Corpus and Cognition Colloquium:
The Relation Between Natural and Experimental Language Data

Gaëtanelle Gilquin[1] and Terry Shortall[2]

Abstract

While the usefulness of corpora for the description of language cannot be denied, it must also be recognised that they are not the only sources for language data. Corpora show how people use language in authentic environments, or what is likely to occur in language, but they do not make it possible to answer questions having to do with, say, grammaticality or language processing, or how, if at all, language is structured in the mind. Hence the suggestion, made by several researchers (e.g. Kennedy 1998), to combine corpus data with other types of linguistic evidence.

One particularly interesting combination is that between corpus analyses and experimental techniques (elicitation, lexical decision, magnitude estimation, eye movement research, reaction time measures, *etc*.). While the former make it possible to study "properties of the linguistic output of language users" (Sandra 1995: 592), the latter give access to "properties of the mental processes and structures underlying language production and comprehension" (ibid.), such as cognitive salience or readability. Bringing together the two approaches, therefore, offers a more holistic view of language.

Depending on the phenomenon investigated and the types of data used (e.g. speech vs. writing, sentence production vs. self-paced reading), one may find that the natural and experimental language data converge (*cf.* Gries *et al*. 2005) or, on the contrary, that they produce different results (*cf.* Roland and Jurafsky 2002). We believe that, by examining such relations more closely, we will learn more about the specificities of each type of data and will thus be able to make informed choices about how the two can fruitfully be combined, in domains such as descriptive linguistics, sociolinguistics or foreign language teaching.

References

Gries, S.Th., B. Hampe and D. Schönefeld (2005) 'Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions'. *Cognitive Linguistics 16*, 635–76.

Kennedy, G. (1998) An Introduction to Corpus Linguistics. London, New York: Longman.

Roland, R. and D. Jurafsky (2002) Verb sense and verb subcategorization probabilities, in S. Stevenson and P. Merlo (eds) The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues, pp. 325–46. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Sandra, D. (1995) Experimentation, in J. Verschueren, J.-O. Östman, J. Blommaert and C. Bulcaen (eds) Handbook of Pragmatics. Manual, pp. 590–95. Amsterdam, Philadelphia: John Benjamins Publishing Company.

[1] Université catholique de Louvain
e-mail: gaetanelle.gilquin@uclouvain.be
[2] The University of Birmingham
e-mail: t.shortall@bham.ac.uk

Generating Well-Formed Compounds: A Corpus-Based Model Tested Against Psycholinguistic Evidence

Marco Baroni,[1] Emiliano Guevara[2]
and Vito Pirrelli[3]

Abstract

Preliminary analysis of a significant sample of compounds acquired from a very large corpus of Italian Web pages provided fresh support to a theoretically and psycholinguistically motivated typology (Bisetto and Scalise 2005, Costello and Keane 2001, Baroni *et al*. 2006), that distinguishes coordinative ("singer songwriter"), relational ("call center") and attributive compounds ("pilot experiment"). Each type exhibits, among other characteristics, different requirements on the substitutability of its constituents with semantic neighbours.

We describe a methodology aimed at integrating corpus-based evidence with acceptability judgments of "surrogate" compounds by Italian subjects. Surrogate compounds are generated algorithmically by replacing the constituents of an attested compound (head, modifier or both) with a set of their respective semantic neighbours, automatically extracted from the corpus through a Latent Semantic Analysis-like technique. We expect acceptability judgments on surrogate compounds to significantly correlate with the above-mentioned typology.

In particular, we focus here on the methodological aspects of our experiments, including:

- issues of experimental design with corpus-derived stimuli characterized by very skewed frequency distributions and high collinearity among independent variables;

- reliable statistical comparison of corpus-attestedness with acceptability judgments;

- usefulness of corpus-based semantic similarity measures in modeling human compound classification tasks.

References

Baroni, Marco, Emiliano Guevara, Vito Pirrelli and Eros Zanchetta (2006) Corpus evidence and compound structure: The case of Italian NN compounds. Paper presented at QITL-2, Quantitative Investigations in Theoretical Linguistics 2, Osnabrück, June 2006.

Bisetto, Antonietta and Sergio Scalise (2005) 'The classification of compounds'. *Lingue e linguaggio 4(2)*, 319–32.

[1] University of Trento
e-mail: marco.baroni@unitn.it
[2] University of Bologna
e-mail: emiliano@lingue.unibo.it
[3] University of Pavia and ILC, Pisa
e-mail: vito.pirrelli@ilc.cnr.it

Costello, Fintan J. and Mark T. Keane (2001) ‘Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts’. *Journal of Experimental Psychology: Learning, Memory and Cognition 27(1)*, 255–71.

Grammaticality Judgments and Language Usage Data: A Case Study on Croatian Clitic Placement

Damir Ćavar[1] and Dunja Brozović Rončević[1]

1. Introduction[2]

In this article we discuss the results of a corpus-based study of clitic placement and clitic clusters in Croatian, accompanied by results from experiments on gradual grammaticality judgments using acoustic sentence repetition, and proof reading tasks with variation in pressure on the subjects via time or instructions. On the one hand, we suggest that the uncertainty about specific low-frequent constructions in corpora, which could be due to errors of writers and reviewers, or to real variation and soft grammatical constraints of native speakers, can be eliminated on the basis of results from psycholinguistic experiments. On the other hand, the rather plausible conclusion that the relative frequency of particular constructions in corpora correlates with the certainty of grammaticality judgments of native speakers, as well as their sensitivity to identify normative or prescriptive deviations, can find empirical support, using quantitative analysis of corpora and psycholinguistic experiments.

1.1 Clitics in Croatian

The domain of clitic placement and properties of clitics in Croatian remains a matter of debate within many theoretical frameworks. It is a particularly interesting phenomenon, since it seems to be subject to phonological, morphological, and/or syntactic constraints (Ćavar 1999, and the literature discussed and cited therein). Clitics in Croatian are function words (e.g. auxiliaries, pronouns, particles) that are default unaccented and mostly mono-syllabic.

The following examples point out classical generalizations assumed in the literature, with the pronominal and auxiliary clitics in bold:

(1) Clitic second

a. *Tko* ***mu ga je*** *jučer* *dao?*
who him it be$_{3sg}$ yesterday give$_{ptc}$
"Who gave it to him yesterday?"

b. **Tko jučer* ***mu ga je*** *dao?*
who yesterday him it be$_{3sg}$ give$_{ptc}$

c. ****mu ga je*** *tko jučer* *dao?*
him it be$_{3sg}$ who yesterday give$_{ptc}$

(2) Clustering in a fixed order

Tko* ***ga je* ***mu*** *jučer* *dao?*
who it be$_{3sg}$ him yesterday give$_{ptc}$

Clitics cluster in second position in the clause, as the examples in (1) show. However, there is little agreement among linguists about the exact definition of the "second position".

[1] University of Zadar and Institute of Croatian Language and Linguistics
e-mail: {dcavar,dunja}@ihjj.hr

[2] This research was made possible with support from the Ministry of Research, Education and Sports of the Republic of Croatia by the research grant No. 2120920–0930.

Further, in prescriptive grammars (e.g. Raguž 1997) and in theoretical linguistics literature (Spencer 1991: 356) it is pointed out that the relative order of clitics with respect to each other is constrained as well, as shown in the contrast between (1a) and (2).

(3) Relative ordering of clitics in Croatian
Q-ptcl. *li* > Aux (not *je*) > Dat. Pron. > Acc. Pron. > Refl. Pron. > Aux. *je*

Usually studying the properties of certain linguistic items, and clitics in Croatian in particular, implies some of the following methods:

- Studying grammars
- Asking native speaker informants (introspection)
- Performing usage studies (corpus-based)

Prescriptive grammars, on the one hand often lack specific information of interest, and in particular with respect to clitics in Croatian a complete documentation seems to be missing in current grammars. On the other hand, they tend to generalize and idealize, dismissing real language usage facts, as well as idiolectal and dialectal variation.

Native speaker judgments are not unproblematic neither,[3] in particular, if the intuition about lexical items like clitics is murky due to the fact that they are inconspicuous, have minimalistic phonological properties, being mostly monosyllabic and unaccented, lack intrinsic semantic properties, and are related to abstract grammatical functions.

Studying the grammatical properties of clitics on the basis of language usage and corpora is problematic due to the fact that corpora might involve typos, and transcription or annotation errors. Using the Croatian Language Corpus[4] (CLC) in a state of approximately eighty million tokens from various genres, we find that violations of the generalizations mentioned above seem to occur surprisingly frequently. In particular, numerous examples that contradict the clitic cluster ordering constraint can be found in the newspaper sub-corpus of the CLC, as for example:

(4)

Sequence	**count**
je ga (“be_{3sg} it”)	32
ga je (“it be_{3sg}”)	1968
je ga je (“be_{3sg} it be_{3sg}”)	24

On the other hand, the proportions of such deviations from the normative or prescriptive order can be found in the fiction sub-corpus of the CLC as well, but seem to be less significant:

(5)

Sequence	**Count**
je ga (“be_{3sg} it”)	4
ga je (“it be_{3sg}”)	6291
je ga je (“be_{3sg} it be_{3sg}”)	0

[3] See Schütze (1996) for a discussion of the problems and pitfalls with introspection and grammaticality judgment approaches.

[4] The corpus can be accessed from the web pages of the Institute of Croatian Language and Linguistics at the following URL: riznica.ihjj.hr.

Given the significant size differences between the two corpora (newspapers with 70 mil. tokens, fiction with 10 mil. tokens), there seems to be a significant difference between the two types of genre with respect to the frequency of these clitics and the clusters in particular.

Other more frequent observed deviations from the prescriptive standard involve an accusative preceding a dative pronoun as in (6a), or an accusative pronoun followed by a clitic auxiliary which is not *je* (be_{3sg}), as in (6b).

(6) a. *da* ***ga mu je***... or *da* ***ga mu*** ***se***...
that it him be_{3sg}... that it him self...
b. *da* ***ga su*** *iskrcali*... or *sluga* ***ga su*** *svoje*...
that it be_{3pl} unload... servant it be_{3pl} his...

More frequent deviations in both genre types can be found with reflexives relative to other pronominal clitics, as for example in the following example:

(7) a. *ako* ***se*** ***ga*** *tko boji*...
if self him who fears
b. *da* ***se*** ***ga*** *treba*...
that self him need

We suspect that newspaper articles contain fewer clitic clusters and combinations in absolute counts, and more deviations from the normative grammar than fiction. There appear to be various possibilities for explanation. It could be the case that these deviations are in fact just typos or transcription errors, in particular the observations in (4)-(5). In this case we would expect some genres that undergo scrutiny in the editing and publication process to contain fewer errors than articles in daily newspapers. On the other hand, it could be the case that daily newspaper articles are in fact closer to real language usage. If certain deviations from the standard grammar are systematic, their occurrence might be idiolectally or dialectally motivated.

In fact, clear and consistent grammaticality judgments for examples with complex clitic clusters seem to be difficult to get from native speakers without linguistic expertise or just common knowledge of normative grammar. This motivates the hypothesis that constraints on sentential placement of clitic clusters and relative position of clitics within the cluster are rather *soft*, being much more subject to idiolectal and dialectal variation than constraints on for example substantives and purely syntactic regularities. On the other hand, we observe an increase of uncertainty in native speakers the more they are confronted with variations in relative order of the clitics in clusters or the relative position of the cluster in the clause.

We also suspect that the reliability of the native speaker's intuition depends on the relative frequency of these particular constructions in real language data. Corpora potentially offer a possibility for approximation of base-measures of familiarity,[5] via simple likelihood estimates for lexical and syntactic types. On the one hand, statistical significance tests could help us in finding support for either hypothesis (typo or unclear judgments). On the other hand, aiming at studying the real performance of speakers, processing experiments seem to be necessary.

[5] Studies of first language acquisition show that frequency plays an important role in memorization during the acquisition period, as for example pointed out in Kidd et al. (2006). In the same way, processing phenomena in adults show frequency sensitivity, as shown for example in Theakston (2004).

2. Experiments

We designed two experiments to test native speaker behaviour with the target constructions, as found in the corpus, trying to avoid intuition-based or direct judgments. The goal is to provide support for the hypothesis that indeed the judgments are murky.

Two types of experiments can help identifying the speaker's unbiased behaviour with the target constructions:

- **Repetition task:** subjects are asked to repeat the sentence they hear. The expectation is that subjects will produce more frequently errors with constructions that deviate from their own intuition, i.e. we should observe conscious or unconscious auto-correction of certain word order constraints.
- **Proof reading task:** subjects are asked to correct a text, given a variation in instructions for increasing or decreasing the pressure and identifying gradual judgments. We expect to subjects to be hyper-corrective with less common or more complicated constructions, with a specific set of instructions, while they should be more liberal given another set of instructions. In any way, gradual judgments should result for the same text and target constructions, determined by the type of instruction.

2.1 Stimuli

The stimuli are constructed as follows:

a. 80 percent unrelated and well-formed constructions
b. 10 percent constructions with syntactic word order violations, that do not contain the lexical target elements (i.e. pronominal and auxiliary clitics), as for example:
   Word order violation:
   **Što taj čovjek priča o svojoj brodici držajući u ruci?*
   what this man $tell_{3sg}$ about his boat holding in hand
   Agreement violation:
   **Taj čovjek pričaju o svojoj brodici.*
   this man $tell_{3pl}$ about his boat
c. 10 percent target structures where only the position of clitics in the clause relative to each other is deviating from the classical generalizations, but otherwise no grammatical deviations occur, i.e. the clitic positions are swapped (marked in bold), as in the following example:
   a. *Taj čovjek **je** **mu** pričao o svojoj brodici.*
      this man be_{3sg} him $tell_{ptc}$ about his boat
   b. *On **je** **ga** opisivao kao starog sijedog gospodina u crnome odjelu.*
      he be_{3sg} him $describe_{ptc}$ as old gray-haired gentleman in black dress

For the repetition task the stimuli represent a list of sentences, recorded as spoken by a native speaker. They were presented to the subjects via headphones monitored by an investigator. The proof reading task uses written text of one page per subject, and the subjects are asked to mark typos and errors in the text.

