studies (Roland & Jurafsky 2002; Nordquist 2004; Shortall, in preparation). Such results are problematic for the conception of language as a window to the mind, since they show that different types of production data may point in different directions. One possibility is therefore to object to language being considered a window to the mind, by arguing that if different types of language production offer different pictures, nothing can be said about conceptualisation. Another possibility is to claim that one of the two types of data is not a reliable source of information on the human mind. However, one would be at a loss to determine which one is unreliable, as both have been claimed to give access to cognitive salience (see above) and both have proved their relevance: elicitation is the method that has traditionally been used in psychology to establish prototypes (e.g. Hampton 1981) and frequency has been shown to have a place in the human cognitive system (e.g. Hasher 1984). Alternatively, and this is the position that will be defended here, one may recognise both corpus data and elicitation data as valid ways of accessing the human mind, but agree that they are governed by different principles and tap into different things. In the next section, two factors will be introduced to explain the differences observed between the two types of production data.

3.4. Explaining the differences

There are at least two factors that may account for the differences observed between the corpus data and the elicitation data. The first one has to do with context, and the second one with phraseology.

Elicitation data and corpus data are produced in different settings. While the former are produced in experimental settings, out of context, the latter are produced in natural settings and are part of a larger (situational and textual) context. In the elicitation test, the subjects were asked to produce a single sentence per stimulus. Roland & Jurafsky (2002) have shown that in isolated sentence production, informants tend to have recourse to a set of default referents, including first person subjects, VP-internal NPs anaphorically related to the subject and predictable head nouns for the direct object. The results of this study suggest that, in addition, people may favour the concrete over the abstract in single sentence production tasks, which might be due to the primacy of the concrete in neural representations (cf. MacLennan 1998). Connected discourse, by contrast, is governed, among others, by the phenomenon of grammaticalisation, that is a process by which some lexical words gradually lose their semantic content and turn into grammatical words, thus becoming, in a way, more abstract (see e.g. Mair's [1994] claim that the verb *see* is on the verge of becoming a function word). This difference would explain why the senses of *give* and *take* that are most common in decontextualised production (elicitation data) are concrete senses ("handing" and "moving", respectively), whereas the most frequent use in contextualised production (corpus data) is the delexical use, where *give* and *take* have lost most of their original meaning and have taken on a more grammatical function.

Another factor that appears to influence the elicitation data and corpus data differently has to do with phraseology. Language relies heavily on multi-word expressions which do not need to be re-assembled every time they are used (Sinclair 1991, Stubbs 1993). This tendency is even stronger with high-frequency verbs such as *give*, *take*, *make* or *do*, which exhibit a large proportion of phraseological uses (see Lewis 1997). The importance of such ready-made chunks for language is reflected in the corpus data, which contain a large proportion of them. By contrast, we would not expect multi-word expressions to play such a crucial role in elicitation data. This is because multi-word expressions are claimed to be stored holistically in the mind and independently of the words they are made up of (cf. Wray 2002). Therefore, as pointed out by Nordquist (2004), subjects prompted for a word access just that, and not the independent units derived from it. Thus, in her own experience, people produced sentences such as *I cough* when given the stimulus *I*, but they failed to produce the unit *I guess*, which is arguably stored independently and hence cannot be elicited by means of the word *I*. In the present case, one may argue that the verb *give* represents one entry in the mental lexicon, and that an expression such as *give a smile* represents another entry, stored as a whole unit and separate from the entry of *give*. As a consequence, when prompted for *give*, people only access the entry of *give*, and not independent expressions such as *give a smile*, *give a call* or *give birth*. This different treatment of phraseological units in natural discourse and in elicitation explains why delexical *give* and *take* are common in the former but not in the latter.

4. Corpus data and elicitation data: two windows to the mind?

It appears from what precedes that corpus data and elicitation data exhibit different characteristics, which can be related, among others, to the context of production (connected discourse vs. isolated sentence) and to the place occupied by phraseological units (inherent part of natural language vs. holistic storage in the mental lexicon). These differences account for the apparent inconsistency displayed by *give* and *take* with respect to the

question of cognitive salience, but they also suggest that, while the two types of data provide different answers to this question, both answers have their relevance and usefulness.