In the initial pilot studies we used examples with deviations from the normative grammar, as shown in the following samples for each type of deviation:

(8) a. *Deset godina protestirao je protiv nadimka i to **ga** **mu** **je** učvrstilo.*
ten years protested be against nickname and this it him be

manifested

b. *Brže-bolje* ***ga*** ***su*** *prodali u Cibonu već sa 14 godina.*
helter-skelter him be sell in Cibona already with 14 years

c. *Očito je August imao svoje kombinacije, za koje još nije*
obvious be August have his combinations for which so-far not-be
smatrao da su zrele da ***ih*** ***nam*** *otkrije.*
considered that be ripe that them us present

2.2 Repetition task

In the first experiment the target sentences are presented as auditory stimuli to subjects, whose task is to repeat the sentence they heard, as accurately as possible, and immediately after the output is finished.

The response is recorded, the time between end of stimulus and start of response is measured, and the deviations are registered.

The target structures are sentences with deviation of clitic placement regularities, in particular relative order constraints between the clitics themselves.

The expectation is that the number of corrections of deviating structures by subjects is proportional to their certainty. In other words, we expect subjects to automatically correct (or swap) wrong clitic sequences, if the sequences are hard constraints, proportional to corrections of other word order violations involving substantives.

The corrections on target structures are relativized on the basis of the proportions of corrected non-target violations from the structures in (2b) and (3).

This task (E1) is expected to partially neutralize the influence of normative or standard grammar rules, thus be closer to the speaker's individual intuition, by involving time pressure on the subject and excluding direct reference to introspection.

2.3 Proof reading task

The second experiment is set up as a text proof reading task,[6] where the subjects are asked to correct a small essay for publication in a local linguistic journal. The proportions of distracter and target structures are kept the same as in the first experiment.

Three subject groups are defined. Each of the groups is confronted with different instructions for the same text:

- The first group is informed that the essay was written by a well-known professor and Croatian linguist, who is pointed out to be a capacity in his field.
- The second group is informed that the text was written by a colleague with less experience in writing essays, who is native-speaker of Croatian.
- The third group is informed that the text was written by a colleague who joined the research group recently, as a non-native speaker of Croatian.

All subjects are asked to correct the essay as soon as possible, mentioning that the publication deadline is due and the results are required urgently.

While the expectations in the spoken language repetition task (E1) are that subjects are more likely to correct stronger grammatical violations, and less so constructions where we

[6] Reinhold Kliegl (Department of Psychology, University of Potsdam) initially suggested such a method to us in the context of a different project.

assume soft constraints being violated, in the proof reading task (E2) we expect to see overcorrection of complex sentences for the third group, i.e. a higher false alarm rate. On the other hand, we should observe a higher acceptance rate for soft violations of grammatical constraints with the first type of instruction, assuming that the authority of a language expert will make the native speakers doubt uncertain intuitions.

In this task the knowledge of normative or standard grammar should be more dominant than in task E1, with the uncertainty being overridden via pressure by instructions of less common construction types.

The results should allow for graduation of judgments on the target structures. If a deviation from the normative constraints on clitic order is observed in all three groups, we might conclude that the normative generalization (3) is either ad-hoc, or ignoring a certain amount of freedom that native speakers tolerate. As for the corpus, the consequences have to be that we would have to distinguish between real errors or typos and deviations from the norm or standard, not just on a dialectal level, but also as idiolectal performance in text.

This method should allow a fine-grained classification of the target structures with respect to complexity and the nature of the underlying constraints.

3. Conclusion

Since we would classify all experiments so far as just pilot studies, we present rough scores for some of the crucial findings, which we interpret as rough tendencies only. More precise results will be presented in the near future.

Our initial pilot studies confirm some of the expectations. Given the special status of dialects and variations in Croatia, with Kajkavian, Čakavian, and New-Štokavian in its Ijekavian variants the major dialects, native speakers tend to have different preferences with respect to both variables, i.e. the position of the clitic cluster within the clause, and the position of clitics within the cluster relative to each other. Further, the certainty of the tested subjects with respect to the normative or prescriptive rules of standard Croatian seems to be low. Subjects tend to be increasingly uncertain about these variables, the more they are confronted with more concrete questions and tasks, that is they do not seem to be overly sensitive to violations of the relative order constraint of clitics with respect to the prescriptive placement rules. On the other hand, their preference for the relative position of the clitic cluster in the clause seems to be guided by their dialectal origin.

In the proof reading task (E2) for example, from twenty-one target structures (ungrammatical clitic order on the basis of the normative grammar) with swapped clitic positions within the clitic cluster, we find that three speakers with their dialectal origin in New-Štokavian Ijekavian are very certain about their suggested corrections, marking one or two target structures as stylistically marked and providing for all the others clear typological correction suggestions that correspond to the normative rules.

On the other hand, all the other subjects marked up to seven target structures as stylistically marked, being uncertain and not suggesting a correction. In the average, even the New-Štokavian Ijekavian speakers accepted three target structures as well formed. The Kajkavian speakers accepted in the average thirty percent of the target structures. The target structures with swapped pronominal clitic sequences Accusative > Dative were mostly within these thirty percent. Sequences with swapped pronominal and auxiliary clitic sequences (Accusative > Auxiliary (non *je*)) were overall less acceptable and rather marked for correction. Speakers with a Čakavian dialectal origin seem to accept rather fifty percent of the target structures as well formed. Within subject analyses show that the majority of the subjects refused the same clitic sequence in some contexts and structures, but accepted it in

others. There is no clear relation between specific clitic sequences and particular dialectal origin or individual subjects.

The pilot studies made clear that another variable seems to be crucial for inclusion in the experimental designs and evaluations. The dialectal origin is obviously an important factor that needs to be considered when defining subject groups.

The high level of uncertainty and reference to marking of target structures as being "stylistically marked" and not necessarily ungrammatical implies a lower consciousness of the subjects with respect to the normative rules of these particular constructions. Thus, their dialectal origin seems to be transparent even in the proof reading task.

Overall, we find support for our hypothesis that the observed phenomena in the CLC are not due to typos and errors only, but rather find some explanation along the lines of idiolectal and dialectal variation that enters different genres in varying forms and amounts. While for some observed combinations of clitics in the cluster it seems indeed to be the case that they are based on errors, others are clearly not.

The placement of clitic clusters seems to follow a general rule of tendency, rather than clearly definable restrictions. Depending on the dialectal origin these tendencies vary. The relative position of clitics in the cluster seems to be subject to similar soft constraints. Such an observation is not surprising, given the nature of clitic elements, being linguistically subject to various constraints from (independent) linguistic subcomponents.

As for the study of clitics in Croatian in the linguistic literature[7] we observe one general mistake, i.e. ignoring the richness and varieties of dialects and languages within what was historically wrongly subsumed under the label Serbo-Croatian. Generalizing from the normative description of some hybrid artefact (i.e. Serbo-Croatian) over the real language facts is not leading to an understanding and explanation of real language phenomena. In addition to that, we clearly observe that certain language phenomena are subject to soft constraints or principles, which cannot be described or captured in terms of clearly deterministic descriptions or explanations in grammar formalisms.

Further investigations, given the pilot study experience, will involve new variations in the time pressure, as well as inclusion of the variable of dialectal origin in the grouping of the subjects.

References

Ćavar, D. (1999) Aspects of the Syntax-Phonology Interface, Doctoral dissertation, University of Potsdam.

Kidd, E., E. Lieven and M. Tomasello (2006) 'Examining the role of lexical frequency in the acquisition and processing of sentential complements'. *Cognitive Development 21(2)*, 93–107.

Raguž, D. (1997) Praktična Hrvatska Gramatika. Zagreb: Medicinska Naklada.

Spencer, A. (1991) Morphological Theory: An Introduction to Word Structure in Generative Grammar. Oxford: Basil Blackwell.

Schütze, C.T. (1996) The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology. Chicago: University of Chicago Press.

Theakston, A.L. (2004) 'The Role of Entrenchment in Children's and Adults' Performance on Grammaticality Judgment Tasks'. *Cognitive Development 19(1)*, 15–34.

[7] Listing all the linguistic work here that makes this mistake of over-generalization on the basis of Serbo-Croatian would extend the formal restrictions on this article.

Integration of On- and Offline Linguistic Evidence for Capturing the Cognitive-Functional Motivations of Syntactic Variation

Gert De Sutter[1]

Abstract

The present paper combines evidence from a multivariate corpus analysis and psycholinguistic experimentation in order to find an adequate cognitive-functional explanation for the coexistence of [part+aux] and [aux+part] clusters in Dutch complement clauses:

a. dat ik een boek *gekocht*$_{part}$ *heb*$_{aux}$
b. dat ik een boek *heb*$_{aux}$ *gekocht*$_{part}$

First, we tried to find out which language-internal factors influence the choice of word order. To that end, we extracted all relevant verb clusters from one of the regionally and stylistically controlled components of the ConDiv-corpus of contemporary written Dutch (n = 2,390), annotated them for nine language-internal (structural, semantic, discursive) factors, and fitted a binary logistic regression model (main effects only). The resulting model reveals a.o. significant effects for eight out of nine variables, with the semantic factor as the most influential one. The model is able to explain and predict 80 percent of the variation.

Building on this robust statistical model, an overarching cognitive-functional explanation was developed and (partially) tested in an eye-tracking experiment: [part+aux] order is the basic word order on which language users fall back in circumstances of heavy production demands, whereas [aux+part] order is considered a socio-stylistic option. The results of the experiment, though premature, point at significant processing differences that are in line with the explanation.

[1] University of Ghent
e-mail: gert.desutter@hogent.be

Reconciling Corpus Data and Elicitation Data in FLT

Gaëtanelle Gilquin[2] and Terry Shortall[3]

Abstract

Because they tap into basically different things, corpora and elicitation tests may at times provide diverging results. In this paper, we will show that, far from being opposed to each other, corpus data and elicitation data should be seen as complementary and that their reconciliation can actually prove useful in a field such as Foreign Language Teaching (FLT). Two case studies will illustrate this. In the first one, we will use data coming from a learner corpus and from fill-in and evaluation exercises to investigate learners' knowledge of *make*-collocations. We will see that only the combination of corpus and elicitation data can give the full picture, i.e. performance and competence, and that an analysis relying on just one source of language data therefore runs the risk of being biased. In the second case study, we continue to investigate the differences between competence and performance through a comparison of the existential structure in elicited data from native speakers and in a spoken corpus. The divergences between the prototype effects displayed by the elicited data and the frequency effects found in the corpus data are discussed in the light of Foreign Language Teaching, and it is argued that prototypes should be taught first and that structures should be revisited in a cyclical fashion as proficiency increases, with all extensions of a structure being eventually introduced to students.

1. Introduction

In contrast to linguists working three or four decades ago, who had to rely mainly on introspective evidence in order to study language, today's linguists have at their disposal large collections of naturally-occurring language, searchable at the click of a mouse. Not only have these corpora made it possible to investigate aspects of language that had been largely unexplored by earlier generations of linguists (e.g. frequency, genre variation or phraseology), but they have also thrown new light on previously studied phenomena. Thus, it is not unusual, when comparing findings based on corpus data with those obtained through intuition or experimentation, to notice important differences (see Berry 1994 for an example of differences between corpus data and intuition, and Roland and Jurafsky 2002 for a study illustrating the differences between corpus data and experimentation). For many, such differences highlight the limitations of more introspective sources of language data and argue in favour of using natural data. In other words, corpora are seen as supplanting other types of linguistic evidence.

In this paper, we will suggest that corpus data and introspective data (in particular, elicitation data), far from being opposed to each other, are in fact compatible, and that their

[2] Fonds National de la Recherche Scientifique, Centre for English Corpus Linguistics, Université catholique de Louvain
e-mail: gaetanelle.gilquin@uclouvain.be
[3] Centre for English Language Studies, University of Birmingham
e-mail: t.shortall@bham.ac.uk

reconciliation can prove enlightening for a field such as Foreign Language Teaching (FLT). More precisely, we will argue that corpora say something about performance, whereas elicitation gives information about competence, and that both performance and competence are necessary to get a full picture of the acquisition of a foreign language. Two case studies will illustrate this. The first one investigates learners' knowledge of *make*-collocations through a combination of (native and non-native) corpora and fill-in and evaluation exercises. It will be shown that, while learners' free production displays a number of problems, especially when contrasted with a corpus of native English, the true extent of their collocational deficiency is only revealed by the analysis of the more constrained data. In the second case study, we will compare the prototypical sense of the existential structure in elicited data with its most frequent sense, as attested in a corpus of spoken British English. We will claim that prototypes should serve as launching pads for learning and therefore as key initial points in the FLT syllabus, whereas performance-driven corpus data should be taught at later stages, when proto-based competence has already been established.

2. Case study I: The use of *make*-collocations by learners of English

As early as 1933, Palmer noted the difficulty combinations such as *to ask a question, to do a favour, to give trouble* or *to have patience* present for learners of English. Since then, many studies have been devoted to this problem, studying learners' collocational knowledge on the basis of corpus data (e.g. Nesselhauf 2005) or through a more controlled method of data collection (*cf*. Bahns and Eldaw 1993, who use translation cloze tasks). This case study, which investigates collocations with *make* in advanced French-speaking learners' interlanguage, relies on the combination of these two types of data, namely natural data extracted from native and non-native corpora, and elicitation data from fill-in and evaluation exercises. The two types of data are shown to be both necessary to gain a thorough understanding of learners' knowledge of *make*-collocations.

2.1 Corpus analysis

The learner corpus on which the study is based is ICLE-FR, the French component of the International Corpus of Learner English (Granger *et al*. 2002), which is made up of argumentative essays written by French-speaking learners, for a total of 202,957 words. ICLE-FR contains 469 occurrences of one of the forms of the lemma *make*, of which 171 are collocations. Incorrect *make*-collocations amount to 12, thus accounting for 7 percent of all the occurrences of a *make*-collocation (based on Borgatti 2006). Some examples are given in (1) to (3).

(1) In the first part of the novel, another activity takes place: Lily is **making a painting** but she cannot complete it. [ICLE-FR]
(2) Progressively, thanks to vivid **descriptions made** in a rich language (…), the picture of a society which is superficial comes before our eyes. [ICLE-FR]
(3) On the one hand, some people are still against the idea of Europe, or other people claim they are for union, but actually they **make separations** in their own country. [ICLE-FR]

Interestingly, of these 12 errors, 10 are potentially due to interference from the mother tongue (i.e. 83.3 percent). Let us consider example (1). French has only one verb to refer to *do*/*make*, viz. *faire*. It is therefore not surprising that French-speaking learners find it hard to distinguish between *do* and *make*, and sometimes use one verb instead of the other, as in (1). The data

contain three such cases. In (2), influence of French may also explain the learner's lexical choice, since both *faire* ('make') and *donner* ('give') can be used with *description* in French.

This error analysis suggests that French-speaking learners are not too bad at using *make*-collocations, and that their main problem is interference from the mother tongue. A comparison of the learner data with a control corpus of native English, however, makes it possible to go further than that. Using comparable data from LOCNESS-US, the American component of the Louvain Corpus of Native English Essays[4] (168,314 words), it appears that French-speaking learners tend to underuse *make*-collocations, in a way that is statistically significant, as appears from Table 1.

LOCNESS-US	ICLE-FR	X^2
123.58 (208)	84.25 (171)	13.95 (p<0.001)

Table 1: Relative frequency per 100,000 words (and absolute frequency) of *make*-collocations in LOCNESS-US and ICLE-FR (based on Borgatti 2006)

But it is not all *make*-collocations that learners underuse. In fact, if we perform a distinctive collexeme analysis (see Gries and Stefanowitsch 2004), with the aim of identifying the collocations that are more distinctive for learner English and those that are more distinctive for native English, it turns out that learners tend to underuse collocations that have no word-for-word equivalent in French, but overuse collocations which are directly translatable into French. This is very clear in Table 2, which gives an overview of the results of the distinctive collexeme analysis that are statistically significant.[5] Of the nouns that are more distinctive of native English (i.e. are underused by the French-speaking learners), only one has a direct equivalent in French, namely ***make*** *an error* (***faire*** *une erreur*). Of the nouns that are more distinctive for French-speaking learners (i.e. are overused by them), on the other hand, all have a word-for-word translation in French (e.g. ***make*** *progress* = ***faire*** *des progrès*, ***make*** *an effort* = ***faire*** *un effort*).

LOCNESS-US	ICLE-FR
Decision (6.19)	Progress (4.00)
Argument (2.66)	Effort (3.56)
Claim (2.12)	Use (3.51)
Case (1.32)	Distinction (2.44)
Error (1.32)	Step (2.44)

Table 2: Most distinctive nouns in *make*-collocations (LOCNESS-US vs. ICLE-FR)

This preference for congruent collocations (that is, collocations having a direct, word-for-word equivalent in French) is confirmed if we apply the technique of reversed translation, which consists in translating interlanguage back into the learner's mother tongue. As shown by Borgatti (2006), over 90 percent of the *make*-collocations used by French-speaking learners have a direct equivalent in French.

[4] See http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm (last accessed on 21 June 2007).

[5] The figure between brackets corresponds to the distinctiveness value (log-transformed *p*-value). The higher this value, the more distinctive the noun is for the group of speakers.

All this seems to suggest that French-speaking learners use *make*-collocations which sound familiar to them because they correspond to a *faire*-collocation in French, but avoid collocations which do not have such an equivalent, hence perhaps the overall underuse of *make*-collocations discovered in the learner data. Such performance phenomena, however, give only a very partial view of learners' actual knowledge of the collocations (i.e. their competence). We saw above that some 7 percent of the *make*-collocations found in ICLE-FR are incorrect, but we cannot say anything about learners' knowledge of the collocations that are **not** found in the corpus. In an attempt to come to a better understanding of learners' competence, the next section investigates elicitation data in the form of fill-in and evaluation exercises.

2.2 Elicitation analysis

The elicitation test was taken by 19 native speakers of French, all of them in their third year of English studies at the Université catholique de Louvain. In the first part of the test, which comprised 25 test items, the students were asked to fill in sentences with a verb of their choice, on the basis of the French translation provided for the sentence, *cf*. (4). They were also required to indicate their degree of certainty, using a scale ranging from 0 ("don't know the answer, made a guess") to 3 ("absolutely sure of the answer"). In the second part of the test, an evaluation exercise, the students were presented with 20 sentences and had to decide whether the underlined elements, corresponding to the collocation, were acceptable or not, *cf*. (5). Again, they had to indicate their degree of certainty, using the same 0-3 scale as in the fill-in exercise. In addition, the students were asked to correct, whenever possible, the sentences they judged unacceptable. All the sentences used in the test were authentic sentences, extracted from LOCNESS for the acceptable collocations and from ICLE for the unacceptable collocations.

(4) They were not even given time to _____ an offer.
= Ils n'ont même pas eu le temps de faire une offre.

(5) They wanted to <u>make an end</u> to these conflicts and maintain pacific relationships within Europe.

The picture that emerges from the analysis of the elicitation data is much gloomier than what the corpus data suggested. The error rates for the fill-in exercise and the evaluation exercise amount to 51 percent and 43 percent respectively (to be compared with the 7 percent error rate established in the free production data). However, these overall figures iron out important differences between the various test items. More precisely, the results show that there is a strong tendency for the learners to do much better with congruent collocations than with non-congruent collocations. Thus, in the fill-in exercise the congruent collocations were completed correctly most of the time (average of 89.5 percent), whereas for non-congruent collocations there were very few correct answers (average of 8 percent). The influence of the mother tongue is even clearer if we consider the incorrect answers provided by the learners. In (6), for example, 79 percent of the respondents chose the verb *take*, which is the equivalent of the verb used in the French collocation (***prendre** un engagement*).

(6) He refused to _____ any kind of commitment.
= Il refusa de prendre quelque engagement que ce soit.

The same tendency is observed in the evaluation exercise, where a majority of the students were able to judge the acceptability or unacceptability of congruent collocations, but

had more problems with the evaluation of the non-congruent collocations. Compare (7) and (8). The former contains a congruent collocation, with a *make*-equivalent in French (***faire*** *des promesses*), whereas the latter, which is incorrect in English, is a direct translation of the French expression ***faire*** *une difference* (i.e. 'make a distinction'). The results are very telling here. While 100 percent of the students correctly accepted (7), the same proportion accepted (8). Similarly, 68 percent of them did not seem to have any problem with the incorrect collocation *make abstraction of*, which is a literal translation of French ***faire*** *abstraction de* ('disregard').

(7) The candidate had made promises to local groups of voters on behalf of the government.
(8) Children are often unable to make the difference between fiction and reality.
(9) In business, one has to solve problems by counting, calculating and making abstraction of any emotional factors.

It should also be pointed out that the respondents were often unable to correct unacceptable collocations in the evaluation exercise. If we also consider those cases where the subject was unable, when necessary, to replace the incorrect collocation by an appropriate alternative, the error rate of the exercise rises from 43 percent to 60 percent.

Finally, it is interesting to examine the degree of certainty the subjects assign to their answers. One clear pattern is that congruent collocations tend to be assigned a higher degree of certainty. In the fill-in exercise, they reach an average degree of certainty of 2.05 (out of a maximum of 3), against 0.95 only for non-congruent collocations. In addition, the learners sometimes appear to be too optimistic or, in contrast, too pessimistic. *Make the difference*, for example, which is (incorrectly) accepted by all the subjects, has an average score of 2.6, with 12 subjects assigning it the maximum degree of certainty. This is even higher than the score for *make a promise*, which is accepted with an average degree of certainty of 2.2. On the other hand, it is not rare to see the learners assign a low degree of certainty to a correct answer. The subjects who judged *make a gain* as acceptable, for instance, did it with an average degree of certainty of 1.3 only (including one guess). A correct answer in the elicitation test, therefore, does not guarantee that the learner feels confident about his/her answer.

2.3 Discussion

The corpus analysis reveals a relatively low error rate in advanced French-speaking learners' use of collocations with *make*. However, it brings to light two major problems in learners' free production, namely an underuse of *make*-collocations and a clear influence of the mother tongue, both in the form of word-for-word translations from French and preference for congruent collocations to the detriment of non-congruent collocations.

The elicitation data confirm the role played by transfer in learners' collocational knowledge. They also provide a possible explanation both for the relatively low error rate and for the underuse of *make*-collocations (in particular non-congruent collocations) observed in the corpus data. Given that the choice of the verb in collocations is largely arbitrary (see Allerton 1984), learners arguably tend to rely on what they know best, namely the corresponding collocation in their mother tongues, hence the importance of (positive and negative) transfer. As a rule, learners are more familiar with congruent collocations, as appears from the higher average degree of certainty in the test. When writing free compositions, they seem reluctant to take risks, preferring to stick to those collocations which they feel safe with, that is, congruent collocations. Not only does it result in few errors, since learners just have to translate the collocation word for word into the foreign language, but it

also leads to underuse, since a whole set of collocations are avoided, namely those that are not congruent.

As Ellis and Barkhuizen (2005: 49) rightly point out, "no one method will provide an entirely valid picture of what a learner knows or thinks". Hence the importance of combining several methods, which each offer a different perspective on the knowledge of a foreign language. The use of elicitation data as a supplement to corpus data in the study of learners' knowledge of *make*-collocations makes it possible, not only to go beyond the relatively low number of errors found in free production, but also to explain some of the tendencies observed in the corpus. This is one way in which FLT can benefit from a reconciliation between corpus data and elicitation data. In the next section, we present the results of another case study which also illustrates the complementarity of these two types of data.

3. Case study II: Prototype and real-time language

In the first study, we saw how fill-in and evaluation exercises can give insights into the nature of underlying language competence of L2 learners and how this differs from authentic native-speaker performance.

In this second study, we continue to investigate the differences between competence and performance through a comparison of the existential structure in elicited data from native speakers and a spoken corpus. The first section offers a review of Prototype Theory (Rosch 1975, 1978), as this affects the analysis of the data, and the second section examines the variation that is to be found within the existential construction. We then examine whether prototype effects are to be found in the elicited data and how this compares with the existential structure as it appears in the Bank of English 20 million word spoken British English corpus (brspok). The implications for language teaching are discussed in the final section.

3.1 Categories and prototypes

Humans categorise events and objects, but also categorise abstractions, experiences, feelings, social relations and so on. And categories abound in our social structure. Illnesses are commonly categorised as contagious or not, life-threatening or not, curable or not. In law, courts are asked to categorise killings as murder, manslaughter, accidental, *etc*.

The most obvious type of categorisation, and one which plays an important role in language learning, is the organisation of items into categories around prototypes. Children readily learn the name of the best example of a category, e.g. *dog*, and from this prototype they create the category ANIMAL, and then extend the category to include other items such as *cat* or *horse*.

The notion of grammar in Cognitive Linguistics is tied very much to the categorisation principle. Just as people categorise objects like *table* into categories like FURNITURE, so too do they seem to categorise grammatical elements into various categories at different levels of complexity: objects and entities are placed in the noun class category, properties are coded as adjectives, and complex grammatical patterns are coded into a range of different construction categories. *Table* is a prototypical noun in that it is a concrete entity with clear reference, while *education* is a more peripheral non-concrete member. Similarly, *I ate dinner* is a good example of a prototypical past tense verb in that it is both past and punctual in time, while *I watched TV* is a more marginal member as it lacks punctuality even though it is past in time.

Each category has a prototype, or Cognitive Reference Point (Rosch 1978), and it is the prototype that people call to mind when asked to provide examples of a category, so that we often think of *table* or *chair* when asked to think of furniture, but are less likely to think of *sideboard* of *bookshelf* as examples of furniture.

Prototypes are "defined operationally by people's judgements of goodness of membership in the category" (Rosch 1978: 36) so that knowledge of prototypes involves knowing, for example, that a robin is more 'birdy' than, say, a chicken or a penguin (as these are 'flightless'). Rosch (1975a, 1975b, 1976) conducted a series of experiments which clearly demonstrated the prominence of prototypes in the minds of her subjects; however, these experiments were restricted to categories of concrete items such as furniture, fruit or vegetables.

Some descriptive work has also been carried out at the level of language constructions, and Taylor (1995: 197) suggests that "constructions, no less than other kinds of linguistic object, also need to be regarded as prototype categories, with some instantiations counting as better examples of the construction than others". Every category has a prototype, with extensions, in Langacker's (1987) terminology, being less representative members of the category. Often strength of membership of a category can be seen as the extent to which members comply with a set of features. The preposition category IN, for example, carries two features, +concrete and +delineated, with different members of the category carrying different feature attributes (Figure 1).