More generally, it may be argued, as Wray (2005) does, that corpora and elicitation tests tap into different things. Wray explains this by means of the metaphor of light. The corpus approach is equivalent to walking around a room, illuminated by an array of electric lights and where everything is visible. Elicitation, on the other hand, corresponds to a torch, which illuminates what it hits, but does not make it possible to see other objects that remain in shadow. Depending on what aspect of conceptualisation one needs to investigate, either the global view of corpora or the focused view of elicitation may prove more appropriate. For some aspects, it may even be useful to combine the pictures offered by the two approaches. Returning to the issue of phraseology, one could use corpus data to identify the most recurrent strings in texts, and elicitation data to establish whether they are processed holistically or not, as suggested by Wray (2005). In other words, corpora and elicitation may be said to offer complementary views of language.

In view of this, is it still acceptable to argue that language production is a window to the mind? The answer that emerges from this analysis is of the “yes but” type. Yes, language production may tell us something about the human mind, but different types of production data may provide different perspectives. It would therefore be more accurate to talk, not of language production as a window to the mind, but of language products as windows to the mind, with corpus data and elicitation data being two windows overlooking different aspects of the human mind.⁶

5. Conclusion

On the basis of a corpus-based analysis and an experimental study of the highly polysemous verbs *give* and *take*, it has been demonstrated that different types of language production do not necessarily overlap, as one would normally expect if language was a window to the mind. More precisely, it has been shown that spontaneous production and elicited production point to different senses of *give* and *take* as the most salient ones, with the corpus data presenting a majority of delexical uses of the two verbs and the elicitation data favouring more concrete senses (“hand” for *give* and “move” for *take*). Such differences, however, do not perforce invalidate the idea of language as a window to the mind. Provided one recognises that spontaneous production and elicited production are governed by different principles and that these principles influence what aspects of the mind corpora and elicitation may claim access to, the extrapolation from language production to the human mind is still justified. In fact, language emerges as an even richer source of information on conceptualisation than is customarily thought, since different types of language production turn out to provide different (but complementary) insights into this issue.

The odds are that the human mind will never reveal all of its secrets and will always, to some extent, remain a “black box”. However, this should not discourage us from studying it and trying to understand it. Linguists, by investigating language production, may contribute to this effort, which should be a joint effort, teaming up with neurologists, psycholinguists and other specialists in the human mind. For, the more doors (and windows) we unlock, the more light we will shed on this black box and the closer we will come to unravelling one of the greatest mysteries of the human being.

References

- Geeraerts, D. “Where does prototypicality come from?” In *Topics in Cognitive Linguistics*, edited by B. Rudzka-Ostyn, 207-29. Amsterdam; Philadelphia: John Benjamins Publishing Company, 1988.
- Gillard, P., ed. *Cambridge Advanced Learner's Dictionary*. Cambridge: Cambridge University Press, 2003.
- Hampton, J.A. “An investigation of the nature of abstract concepts.” *Memory and Cognition* 9, no. 2 (1981): 149-56.
- Hasher, L. “Automatic processing of fundamental information: The case of frequency of occurrence.” *American Psychologist* 39, no. 12 (1984): 1372-88.
- Keller, F., M. Corley, S. Corley, L. Konieczny and A. Todirascu. *WebExp: A Java toolbox for web-based psychological experiments*. Technical Report HCRC/TR-99. Edinburgh: Human Communication Research Centre, University of Edinburgh, 1998. See <http://www.webexp.info>.

⁶ One could even argue that corpora and elicitation are windows made up of several panes, each offering a slightly different view depending on the type of language that is included in the corpus (telephone conversations, political speeches, newspaper extracts, poems, etc.) or the task that is proposed in the elicitation test (sentence production on the basis of a stimulus, sentence completion, etc.).