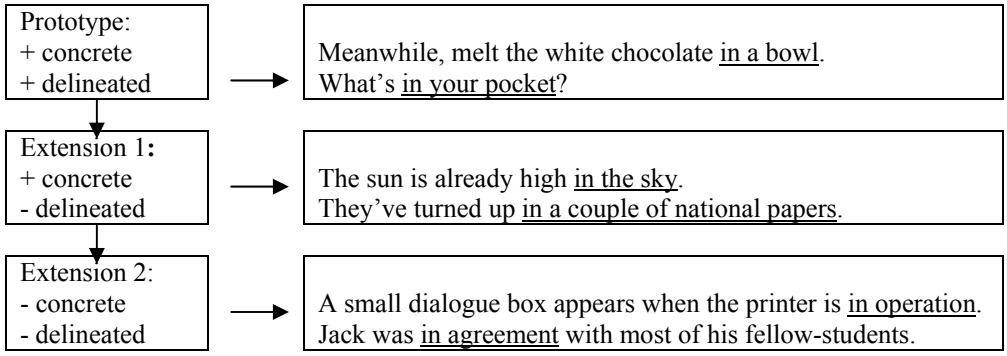

Figure 1: Prototype and extensions for IN (examples from brspok)

Tense can also be seen as a category with prototypes. A standard grammatical description of past tense suggests that "[t]he Past [tense] refers to a definite event or state that is seen as remote in time or as unreality or for reasons of politeness" (Downing and Locke 2002: 353). The categorial description is similar: the past tense prototypically assigns an event or state to some point in time prior to the moment of speaking or writing (Langacker 1987, Taylor 2002); extensions from the prototype include counterfactuality, i.e. the unreality of a state or event, and pragmatic softening, i.e. the cushioning of the impact of an utterance (Figure 2).

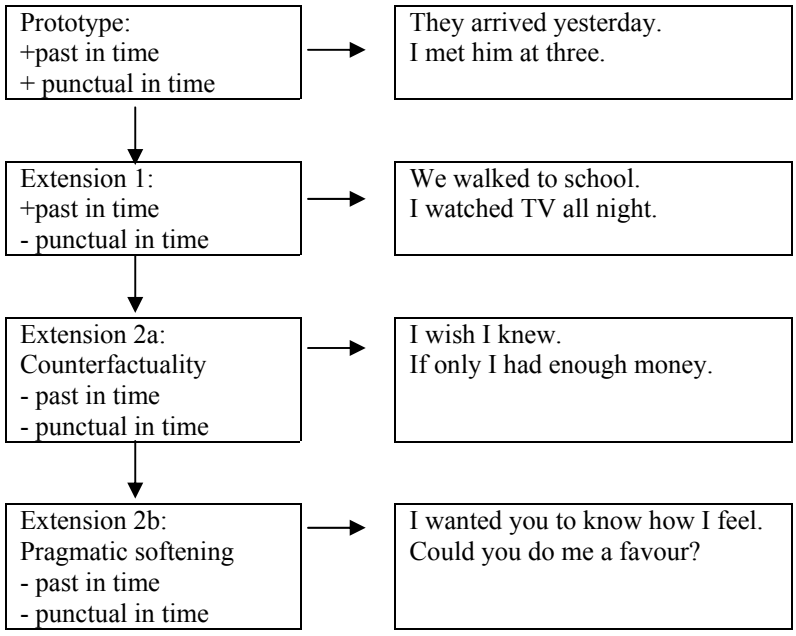

Figure 2: Prototype and extensions for past tense

In the next section, we take a brief look at the kind of variation to be found in the existential structure.

3.2 The existential construction

The only corpus description carried out to date of the existential construction is by Sasaki (1991). The existential pattern consists of there+V+NP+X, whereby the variation within the category focuses around 'X', which may be a PP, VP, that-clause, and so on. Sasaki's work involved identifying the different variations of the there-construction and detailing the frequency of occurrence of these in corpora taken from three sources: Informal Conversation, Radio, and Narrative American English (the UCLA Oral corpus of approximately 140,000 words).

Sasaki (1991: 168) has identified five existential variations. We quote below her descriptions, and also her examples.

1. There + be + NP + post-NP modifier(s) (Type PM)

   The post-NP modifiers in this category can be relative clauses, adjective phrases, infinitives, or prepositional phrases.

```
There are not too many things that will do very very well (relative clause)
There is something special (adjective phrase)
There are some special ways to cut the climbing roses (infinitive)
Is there any problem with it? (prepositional phrase)
```

2. There + be + NP (Type BA)

   This type of NP *there* sentences is called "bare" because it does not have any post-logical subject element.

They tend to get pretty sad looking if there's cold weather (Bare NP)

3. There + be + NP + expression of place (Type PL)

This type has generally been regarded as a "prototype" of the *there* sentences. The expression of place can be an adverbial or a prepositional phrase. [Sasaki gives no indication or evidence as to why this pattern should be considered the prototype.]

There's lots of 'em here
There are some restaurants in this town.

4. There + be + NP + adverbial phrase (Type AD)

This category contains adverbial phrases other than the expression of places as the post-logical subject elements.

Was there a fire a couple days ago?

5. There + be + NP + participle (Type PA)

Either a present participle or a past participle as in:

there couldn't have been enough water coming in
There was just so much money appropriated

Sasaki's description sees the existential construction as having five variations. The analysis of data from the Bank of English led us to make eight classifications; this difference may be because Sasaki is largely concerned with a functional classification, while we are also interested in structural variation (e.g. locative adverbials and locative PPs were classified as different items).

Lakoff (1987: 549) also notes a number of patterns in the existential construction. These are shown in Table 3, with the terminology Lakoff uses, and the examples he gives, along with the syntactic patterns we have used in the analysis of the brspok data.

EXAMPLE	LAKOFF TERM	SYNTACTIC PATTERN
There isn't anyone taller than Harry.	Adjective phrase	There+be+Comp
There is someone in the yard.	Locative phrase	There+be+NP+PP
There was no one with his shirt on.	Special *with*-phrase	There+be+NP+PP
There is a concert at noon.	Temporal phrase	There+be+NP+PP
There is a man about to leave.	*About to*-phrase	There+be+NP+PP
There wasn't any money stolen.	Passive phrase	There+be+NP+VP-ed
There's a boy running away.	Progressive participial phrase	There+be+NP+VP-ing

Table 3: Lakoff's existential phrase variations

EXAMPLE	PATTERN
there is a very wild passionate kinky lover in this man	There+be+NP+PP
There's so much joy so much peace so much blessing here	There+be+NP+Adv
There is no right or wrong answer	There+be+Bare NP
there is no er net asset created in the private sector there's another conveyor belt running in the opposite sense there's still a bit to do on that	There+be+NP+VP (-ed/-ing/inf)
There's a couple who are really naughty	There+be+NP+Wh-clause
there's no doubt we will beat them	There+be+NP+that clause
there's nothing worse than a wobbly desk	There+be+NP+Comparison
There's no such thing in life as a free lunch	Idioms

Table 4: Brspok examples of the existential construction

Table 3 shows that Lakoff has distinguished between four different phrase types (Locative phrase, Special *with*-phrase, Temporal phrase, *About to*-phrase) which have been collapsed into a single pattern here, viz. there+be+NP+PP. Similarly, he has distinguished between Passive phrase and Progressive participial phrase, both of which are considered here to be variants of the there+be+NP+VP pattern (i.e. VP-ed and VP-ing). This means that Lakoff's seven patterns are equivalent to only three from our own inventory. Our own list, taken from an analysis of brspok, is more extensive, with eight variations of the existential construction (Table 4) and includes patterns not found in Lakoff or Sasaki (all examples in Table 4 are taken from brspok). This analysis is based on just over 2,000 randomly sampled examples, around 15 percent of the total number found in brspok, the 20 million word sub-corpus of the Bank of English consisting of interviews, speeches and spontaneous interactions among speakers of British English.

3.3 Informant data for the existential structure

In this section we investigate whether any prototype emerges in existential sentence data elicited from 35 native-speakers of English.

3.3.1 Informants

The informants of this study were thirty-five native speakers of English. All were high school students in Ireland. This group was chosen because, being neither language teachers nor linguists, there would be no dangers that any answers they gave would be influenced by prior knowledge of the linguistics of there-constructions.

3.3.2 Research technique

Informants were asked to write out five sentences beginning with *there is/are*. The survey was carried out in the informants' high school, and was administered by the second author of this paper. The objective of this exercise was to check informants' instincts regarding the there-construction. This kind of elicitation is widely used in semantic memory research in psychology, and informant output in this kind of research is "normally taken to reflect some aspect of storage, retrieval, or category search" (Rosch 1978: 38). Rosch *et al.* (1976) have also shown that prototypes of categories are produced first and more frequently in this kind of research (although their research was restricted to artificial categories representing concrete objects).

3.3.3 There-construction data from native speakers of English

THERE+BE+X	NO. OF SENTENCES	PERCENTAGE
there+be+NP+PP	117	67.63%
there+be+NP+VP	30	17.34%
there+be+Bare NP	8	4.62%
there+be+NP+Comparison	6	3.47%
there+be+NP+that clause	5	2.89%
there+be+NP+Adv	4	2.31%
there+be+NP+wh-clause	2	1.16%
there+be+NP+Idiom	1	0.58%
Total	173	100.00%

Table 5: Structural patterns produced by subjects

Table 5 shows the there-construction sentences produced by the thirty-five informants. Each informant was asked to produce five sentences, but one informant only produced four sentences, and one sentence, *There is, is there, yes there is,* was unclassifiable and therefore excluded, leaving a total of 173 sentences. At 68 percent, the there+be+NP+PP pattern predominates; there+be+NP+VP ranks second at 17 percent. All other patterns account for only 15 percent of all sentences. The spread of the there+be+NP+PP pattern was wide: one informant produced this pattern only once, one student only twice, with all other informants producing the there+be+NP+PP pattern in between three and five of their sentences. This indicates that the there+be+NP+PP frequencies are a product of the group as a whole, and that there are no individuals whose production is skewing results.

Examples of informant-produced sentences are shown below (the X patterns are underlined):

there+be+NP+X	Examples	
there+be+NP+PP	There is	not one nice looking bloke in this school.
	There are	grapes in Moore Street every day.
there+be+NP+VP	There are	ways and means to do what has to be done.

	There is	a man screaming.
	There is	a great movie called The last of the High Kings.
there+be+NP+NP	There is	no correct answer, it is a matter of opinion.
	There is	no food.
there+be+NP+wh-clause	There is	a cat with one leg that lives down the road.
	There is	nothing here that I want.
there+be+NP+Comparison	There is	more trouble in the North than there was years ago.
	There is	a lot more Irish than English in Liverpool.
there+be+NP+ADv	There is	a football match today.
	There is	another train later, don't worry.
there+be+NP+that clause	There is	no way we can stop now.
	There are	ten CD's I want to buy.
there+be+NP+Idiom	There is	no place like home.

Two of the three set expressions involve comparisons, and so could also be listed with the comparison group:

```
There is no fool like an old fool.
There is no place like home.
There is method in her madness.
```

The ranking of items in the there-construction category in Table 5 suggests the there+be+NP+PP pattern as being the mental representation of the there-construction for a considerable majority of our informants, and it seems reasonable to posit this structure as the prototype of the structure on the basis of the data we have looked at. In the next section we examine the occurrence of existentials in brspok.

3.4 Existential data from spoken British English corpus

The brspok data consisted of 2,378 lines of concordances, representing a random sample of 15 percent of all occurrences of the existential construction. The search involved keying in there+is and there+are, producing 7,982 concordance lines for the former and 7,698 for the latter.

3.4.1 Research procedure

The concordance lines were examined for different variations in the existential grammar patterns. The objective was to make a comparison of these results with those obtained for the elicited data.

One hundred lines were not analysed as these were unclassifiable. Typically, these lines were full of hesitations and slips of the kind normally found in spoken discourse:

```
hink. Erm I think that there's there's a lot of the er the the
there's there's there's it's sort of simi similarity but then you' ean 'cos
there's a lot of ch I mean it's like with messy play as
```

3.4.2 Existential patterns in brspok

Table 6 shows the frequency of the existential patterns in brspok. Although the there+be+NP+PP pattern predominates, it does so only to a small extent. There+be+Bare NP is only slightly below there+be+NP+PP in the ranking, and both there+be+NP+VP and there+be+NP+wh-clause also have substantial presence. Overall, there is a much more even distribution of patterns in brspok than in the elicited language data.

Pattern	No.	%
There+be+NP+PP	769	33.76%
There+be+NP	666	29.24%
There+be+NP+VP	305	13.39%
There+be+NP+wh	292	12.82%
There+be+NP+Adv	135	5.93%
There+be+NP+that	94	4.13%
There+be+NP+Comp	17	0.75%
There+be+NP Idiom	1	0.04%
Total	2278	

Table 6: Existential patterns in brspok

Examples of the different sentence patterns are shown below.

There+be+NP+PP:

The most frequent nouns in this pattern were *way, lot, problem* and *point*:

I I I don't think there's any way out of it now they've got one ably it's
solvents. I # mean there's been a lot in the press recently about Ecstasy
Mr MX says until now there's been no problem with the mortgage for four years
do. Erm there's there's one point erm on the form it sort of says er asks

There+be+Bare+NP:

In the Bare NP pattern, there is no adverbial, prepositional phrase, or other postmodification. These 'missing' elements are often part of shared knowledge and are mutually understood by the interlocutors, and therefore do not need to be made explicit. In the first example below, for example, the speaker feels that it is clear to the listener where there is no space, so there is no need to make this explicit:

no I'm not being obje I mean there's no space. All right. Unless you have it
about black people and There's lots of opportunity Yeah. in town
ression Do you think there's a community spirit? I've probably # asked

The Bare NP pattern also often has a dislocated PP, adverbial, VP, *etc*. In the first extract below, the dislocated element (in bold) is a VP; in the second extract the dislocated element is

a PP. This kind of dislocation does not occur in the elicited language data as this was elicited at sentence level.

> Iceland. <F0X> Iceland. <F0X> Oh <ZGY> MX brought that. <F0X> Mm. <F0X> Oh they're reasonable are Iceland. <F0X> Does anybody want more wine? Is there any **left** or <ZGY> <M0X> Yes. There's some more wine. Yes. <M0X> Anybody want any? <F0X> We've no <ZGY> at all. <ZZ1> <!--unintelligible--> <ZZ0> <F0X> No I haven't. <ZG1> I hide them all. <ZG0> <F0X> I'll have some more if there's any going.
>
> be taking their leadership and advice and guidance from you. Now that is the way that it may well be. Er but out of the words er **out of the mouths of babes and children** frequently there is some [old] wisdom. And I'm not necessarily saying that a younger person or younger people erm especially if we are talking here about a a a younger man rather than a younger woman may inadvertently erm

There+be+NP+VP:

Three types of VP were evident in the corpus, infinitive, -ing form and past -ed form:

> on't mean to be flippant but there's no nice way to kill somebody in a war.
> ed 381-418 Good. There's a lot going on there isn't there
> eare's heroes the sense that there is nothing left in the world at all and

There+be+NP+wh-clause:

Different wh-words appear in the relative clauses: *that, who, which*:

> ment. And there's one other thing that affects this for us.
> Yeah but there's a load of people who only come when they think there
> eems. And there's a whole load of areas which have got to be looked at

There+be+NP+Adv:

Both adverbials of time and space occur:

> Yes. there's no problem there. But er erm a a lot of
> what they're saying is there is a risk here <tc text=pause> there there
> Erm there's a very big Sunday School now. And it's nice 'cos
> unity spirit. Do you think there's still community spirit today

There+be+NP+that:

The NP that-clause appears with a small number of highly frequent nouns like *way, doubt, reason*:

> Yeah. because there's no way that they were doing thirty miles an hour.
> vation is there. There's no doubt that he is one of our most talented players
> ver is that that there's some feeling that syllabus means course content

There+be+NP+comparison:

The fixed phrase *no such thing* was common among existentials with comparisons:

> ollisions that occur. I mean there is no such thing as an accident really.

But now it's all different there's no such thing as a Black Friday. They

Other comparisons were more varied:

<F01> And there's nothing worse than a wobbly desk. <M02> No than people apparently there's eight times more sheep than there are people in New

There+be+Idiom:

There was one example of an idiomatic expression, in a comparison phrase:

American adage isn't it there is no such thing as a free lunch. Well the

3.5 Discussion

In this second case study, we noted that the elicited data showed strong prototype effects, with the prototype there+be+NP+PP occurring in almost 70 percent of all sentences. No such strong effects are seen in the corpus data. While the prototype dominates, it only occurs in 34 percent of sentences, and is only 4.5 percent above the Bare NP pattern. In addition, there is also a strong showing for other patterns.

The overall impression here is that language competence and language performance are related but different phenomena. Language structures seem to be organised in the mind as categories with strong prototypes. Language usage, on the other hand, seems to involve the use of a wide variation of extensions. The prototypes of language constructions seem to form a basic pattern, a mental representation that is our Cognitive Reference Point for the structure. Real-time communication appears to be variation on this theme, i.e. extensions from the prototypes of categories. Ongoing research into the existential structure in other languages, namely Portuguese, Mandarin and Japanese, is producing similar results, suggesting that prototypes of language structure categories may be universal.

The implications of the above findings for language teaching may be of some importance. If prototypes are our Cognitive Reference Points, then these are what learners will expect to find when they come to study a second language. This being the case, prototypes of structures should be taught first in second language programmes. At the same time, corpus evidence suggests that there is wide variation of structural use, i.e. extensions, in real-time communication. It may therefore be valid to propose a cyclical language syllabus, with prototypes taught first, and structures being revisited in a cyclical fashion as proficiency increases, with all extensions of a structure being eventually introduced to students.

4. Conclusion

Using two case studies as an illustration, this paper has shown that, while corpus data and elicitation data may present marked divergences, they are in fact complementary and their combination may prove worthwhile for Foreign Language Teaching. More precisely, corpora give access to performance, whereas elicitation gives access to competence, and this twofold approach makes it possible to get a more comprehensive picture of the acquisition of a foreign language.

The power of corpus data to make new findings about language should not obscure the fact that other methods and other sources of data are also available and may have important

contributions to make to the study of language. Each type of linguistic evidence has its advantages, but it also has its own limitations. By combining different sources, one may capitalise on the strengths of each source and make up for its weaknesses, thus gaining a better understanding of the phenomenon investigated.

References

Allerton, David J. (1984) Three (or four) levels of word cooccurrence restrictions. *Lingua 37*, 17–40.

Bahns, Jens and Moira Eldaw (1993) Should we teach EFL students collocations? *System 21*, 101–114.

Berry, Roger (1994) "Blackpool would be a nice place unless there were so many tourists": some misconceptions about English grammar. *Studia Anglica Posnaniensia 28*, 101–112.

Borgatti, Edwin (2006) *The Use of the Verbs 'Make' and 'Do' by French- and Dutch-Speaking EFL Learners*. Unpublished M.A. Thesis. Université catholique de Louvain, Louvain-la-Neuve.

Downing, Angela and Philip Locke (2002) A University Course in English Grammar. London and New York: Routledge.

Ellis, Rod and Gary Barkhuizen (2005) *Analysing Learner Language*. Oxford: Oxford University Press.

Granger, Sylviane, Estelle Dagneaux and Fanny Meunier (eds) (2002) *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Lakoff, George (1987) Women, Fire, and Dangerous Things. Chicago: University of Chicago Press.

Langacker, Ronald W. (1987) Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites. Stanford: Stanford University Press.

Nesselhauf, Nadja (2005) *Collocations in a Learner Corpus*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Palmer, Harold E. (1933) *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Roland, Douglas and Daniel Jurafsky (2002) Verb sense and verb subcategorization probabilities, in Paola Merlo and Suzanne Stevenson (eds) The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues, pp. 325–46. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Rosch, Eleanor (1975a) 'Cognitive reference points'. *Cognitive Psychology 7*, 532–47.

Rosch, Eleanor (1975b) 'Cognitive representations of semantic categories'. *Journal of Experimental Psychology*: *General 104*, 192–233.

Rosch, Eleanor (1976) 'Structural bases of typicality effects'. *Journal of Experimental Psychology: Human Perception and Performance 2*, 491–502.

Rosch, Eleanor (1978) Principles of Categorization, in Eleanor Rosch and Barbara B. Lloyd (eds) Cognition and Categorization, pp. 27-48. Hillsdale, NJ: Erlbaum Associates.

Rosch, Eleanor, Carolyn Mervis, Wayne Gray, David Johnson and Penny Boyes-Braem (1976) Basic objects in natural categories. *Cognitive Psychology 8*, 382–439.

Sasaki, Megumi (1991) 'An analysis of sentences with nonreferential *there* in spoken American English'. *Word 42*, 157–78.

Taylor, John R. (1995) Linguistic Categorisation (Second Edition). Oxford: Clarendon Press.

Taylor, John R. (2002) Cognitive Grammar. Oxford: Oxford University Press.

Teachers' Lexical Intuitions Versus Corpus Data: Differences, Similarities and Explanations

Dr Iain McGee[1]

Abstract

In this paper I discuss the different explanations which have been forwarded to explain why lexical intuitions may differ from corpus data, explanations forwarded by both corpus linguists (e.g. Sinclair 1991) and psycholinguists (Wray 2002, Bybee and Hopper 2001). In addition, relevant research from word frequency estimation studies and word association studies are considered for the light that such data can shed on this subject. An experiment is then reported, designed to compare BNC data and EAP teacher intuitions about the most frequent collocates of some fairly common adjectives. The results indicate, perhaps surprisingly, that intuitions are, at times, very similar to the BNC data in a statistically significant way. However, at other times, the intuitions are quite different from the corpus data. The data are considered in the light of the theories previously discussed, and it is argued that they add support to the view that a key factor affecting the 'quality' of lexical intuitions may be the employment of an 'availability heuristic' in judgements of frequency. It is argued that some collocates of words (particularly those typically occurring together with the stimulus word in a larger language chain) may be more hidden from memory searches than other collocates which tend to occur with the stimulus word as a dyad.

1. Background

It is not particularly surprising that in the increasingly specialized and fragmented world of research that academics from a particular discipline may be unaware of studies from other specialist fields which may, be it directly or indirectly, touch upon their own research interests. A case in point, and the focus of this paper, is the debate concerning intuitions about language use and corpus data – actual records of language use. Many corpus linguists have challenged the reliability of language intuitions: as opposed to actual language data that can be observed in a corpus, intuitions are deemed subjective and undisciplined, being termed by one authority, "our random and incomplete access to our experience of language" (Cook 1998: 59). The actual evidence forwarded to support such claims is rather thin: typically being either anecdotal (Beaugrande 1996: 523; Renouf 1997: 259, 260) or indirect (Hunston 2002: 21; Willis 1990: 49, 55, 124). On the other hand, there is a large body of relevant research data readily available in two specialist areas of psychology - word frequency estimation research and word association research. Such research has spawned theories to explain the data and when such theories are considered alongside models of the mental lexicon, a fuller and more rigorous understanding of the intuition-corpora debate can be appreciated.

Before proceeding further, it is helpful to highlight what exactly some corpus linguists have said about the 'intuition-corpus data' divide. Hunston (2002) mentions four areas of intuition weakness (frequency, collocation, semantic prosody and phraseology) though here I

[1] Department of English Language Studies, King Fahd University of Petroleum and Minerals, Saudi Arabia
mcgee@kfupm.edu.sa or
e-mail: idmsjm96@muchomail.com

shall focus on just the first two of these. Regarding frequency, Hunston argues that, "It is almost impossible to be conscious of the relative frequency of words, phrases and structures except in very general terms" (2002: 21). Other challenges to word frequency estimation (particularly the identification of the most frequent use of a particular word) have been made by Willis (1990), Renouf (1997) and Kennedy (1991). With regards to the second of the language intuition weaknesses noted by Hunston, collocations, Biber *et al.* (1996: 120) believe that, "Intuitions regarding lexical associations are often unreliable and inaccurate" and Stubbs (1995: 24) argues that although examples of collocates can be given "sometimes accurately", on the whole the production of collocates on demand is weak: "[native speakers] certainly cannot document collocations with any degree of thoroughness, and they cannot give accurate estimates of the frequency and distribution of different collocations" (1995: 24, 25). The problem with such views is that they lack a strong empirical base: they are founded on hunches, and omissions from text books: indeed, they could be described as intuitions about intuitions. In the section that follows I focus on two fields of relevant research to help investigate this subject further: firstly, investigations of word frequency estimation and secondly, word association research.

2. Tapping into the relevant research

2.1 Word frequency research

Research conducted by psychologists into word frequency estimation over the past four decades has suggested that word frequency estimates are accurate, i.e. that native speakers of the language can rank words according to their relative frequencies in language, or estimate how often they are used (e.g. Tryk 1968, Shapiro 1969, Carroll 1971, Backman 1976, Frey 1981, Ringeling 1984, Arnaud 1990, Desrochers and Bergeron 2000, and Balota *et al.* 2001).

Some researchers who have investigated word frequency estimation skills have espoused diametrically opposed suppositions to those of some corpus linguists who question frequency estimation abilities. For example, Tryk states:

> This study was generated by the assumption that individuals are able to give valid and reliable reports reflecting the degrees to which they have processed given words. That is, it was assumed that people carry with them a kind of subjective 'yardstick' of word frequency enabling them to measure the magnitude of words on a dimension of word frequency, much as they give quantitative estimations of perceived intensity, length, duration, and numerosity in psychophysics (1968: 170).

Indeed, when differences were found between corpus data and subjective frequency estimates (SFEs) it was even suggested by some of the above noted researchers that the corpus was 'the problem', not the SFEs (Carroll 1971: 728; Frey 1981: 401; Ringeling 1984: 68).

Digging a little deeper into frequency estimation abilities, two distinct factors should be considered. They are the representation of frequency in memory and the ability to access this information. Regarding the former of these, indirect coding theory posits that it is not frequency per se which is coded, but the traces of an event which are recorded. The repetition of the event leads to a trace being multiplied or simply strengthened over time. In this model frequency information is different from 'normal' propositionally encoded information (e.g. that Jack's birthday is in March). One of the reasons why some researchers doubt that frequency is coded directly (i.e. like 'normal' propositional information) is that if this were so,

the encoding would be optional. Part of the distinctiveness of frequency information, Hintzman (1978: 548) argues, lies in its being 'obligatory', i.e. the coding of frequency is an automatic process.

Assuming then, that frequency information is automatically encoded, the second factor that should be considered is how the frequency information is accessed. Brown (1995: 1540) believes that a variety of strategies can be employed when trying to access frequency information. In the area of word frequency estimation research, non-enumeration memory assessment strategies are considered to be the most relevant estimation technique, and heuristic strategies are the most important of these. Heuristics have been defined as, "strategies that simplify complex tasks and get the job done well enough – they don't optimize they do 'satisfice'" (Cosmides and Tooby 1996: 11). Tversky and Kahneman's (1973; 1982: 18) view is that three heuristics are employed in judgements under uncertainty: availability, representativeness and anchoring and adjustment. Of particular interest to us is the availability heuristic. Tversky and Kahneman explain how this operates in the following way:

> A person could estimate the numerosity of a class, the likelihood of an event, or the frequency of co-occurrences by assessing the ease with which the relevant mental operation of retrieval, construction, or association can be carried out. A person is said to employ the availability heuristic whenever he estimates frequency or probability by the ease with which instances or associations could be brought to mind (1973: 208).[2]

They go on to note that availability, while positively related to frequency (i.e. what is more frequent is more available), is also affected by other factors, (e.g. salience) and such factors may affect how frequent an event appears to be (1973: 207, 208; 1982: 11). Tversky and Kahneman (1982: 11) note, for example, that seeing a house on fire, (as opposed to reading about a house burning down) is likely to affect one's ideas about how common or rare such an event is, and Taylor (1982: 192) explains this idea in the following way: "Salience biases refer to the fact that colorful, dynamic or other distinctive stimuli disproportionately engage attention and accordingly disproportionately affect judgement".