- Langacker, R.W. *Foundations of Cognitive Grammar. Vol. I. Theoretical Prerequisites*. Stanford, California: Stanford University Press, 1987.
- Lewis, M. *Implementing the Lexical Approach. Putting Theory into Practice*. Australia: Thomson Heinle, 1997.
- MacLennan, B.J. "Finding order in our world: the primacy of the concrete in neural representations and the role of invariance in substance reidentification." *Behavioral and Brain Sciences* 21, no. 1 (1998): 78-9.
- Mair, Ch. "Is *see* becoming a conjunction? The study of grammaticalisation as a meeting ground for corpus linguistics and grammatical theory." In *Creating and Using English Language Corpora*, edited by U. Fries, G. Tottie and P. Schneider, 127-37. Amsterdam: Rodopi, 1994.
- Newman, J. "The semantics of giving in Mandarin." In *Conceptualizations and Mental Processing in Language*, edited by R.A. Geiger and B. Rudzka-Ostyn, 433-85. Berlin; New York: Mouton de Gruyter, 1993.
- Newman, J. "Three-place predicates: A cognitive-linguistic perspective." *Language Sciences* 27, no. 2 (2005): 145-63.
- Nordquist, D. "Comparing elicited data and corpora." In *Language, Culture and Mind*, edited by M. Achard and S. Kemmer, 211-24. Stanford: CSLI Publications, 2004.
- Norvig, P. "On defining polysemous words." *Proceedings of the Third Annual Conference on Theoretical Issues in Conceptual Information Processing* (Philadelphia, August 1986), 1986.
- Norvig, P. and G. Lakoff. "Taking: A study in lexical network theory." In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society, February 14-16, 1987. General Session and Parasession on Grammar and Cognition*, edited by J. Aske, N. Beery, L. Michaelis and H. Filip, 195-206. Berkeley: Berkeley Linguistics Society, 1987.
- Nuyts, J. *Aspects of a Cognitive-Pragmatic Theory of Language. On Cognition, Functionalism, and Grammar*. Amsterdam; Philadelphia: John Benjamins Publishing Company, 1992.
- Pederson, E. and J. Nuyts. "Overview: on the relationship between language and conceptualization." In J. Nuyts and E. Pederson (eds) *Language and Conceptualization* (pp. 1-12). Cambridge: Cambridge University Press, 1997.
- Roland, R. and D. Jurafsky. "Verb sense and verb subcategorization probabilities." In *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, edited by S. Stevenson and P. Merlo, 325-46. Philadelphia; Amsterdam: John Benjamins Publishing Company, 2002.
- Rosch, E. "On the internal structure of perceptual and semantic categories." In *Cognitive Development and the Acquisition of Language*, edited by T.E. Moore, 111-44. New York; San Francisco; London: Academic Press, 1973.
- Rosch, E. "Cognitive representations of semantic categories." *Journal of Experimental Psychology, General* 104, no. 3 (1975): 192-233.
- Rundell, M., ed. *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan Education, 2002.
- Schmid, H.-J. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin: Mouton de Gruyter, 2000.
- Shortall, T. *Corpus, Cognition, and Curriculum*. PhD Thesis, The University of Birmingham, in preparation.
- Sinclair, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.
- Sinclair, J., ed. *Collins Cobuild English Dictionary for Advanced Learners*. Glasgow: HarperCollins Publishers, 2001.
- Stubbs, M. "British traditions in text analysis. From Firth to Sinclair." In *Text and Technology. In Honour of John Sinclair*, edited by M. Baker, G. Francis and E. Tognini-Bonelli, 1-33. Philadelphia; Amsterdam: John Benjamins Publishing Company, 1993.
- Summers, D., ed. *Longman Dictionary of Contemporary English*. Harlow: Pearson Education, 2001.
- Switchboard-1*. Philadelphia: Linguistic Data Consortium, 1993.
- Taylor, J.R. *Linguistic Categorization. Prototypes in Linguistic Theory*. Oxford: Clarendon Press, 1989.
- Wehmeier, S., ed. *Oxford Advanced Learner's Dictionary*. Oxford: Oxford University Press, 2000.
- Wray, A. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press, 2002.
- Wray, A. "Making sense of patterns: psycholinguistic perspectives on corpus linguistic research." Paper presented at the *Corpus Linguistics* conference (Birmingham, 15-17 July 2005), 2005.