So, researchers investigating subjective frequency estimation start out with the opposite assumption of corpus linguists: that frequency estimates are reliable. However, as noted, there is an important qualification too – estimates are only as good as the availability of the key information required to make the judgement. Estimates may be wrong, either because some relevant information may be less available than other relevant information, or other information may be more 'salient' than it is 'frequent'.

2.2 Word association research

The frequency estimation research referred to above does not investigate production ability. We must refer to word association research to help investigate this type of knowledge. In perhaps the most well-researched type of word association test – free association testing - the subject is required to provide the first word that comes into his or her mind when provided with a stimulus word. Psychologists and psycholinguists believe that such data are valuable in helping us understand how words are connected in the mind. Groot (1989: 824), for example, terms word association responses, "relatively pure indicators of the way human knowledge is

[2] Note also N. Ellis (2002: 317): "We have no conscious access to the frequencies represented in our language processing systems, so we have to generate some exemplars in order to scrutinize them". The ease of generating those exemplars is the distinctive contribution of Tversky and Kahneman's availability theory.

mentally represented". Collocation type responses are fairly common in free word association tests; however, because more interest has been shown in analyzing and classifying paradigmatic type responses, there has been little investigation into what type of collocates are produced. The reason why syntagm responses have been largely 'overlooked' is that paradigmatic responses are viewed as being more typical and because syntagm responses are viewed by many researchers as not fully 'mature' responses (e.g. Carter 1987: 158; Söderman 1993: 157; Meara 1980: 235). In my own analyses of 'primary' collocate responses (i.e. stereotypical responses) from the Moss and Older (1996) data I have identified seven types of relationship between the adjective-noun and noun-adjective responses. It should be noted that it is far more typical for an adjective to elicit a noun, than for a noun to elicit an adjective in a free association response. Below the classes are noted.

- The adjective is *the* salient feature of the noun or the noun is the adjective stereotypically (e.g. *green grass*).
- The adjective is the opposite of the stereotypical feature of the noun (e.g. *blunt knife*).
- The resulting noun is a compound noun, i.e. it has a specific meaning with the adjective and a separate dictionary entry (e.g. *broad beans*).
- The combination is an idiom (e.g. *humble pie*).
- The resulting collocation is 'restricted'. The adjective qualifies the noun from a small number of possibilities (e.g. *juvenile delinquent*).
- The adjective is a 'polar' quality of the noun (e.g. *blond girl*).
- The collocation is a quotation (e.g. *silent night*).

It should be noted that subjects in free association tests are not being asked to produce frequent collocates. They are just asked to produce the first word that comes to mind. However, it is sometimes the case that the primary collocate response, i.e. the most stereotypical response, produced by the subjects, is actually the most frequent collocate of the word according to corpus data. For example, at least 20 percent of the subjects in the Moss and Older data provided the following responses (noun in response to the adjective): *blond hair, blunt instrument, candid camera, classical music, cosmetic surgery, curly hair, elastic band, fertile soil, merry Christmas, old man, shallow water, straight line, tepid water, tidy room, wavy hair*. According to the BNC the nouns in these collocations are the most frequent noun collocates of the adjectives. Many of these collocations would be classified as 'restricted', i.e. the adjective typically co-occurs with very few collocates. These data provide empirical support to Fox's (1987) belief that typical collocates of words in restricted collocations can be provided by native speakers.

Clark (1970) is one of the few writers who has tried to make sense of the different syntagmatic responses in word association tests. He argues that two rules deal with the bulk of the syntagmatic responses. The first is what he terms 'the selectional feature realization rule'. For example, the word *young* has selectional restrictions, i.e., it is used to describe animate things that are not adult. This being so, Clark argues that syntagmatic responses to this word in free word association tests (e.g. *boy, child, etc.*), simply 'realize' the above noted features. Of course, many adjectives (particularly frequent ones) do not have such narrow selectional restrictions as the *young* example provided above. They occur in attributive position before a wide range of nouns (Biber *et al.* 1999: 509) and so how important this 'rule' is, in helping us

analyse syntagmatic associations is not very clear. The second rule that Clark forwards to explain syntagmatic responses is 'the idiom-completion rule'. Clark explains how this rule works in the following way: "Find an idiom of which the stimulus is a part and produce the next main word" (1970: 282). It is unclear what exactly Clark means by the term 'idiom' in the quote above as the examples he forwards (i.e. *cottage cheese, white house, so what, ham eggs, stove pipe, justice peace, how now, whistle stop*) are quite a 'hotchpotch' of combinations (including idioms, restricted collocations, sayings, *etc.*). However, it can be seen that my own categorization of syntagm responses from the Moss and Older data, provided above, do concur, to a certain extent, with the second of Clark's rules.

3. The mental lexicon and its role in affecting intuitions

It is at this point that we need to consider the mental lexicon in the light of the above research, and insights that might, together with the above information, help us put some of the jigsaw pieces together. Before looking at two specific and developed theories, an explanation forwarded by a corpus linguist to explain the corpus-data elicited data mismatch is noted.

Some corpus linguists have forwarded the idea that intuitions about frequent collocates are governed by the fully lexicalized meaning of a stimulus word. For example, Sinclair (1991: 113) believed that the core meaning of a word for people would be 'the most frequent independent sense'. As a result of this, we would expect there to be a mismatch between people's views about how words are used and corpus data. To put this into the availability heuristic theory framework terminology, delexicalised senses of words would be less available in searches of the lexicon because the fully lexicalized sense is more salient. As a consequence, the collocates that respondents might suppose to be frequent collocates of a particular word may actually be infrequent, as the 'wrong' sense of the stimulus word has been drawn on in making the judgement.

3.1 Mental lexicon theories - Bybee

What is it that might make some collocates less available when searching for typical collocates of a word? Bybee, and her colleagues and followers (e.g. Bybee and Hopper (2001), Sosa and MacFarlane (2002), Nordquist (2004)) have forwarded a model of the mental lexicon, which, inter alia, might help explain this. Bybee believes that frequency of co-occurrence leads to chunking – i.e. frequent collocations take on a form of their own in the lexicon. As a result, the component parts of the 'chunk' may become autonomous from the constituent parts. For example, Sosa and MacFarlane (2002), following Bybee's theory, argue that the frequent combination *kind of* may become autonomous from the constituent units *kind* and *of* (2002: 234). In the area of lexical collocations, Nordquist (2004) has forwarded the same kind of argument. Nordquist, required twenty-five students to provide three sentences (orally) for each of eighteen stimulus words. With regards to the stimulus *small*, she notes that the attributive collocates provided by her subjects were quite different from the typical noun collocates noted by Biber *et al.* (1998), which she notes are often 'quantity' collocates (e.g. *amount, piece, sum, etc.*). With regards to this difference she comments, "The specific, high frequency *small*-NOUN dyads…are likely to have separate storage in the lexicon. If highly entrenched, these phrases will have lost connections to lone *small*….decreasing the likelihood of being accessed in the elicitation task" (2004: 220).

It should be noticed that the key factor affecting 'hiddenness' in the 'Bybeean' theory is frequency of co-occurrence. This theory can deal with the available word association data,

as the frequent collocates provided in such data are to relatively infrequent stimuli: the combinations are actually quite rare.

3.2 Mental lexicon theories - Wray

Wray defines formulas as holistically stored multi-word items, which, rather than having been fused together, have been stored 'word-like' *from the beginning* (2002: 138; see also Peters 1983: 89; Widdowson 1989: 131). From the formula there may be subsequent analysis, i.e. the language may be broken down (i.e. segmented), but this only happens when needs require it (Wray 2002: 122, 130); it is not a default operation. Wray argues for holistic (non-analysed) representation of a considerable amount of language, not just/only language that is frequent or irregular. In this respect, her theory is very different to that of Bybee's.[3] Wray believes that collocations may be stored holistically, giving *major catastrophe* as an example (2002: 206-209). She argues that this collocation would be both, "noticed and remembered as a sequence", for the native speaker. For Wray this means that the individual components are not analysed and that the collocation is stored holistically with its associated meaning. Wray does allow for collocations to be analyzed and broken down by the native speaker when necessary or desired (2002: 211) believing that segmentation of a formula will occur, "where the word occurs in a context of actual or potential paradigmatic variation" (2002: 277).

Contrary to the view of Cook, noted earlier on page 1, Wray suggests that intuition is "a legitimate indicator of lexical organization" (2002: 281), arguing that her model can explain the differences between intuitive knowledge and corpus data. She believes that 'patterns of knowledge' (i.e. intuition) cannot be equated with 'patterns of use' (i.e. evidence available to us from corpora) because the former accesses only a subset of the latter (2002: 277). For Wray, intuition is incomplete in a principled way: the constituent parts of an unanalyzed unit are not as available to us as more analysed units: what is segmented is more analysed and therefore more available. Accordingly, if respondents were asked to use *by* in a sentence, Wray would suppose that it would be given its *next to* meaning: Wray would not expect subjects to produce a sentence containing the expression *by and large* – a formula in which there is no paradigmatic variation.

However, research by Gilquin (2005) has called into question Wray's theory. In her research, requiring subjects to use stimulus words in sentences, it was sometimes the case that non-prototypical senses of words were the sense in the sentences provided. For example, for *take* it was **not** the case that prototypical *take (i.e. grab)* was dominant in the elicited data. Gilquin (2005) notes that the most common uses in the elicited data were the 'move' sense of *take* (e.g. *I will take you home*) and phrasal verb instances, (22.5 percent each respectively of the elicited data). This is a particularly difficult case for Wray's theory as Wray specifically argues that, "an intuitive definition of *take* will home in on its concrete meaning of 'grasp' or 'capture' because it is in this meaning that it is most segmentable. Its common occurrences as an abstract carrier verb, in for example *take part*, *take on* ….are much less visible to our intuition, because there will have been little if any drive to segment *take* out of these strings" (2002: 277). The Gilquin data, plus the ability to produce highly frequent collocates of words in frozen collocations and idioms, as noted above in the section on word association research, challenge the Wray theory.

[3] Though there are a number of similarities too, about the consequences of chunking: resource conserving (Bush 2001: 268; Wray 1999: 215; Wray 2002: 16, 69); loss of meaning of individual elements of holistically stored units (Bush 2001: 269; Wray 2002: 200); and multiple representation (Bush 2001: 277; Wray 1999; Wray 2002: 262).

4. Research

4.1 Experiment design

Goals and hypotheses

Having covered the important background information above, I report on an experiment below, designed to investigate which theory is best able to account for the data collected in an experiment investigating the ability of language teachers to produce the most frequent collocates of twenty common adjectives. The reason for investigating frequent adjectives and their collocates is twofold: in word association data frequent adjectives rarely elicit nouns, so there is very little available data on this subject; and, secondly, it enables us to investigate the 'Bybeean' theory with its focus on frequent co-occurrence.

The starting point hypothesis underpinning this research is that respondents encode frequency automatically, and use an availability heuristic in their attempt to provide a frequent collocate. Further, the corpus against which the intuitions are compared (the BNC) is assumed to be broadly representative of language use.

Subjects

The respondents who participated in the test were all male EFL/EAP lecturers at King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. Ten of the subjects were British. The other nationalities were: American (5), Irish (2), Canadian (1), Australian (1) and South African (1). The above noted group was homogeneous educationally, and furthermore, all worked in the same teaching environment.

Task description

This experiment is a single response (doubly) controlled word association task.[4] The subjects are challenged to provide what they consider to be the most frequent noun collocate of twenty different adjectives. Their intuitions are compared to BNC data for the same adjectives.[5] The adjective stimuli are all high frequency and well distributed throughout different text types in the BNC and also in the Brown corpus.

BNC search procedure

Part of speech tagging was utilized to restrict the search to adjective occurrences of the stimulus words (given below). To make a listing of the word's different noun collocates (according to raw frequency), the output from the BNC search was modified - pronouns, determiners, articles, prepositions, adjectives and proper nouns were excluded (from a +1 search window) and a list of the most frequent noun collocates of the adjectives was obtained. In addition, collocating nouns were excluded from the list if they were not well distributed throughout the corpus.

[4] 'Doubly', in the sense that the response has to be both a noun collocate *and* a highly frequent one.
[5] The intuitions are compared to raw frequency of co-occurrence data, rather than z-score, MI score data.

The adjective stimuli were:

different, difficult, full, good, great, important, large, main, old, particular, personal, possible, real, recent, similar, small, special, strong, various, young

Task methodology

Subjects were informed in the instructions preceding the stimulus words to provide a noun in the slot next to each adjective, with the additional requirement that the noun should be one that the respondent believed to be the most frequent collocating noun of the adjective in the English language as attested by the BNC. Only one response was required.

4.2 Results

The Mann-Whitney U test was used to evaluate the 'reliability' of the intuitions vis a vis the corpus data. The assumption is made that twenty subjects who could not all guess the top collocate might reasonably be expected to produce the top twenty collocates between them. Since the BNC always gets the top collocate, the responses can't be ranked against it. Instead, the BNC data is treated as if it provided twenty different guesses. N was calculated not as the number of subjects, but the number of different responses given. The final N, the number of different responses, determined how many of the BNC's top collocates were used in the calculation. For example, if there were ten different responses, the ranking was conducted against the ten most frequent noun collocates of the stimulus word from the BNC. Perfect correspondence is not required for U to be non-significant. No two corpora will have exactly the same relative frequencies for words or for collocations (see for example Takaie 2002 on the first of these points); however, general corpora would be similar enough for us to say that there is no *significant* difference between them in the information that they provide on frequent collocates (see, for example, Keller and Lapata's (2003) study on adjective-noun bigrams in BNC, Altavista and NANTC and the high correlations between their relative frequencies). The focus of investigation here is to see whether the respondents' data is similar enough to the BNC for us to conclude that the samples come from the same 'language pool'.

With the exception of the words *difficult*, *real* and *young*, there was a significant difference between the BNC data and the elicited data for the responses to the stimulus words: i.e. the teachers' ideas were not, generally, statistically similar to the BNC. A check was made against the BNC spoken subcorpus, but this did not lead to greater agreement between the two sets of data. In Tables 1-3 below I note the BNC's most frequent collocates for each of the three stimulus words noted above, and the responses of the teachers (Ts) to the same words. N is given in the first column, and is always lower than twenty indicating that some responses were provided by more than one of the subjects. Where the same words occur in both lists they are underlined.

Difficult (N=10)	Ts: task, time, times, situation, problem, job, problems, decision, proposition, choice
	BNC: task, time, question, times, situation, questions, problem, job, problems, thing

Table 1: BNC data and teacher collocate data to the stimulus word *difficult*

Real (N=12)	Ts: world, life, thing, problem, time, estate, ale, situation, events, livewire, man, food
	BNC: world, life, terms, thing, problem, wages, reason, time, name, value, sense, danger

Table 2: BNC data and teacher collocate data to the stimulus word *real*

Young (N=9)	Ts: people, man, men, girl, person, boy, generation, kid, adult
	BNC: people, man, children, men, woman, women, girl, lady, girls

Table 3: BNC data and teacher collocate data to the stimulus word *young*

More will be said about the data above and suggestions forwarded as to why the responses to these words were statistically similar to the BNC data in the discussion in section 4.3 below.

While the above analysis helps us to compare the two sets of data and the similarities and differences between them, it hides the fact that a particular word may have been produced on more than one occasion by the subjects. It is helpful to focus on the primary response from word association data as, arguably, this is more important than 'lone' responses for the insights that it can provide about 'mental wiring' for the informants as a group. In Table 4, below, the stimulus words are listed, and, when at least three respondents provided the same collocate response this is noted in column two. For example, the most common response to *different* was *people*. This is the seventh most frequent collocate of *different* in the BNC and five of the twenty respondents provided it.

Stimulus word	Dominant response	BNC Collocation Rank of dominant response	No of respondents providing dominant response
A. Different	People	7	5
B. Difficult	Task Problem	1 7	3 5
C. Full	Time	1	3
D. Good	Idea	1	6
E. Great	Time Men	>20 >20	6 3
F. Important	Issue Person Information	9 >20 >20	3 4 3
G. Large	-	-	-
H. Main	Idea Street Event	>20 6 >20	5 3 3
I. Old	Man	1	4
J. Particular	-	-	-
K. Personal	-	-	-
L. Possible	Answers Answer	>20 >20	3 3
M. Real	Time	8	7
N. Recent	Events Event	17 >20	6 4
O. Similar	Ideas	>20	3
P. Small	-	-	-
Q. Special	Occasion Event	>20 >20	4 5
R. Strong	-	-	-
S. Various	Items	>20	3
T. Young	Man People	2 1	5 7

Table 4: Dominant responses (Minimum 15 percent response attestation required for a particular word to be recorded in the table)

As can be observed, there was a dominant response (defined as one which was provided by at least three subjects) for fifteen of the words. Of the twenty-five dominant responses provided,[6] thirteen are not particularly frequent according to the BNC (i.e. they are outside the twenty most frequent collocates of the stimulus word). Such responses are, on average, around ten times less frequent than the most frequent collocate of the stimulus word. These less frequent collocations, which the respondents believed to be highly frequent are: *great time, great men, important person, important information, main idea, main event, possible answers, possible answer, recent event, similar ideas, special occasion, special event, various items*. The dominant responses which **were** the most frequent collocates according to the BNC were: *difficult task, full time, good idea, old man* and *young people*.

[6] Note that for some of the stimulus words there is more than one 'dominant' response.

4.3 Discussion

One of the most important matters to investigate, following on from the discussion in section 3 above, is whether there is any evidence that the denotational, lone 'dictionary meaning' sense of the stimulus word influenced respondents in their choices, and that, as a result, the collocates that they produced were, actually, not particularly frequent – i.e. investigate the view of Sinclair, noted earlier.

The evidence is mixed about the importance of this factor. Perhaps surprisingly, there are cases where the non-denotational meaning of the stimulus word is in the resulting collocate

on. For example, in only two of the eleven different responses to *full* (*glass*, *stomach*) from the subjects does *full* have its prototypical meaning 'no space'. Interestingly, Sinclair (2004: 21, 22) argues that in *full range*, *full* is delexicalised, but two subjects provided this response. Further, the meaning of *full* in *full time* (the dominant response) is not the typical/prototypical meaning of *full*. Another example, suggesting that the influence of the denotational meaning of the word is not so significant in 'driving' the responses, is the dominant association to *real - time*. Tognini-Bonelli (1993: 118) believes that the word *real* is "usually taken to mean 'existing in reality'". This is not, however, the meaning of the word in the dominant response. It seems then, that in these cases, a collocation is produced where the meaning of the stimulus word, (the adjective) is either delexicalised or does not have its primary denotational meaning in the resulting collocation.

However, there is some support for the denotational, stereotypical meaning of the word affecting the responses on occasion. This is particularly clear in the case of *great*. The 'excellent' meaning of *great* is its stand alone meaning – if something is *great* it is stereotypically *good, excellent, etc.* The dominant response for this word was *time*. In the resulting collocation, the meaning of *great* is clearly the 'excellent' meaning. However, the response is outside the twenty most frequent collocates of *great*. In many of its most frequent collocations the meaning of *great* is 'large / big' (e.g. when it collocates with *majority, interest, importance, care, etc.*) i.e. in combination with certain nouns, it does not have its stereotypical meaning. One can argue that the denotational, stand-alone meaning of *great* may have influenced respondents in their production of the dominant collocates[7] which are actually not very frequent according to BNC data.

But why is the evidence so mixed for the role of the prototypical meaning of the stimulus word in influencing the associates? This explanation *can* explain why the responses to *great* were as they were, but not the responses to *full* or *real.* What else might be affecting the production of the associates?

The adjectives used as stimuli in the experiment are all frequent, and many words are frequent because they are present in frequent phrases (see e.g. Coulmas 1979: 239; Summers 1996: 262, 263; Stubbs 2002: 235). Some of the high frequency collocates, according to the BNC, occur 'embedded' with the stimulus word in a phrase, and others (usually with an article preceding the collocation) are more 'complete'. All of the dominant collocations in Table 1 above have one thing in common: they typically function as a complete unit, i.e. they are not embedded: they do not occur in larger phrasal chains. This is a crucially important finding. When we look at a large number of the most frequent collocates, they *do* typically occur in larger chains of language. There are three types of chain, as noted below.

[7] It should be noted, however, that the denotational meaning of the word is sometimes its meaning in a frequent collocation. The best examples from the dominant responses which support this are the collocations *main street* and *difficult problem*. In both of these cases the adjective has its denotational meaning: a *main street* is not 'minor' and it is 'large', and a *difficult problem* is a 'hard' problem, not an 'easy' one.

Some of the frequent 'bare' collocations typically occur in frameworks of the type *DET ADJ NOUN of NP* (e.g. *a* ***large amount*** *of money*). Sinclair (1991: 89) notes that one of the types of relationship between the two nouns in this type of chain is that the first noun phrase is supportive of the second, and that, as such, the second noun tends to be the most salient. For example, in the chain, *the usual kind of problem*, the second noun is more important in the information that it conveys than the first. Many of the noun collocates *not* provided by the respondents are these kinds of supporting nouns. For example, the dominant response to *different* was *people*. This is a reasonably frequent collocate. However, the typical syntactic pattern for the most frequent collocates of *different* is *DET different N of NP*. Indeed, according to the BNC *different types, different kinds, different parts* and *different aspects* occur around 90 percent of the time as the first noun phrase in the chain *DET different N of NP*. However, the case is very different for *different people* (the dominant response): it has no typical embedding patterns. Might it be that the most frequent nouns in the larger chains were not so available to the respondents, in their typical noun searches[8]?

A similar case is the dominant response *task* to *important*. As with *different*, there is a strong tendency for many of the most frequent collocates of *important* to occur in the first noun position in *DET ADJ NOUN of NP* chains. For example, *important aspect* occurs in this chain in 90 percent of its occurrences in the BNC, *important feature* 66 percent, *important source* 65 percent, and *important aspects* 71 percent. The dominant response *task* does not have this tendency to be embedded and neither do the other subject provided collocates (e.g. *important person, important issue, important thing, important idea,* and *important information*).

In addition, there are cases where words from a particular semantic field fill the supporting noun slot. For example for *large*, if a 'number type' noun follows (e.g. *sum, quantity, amount*) it typically occurs in the frame *a large [NUMBER NOUN] of NP*, e.g. *a large amount of NP*. However, when we examine the respondents' answers, we find that they did not typically provide words from the appropriate semantic field to fill the framework slot.

Secondly, some of the collocations occur in adverbial chains. For *recent*, items from a particular semantic field (time period) fill the slot in the adverbial chain *in recent [time period]*, e.g. *in recent years*, *in recent months, etc.* For *recent*, if the noun that follows this word is a time related noun, then it typically occurs in this chain. The dominant response to *recent* was *events* (six responses, and there were also four *event* responses); however, the most frequent collocates are time related and they are very strongly embedded in the framework noted above, e.g. *recent years* (84 percent), *recent months* (86 percent), *recent weeks* (88 percent). The fact that there was only one 'time' association – *times* – to the word *recent* indicates that while *recent* is clearly a time related adjective, it did not easily elicit very frequent time related nouns. Like *recent*, *similar* has very frequent noun collocates that typically occur in adverbial chains, e.g. *in a similar way, in a similar fashion, in a similar vein*. The dominant response to *similar* was *ideas*: the resulting collocation is not found in the adverbial framework.

Thirdly, and finally, some nouns occur in collocational frameworks which are unique to that noun, in the sense that the other frequent noun partners of the adjective cannot fill that framework slot. For example, *possible exception* typically occurs in the phrase *with the possible exception of*, while *with the possible way of*, with the possible solution of** are not typical. While this is a less important observation, it might explain why some frequent collocates were not provided.

Following on from these observations, it would seem that a sound explanation for the dominant productions (and omissions), would be the possibility that either the noun, the

[8] One might argue that because *people* is a more 'concrete' noun that it is more likely to be produced; however, a 'concreteness advantage' cannot be forwarded for many of the dominant responses.

(stimulus) adjective, or the 'bare' collocation within the larger chains is not so salient or accessible as collocates which do not occur in the larger chains. For example, assuming that *in recent years* is stored holistically, either *recent*, *years* or *recent years* appears to be 'hidden' in the chain, when respondents are searching for frequent collocates of *recent*. The respondents show a preference to provide a 'complete' collocation: for example, *good idea* is 'unit-like' (even without the determiner), and so too is *main street*. However, many of the most frequent noun collocates typically combine with the adjective in a larger chain of language, and they are incomplete as bare collocates: for example, *similar vein* is not complete in any real sense, and *large number* typically has a supporting role e.g. *a large number of problems*.

It would seem, therefore, that it is not the denotational meaning of the stimulus word that is the critically important driving force in affecting the quality of our lexical intuitions, but rather accessibility – and if a 'bare' collocation (i.e. a dyad) is typically embedded in a larger chain, then it seems to be not so accessible. Rather than arguing that respondents produce nouns which, in effect, delexicalise the stimulus adjective, it appears to make more sense to argue that some nouns are simply more accessible than others – those that occur with the adjective as a 'stand -alone' dyad, rather than with the stimulus word in a chain.

It may also be that this explanation could help in determining how it is that a particular meaning of a word becomes salient. Could it be the case that the availability of nouns for a particular adjective may affect our perception of the adjective's typical meaning? This explanation is consistent with Wray's theory that segmented material is analysed and the constituent material in formulaic language is less analysed. For example, returning to the *great* example noted earlier, the denotational, stand-alone, salient meaning of *great* is good/excellent. Why is this the salient meaning and why is the 'large/big' meaning not salient? It could be because *great* means 'large / big' when it is in combination with a set of nouns in holistically stored formulas. Because these are less prone to segmentation, the constituent parts have not been analysed, and as a result, the *great* = 'large / big' meaning of *great* does not occupy first place in the productive lexicon.

Can this explanation for what is going on in these responses account for why the native speaker responses to *difficult*, and *real* were so similar to the BNC data?[9] It can. *Difficult* is quite different from the other adjectives used in the experiment because none of its most frequent noun collocates show any tendency to occur in fixed, invariable phrases / frames. This, we would assume, assists the respondents in providing frequent associates. If we believe that an availability heuristic is being employed in the frequency searches, we would hypothesize that the responses would indeed be more accurate and less 'biased' in such a case. With regards to *real*, while *real time* does occur in the chain *'in real time*' it is only so in 41 percent of its occurrences in the BNC. This suggests that the words in this chain may be segmented, and as such, this would make the individual components more accessible than cases where the 'bare collocation' shows much more dominant embedding tendencies. Interestingly, while the respondents produced a set of associations that did not differ significantly from the BNC data for *real,* a very common collocate of *real* was not produced – *terms*. The dyad *real terms* occurs 97 percent of the time in the chain *in real terms*, according to the BNC. It is not surprising, given the argument above, that while the associates to *real* were, generally, very similar to the BNC data, this, the third most frequent collocate was not produced once.

[9] I exclude discussion of the responses to *young* and *old* here, as the semantic preference restriction [+ animate] seems to be the key reason for this response: see the earlier discussion on Clark's (1970) 'first rule'.

5. Conclusion

This experiment suggests that the noun collocate associations provided to frequent adjective stimuli by the respondents are, typically, not the same as those from the BNC data. It has been argued that accessibility problems in particular, may be responsible for: (a) the under/non-production of non-salient, supporting nouns in *NP of NP* chains (e.g. *large amount, different way*); (b) the failure to provide typical collocates of words in adverbial chains (e.g. in *recent years, in a similar way*); and (c) the failure to provide nouns which are unusual, in the sense that other noun partners of the adjective do not typically fill the noun slot in the frame (e.g. *with the possible exception of*).

It would seem from this experiment that the better candidates for formulaic language status (i.e. less analysed, holistically stored lexical combinations) are not dyad collocations, but rather fixed or semi-fixed phrases / language chains. This is because, at times, the respondents do seem to access a strong (frequent) collocate of a stimulus word (e.g. *good idea*); however, the data is fairly convincing in showing that the noun items in the frameworks are not so accessible.

It follows from the above, that the previous explanation for the failure to access the frequent collocates of *small* (Nordquist 2004) may not have captured what was happening in that experiment. Nordquist noted that number collocating nouns of *small* were not produced by respondents in her elicitation task as much as predicted by her corpus data (a finding with which this experiment concurs, even though the methodology she employed in testing the knowledge of the collocates was different). Her explanation was that the 'number' collocates of *small* were fused with *small*, because of their high frequency of co-occurrence, and that the words in the resulting fused combinations had become autonomous. The experiment reported on here suggests an alternative explanation for that finding: it is not, for example, *small part* that is stored holistically, but rather *a small part of NP*. The reason for forwarding this suggestion is that some high frequency dyads are produced (e.g. *full time, good idea, difficult task*). It is hypothesized, therefore, that frames, rather than dyads are formulaic and hence less accessible in elicitation experiments: as such, this may be a key explanation for why corpus data and elicited data may differ on the subject of frequent adjective-noun collocations.

Acknowledgement

I would like to thank King Fahd University of Petroleum and Minerals for its support in presenting this paper.

References

Arnaud, P.J.L. (1990) 'Subjective word frequency estimates in L1 and L2'. Paper presented at the 9th World Congress of Linguistics, Thessaloniki. *ERIC Document* ED329120, 1–15.

Backman, J. (1976) 'Some common word attributes and their relations to objective frequency counts'. *Scandinavian Journal of Educational Research 20*, 175–86.

Balota, D.A., M. Pilotti and M.J. Cortese (2001) 'Subjective frequency estimates for 2,938 monosyllabic words'. *Memory and Cognition 29(4)*, 639–47.

Beaugrande, R. de (1996) 'The pragmatics of doing language science: The warrant for large corpus linguistics'. *Journal of Pragmatics 25*, 503–535.

Biber, D., S. Conrad and R. Reppen (1996) 'Corpus based investigations of language use'. *Annual Review of Applied Linguistics 16*, 115–35.
Biber, D., S. Conrad and R. Reppen (1998) Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.
Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) Longman Grammar of Spoken and Written English. Harlow: Longman.
Brown, N. (1995) 'Estimation strategies and the judgment of event frequency'. *Journal of Experimental Psychology: Learning, Memory and Cognition 21*, 1539–53.
Bush, N. (2001) Frequency Effects and Word-boundary Palatalization in English, in J. Bybee and P. Hopper (eds) Frequency and the Emergence of Linguistic Structure, pp. 255–80. Amsterdam: John Benjamins Publishing Company.
Bybee, J. and P. Hopper (2001) Introduction to Frequency and the Emergence of Linguistic Structure, in J. Bybee and P. Hopper (eds) Frequency and the Emergence of Linguistic Structure, pp. 1–24. Amsterdam: John Benjamins Publishing Company.
Carroll, J.B. (1971) 'Measurement properties of subjective magnitude estimates of word frequency'. *Journal of Verbal Learning and Verbal Behavior 10*, 722–29.
Carter, R. (1987) Vocabulary Applied Linguistic Perspectives. London: Routledge.
Clark, H.C. (1970) Word Associations and Linguistic Theory, in J. Lyons (ed.) New Horizons in Linguistics, pp. 271-286. Harmondsworth, Middlesex: Penguin Books.
Cook, G. (1998) 'The uses of reality: a reply to Ronald Carter'. *English Language Teaching Journal 52(1)*, 57–63.
Cosmides, L. and J. Tooby (1996) 'Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty'. *Cognition 58*, 1–73.
Coulmas, F. (1979) 'On the sociolinguistic relevance of routine formulae'. *Journal of Pragmatics 3*, 239–66.
Desrochers, A. and M. Bergeron (2000) 'Valeurs de fréquence subjective et d'imagerie pour un énchantillon de 1916 substantifs de la langue française'. *Canadian Journal of Experimental Psychology 54(4)*, 274–325.
Ellis, N. (2002) 'Reflections on frequency effects in language processing'. *Studies in Second Language Acquisition 24*, 297–39.
Fox, G. (1987) The Case for Examples, in J. Sinclair (ed.) Looking Up, pp. 137–49. London: Collins.
Frey, E. (1981) 'Subjective word frequency estimates and their stylistic relevance in literature'. *Poetics 10(4-5)*, 395–407.
Gilquin, G. (2005) What you think ain't what you get: Highly polysemous verbs in mind and language. Paper presented at "From Gram to mind: Grammar as Cognition". Bordeaux 19-21 May 2005.
Groot, A.M.B. de (1989) 'Representational aspects of word imageability and word frequency as assessed through word association'. *Journal of Experimental Psychology: Learning, Memory and Cognition 15(5)*, 824–45.
Hintzman, D.L. (1978) 'Contextual variability and memory for frequency'. *Journal of Experimental Psychology: Human Learning and Memory 4(5)*, 539–49.
Hunston, S. (2002) Corpora in Applied Linguistics. Cambridge: Cambridge University Press.
Keller, F. and M. Lapata (2003) 'Using the web to obtain frequencies for unseen bigrams'. *Computational Linguistics 29(3)*, 459–84.
Kennedy, G. (1991) Between and Through: The Company They Keep and the Functions They Serve, in K. Aijmer and B. Altenberg (eds) English Corpus Linguistics, pp. 95–110. Essex: Longman.

Meara, P. (1980) 'Vocabulary acquisition: a neglected aspect of language learning'. *Language Teaching and Linguistics: Abstracts 13(4)*, 221–46.
Moss, H. and L. Older (1996) Birkbeck Word Association Norms. Hove, UK: Psychology Press.
Nordquist, D. (2004) Comparing Elicited Data and Corpora, in M. Achard and S. Kemmer (eds) Language, Culture and Mind, pp. 211–23. Leland Stanford University: CSLI Publications.
Peters, A.M. (1983) Units of Language Acquisition. Cambridge: Cambridge University Press.
Renouf, A. (1997) Teaching Corpus Linguistics to Teachers of English, in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds) Teaching and Language Corpora, pp. 255–66. Harlow: Longman.
Ringeling, T. (1984) 'Subjective estimations as a useful alternative to word frequency counts'. *Interlanguage Studies Bulletin 20(8)*, 59–69.
Shapiro, B.J. (1969) 'The subjective estimation of relative word frequency'. *Journal of Verbal Learning and Verbal Behavior 8*, 248–51.
Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford: Oxford University Press.
Sinclair, J. (2004) The Lexical Item, in J. Sinclair (ed.) Trust the Text: Language, Corpus and Discourse, pp. 131–48. London and New York: Routledge.
Söderman, T. (1993) Word Associations of Foreign Language Learners and Native Speakers-Different Response Types and Their Relevance to Lexical Development, in B. Hammarberg (ed.) Problem, Process and Product in Language Learning, pp. 157–69. Stockholm: Stockholm University, Department of Linguistics.
Sosa, A.V. and J. MacFarlane (2002) 'Evidence for frequency-based constituents in the mental lexicon: collocations involving the word *of*'. *Brain and Language 83*, 227–36.
Stubbs, M. (1995) 'Collocations and semantic profiles: On the cause of the trouble with quantitative studies'. *Functions of Language 2(1)*, 23–55.
Stubbs, M. (2002) 'Two quantitative methods of studying phraseology in English'. *International Journal of Corpus Linguistics 7(2)*, 215–44.
Summers, D. (1996) Computer Lexicography: The Importance of Representativeness, in J. Thomas and M. Short (eds) Using Corpora for Language Research, pp. 260–66. London: Longman.
Takaie, H. (2002) 'A trap in corpus linguistics: The gap between corpus based analysis and intuition based analysis'. *Language and computers 38*, 111–30.
Taylor, S.E. (1982) The Availability Bias in Perception and Interaction, in D. Kahneman, P. Slovic, and A. Tversky (eds) Judgment Under Uncertainty: Heuristics and Biases, pp. 190–200. Cambridge: Cambridge University Press.
Tognini-Bonelli, E. (1993) Interpretive Nodes in Discourse: Actual and Actually, in M. Baker, G. Francis and E. Tognini-Bonelli, (eds) Text and Technology. In Honour of John Sinclair, pp. 193–212. Amsterdam: John Benjamins Publishing Company.
Tryk, H.E. (1968) 'Subjective scaling of word frequency'. *American Journal of Psychology 81*, 170–77.
Tversky, A. and D. Kahneman (1973) 'Availability: A heuristic for judging frequency and probability'. *Cognitive Psychology 5*, 207–232.
Tversky, A. and D. Kahneman (1982) Judgment Under Uncertainty: Heuristics and Biases, in D. Kahneman, P. Slovic and A. Tversky (eds) Judgment Under Uncertainty: Heuristics and Biases, pp. 3-19. Cambridge: Cambridge University Press.
Widdowson, H.G. (1989) 'Knowledge of language and ability for use'. *Applied Linguistics 10(2)*, 128–37.
Willis, D. (1990) The Lexical Syllabus. London: Harper Collins.

Wray, A. (1999) ‘Formulaic Language in learners and native speakers’. *Language Teaching 32*, 213–31.
Wray, A. (2002) Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Salience and Frequency of Meanings: Comparison of Corpus and Experimental Data on Polysemy

Daniela Marzo, Verena Rube and Birgit Umbreit[1]

Abstract

In this paper we try to contribute to the discussion about the interdependence of frequency and salience with an empirical study in which we compare two methods of researching polysemy on the basis of fifteen Italian high frequency words: a production experiment and a corpus analysis. Our findings support the hypothesis that frequency researched by corpus analysis and salience researched by psycholinguistic experiments are not the same, but are closely connected. We found particularly two important convergences between the corpus and the experiment results:

(i) In almost two thirds of the data the most important meanings in the corpus analysis and in the experiment are identical.
(ii) If (i) is not matched there are still correspondences between the two most important meanings.

Another interesting result is that there is a slight preference of concrete over abstract meanings in the experiments.

1. Introduction

In spite of the vast amount of literature on frequency and salience the relation between these two terms is still controversial. We would like to contribute to that discussion with an empirical study in which we compare two methods of researching polysemy on the basis of fifteen Italian high frequency words: a production experiment and a corpus analysis.
The main goal of our talk is to discuss the status of the respective results: Are the meanings found in corpora and by experiments the same? If they are, what conclusions can we draw from that? If they are not, does this support the hypothesis of the difference of salience (experiment) and frequency (corpus)? Gilquin (2006 and 2007) who adheres to this hypothesis claims that there are nevertheless some links between the two notions: Do our results confirm her findings?

Our paper is organized as follows: Chapter 2 is dedicated to some preliminary reflections: Section 2.1 contains a literature survey on similar corpus vs. elicitation studies, while section 2.2 defines and discusses the notions of frequency and salience. In chapter 3, we explain our own data sources and their analysis, i.e. the Sentence Generation and Definition Task (3.1) and the Corpus Analysis (3.2). Chapter 4 presents our results together with some general observations concerning frequency and salience (4.1), followed by a comparison with the findings in Gilquin (2006 and 2007). Chapter 5 concludes.

[1] SFB 441: Linguistic Data Structures, University of Tübingen
e-mail: daniela.marzo@uni-tuebingen.de, verena.rube@uni-tuebingen.de, birgit.umbreit@uni-tuebingen.de

2. Preliminary reflections

2.1 Comparing corpus data and elicitation data: a survey of similar studies

Even though the problem of data types has seen an increase in interest recently,[2] "there are few studies in which an individual topic is tackled from more than one methodological perspective, producing what is commonly referred to as «converging evidence»" (Gries, Hampe and Schönefeld, 2005).

It is therefore no wonder that there are only few studies comparing the frequency of word meanings in corpora and the salience of word meanings in psycholinguistic experiments. Generally, rather than word senses, verb subcategorizations are compared (*cf*. Roland and Jurafsky, 2002; Merlo, 1994). Word sense analysis comparisons are made, e.g. by Gilquin (2006 and 2007) and, as a by-product, by Roland and Jurafsky (2002). Both studies confirm that there are differences between verb sense distribution in corpus studies and in experiments. The comparison of the study of Roland and Jurafsky though is quite unreliable. As experimental data they use the results of a sentence production test. In the test subjects were presented with stimuli and topics and they were asked to write sentences using the stimuli based on the given topic. Certainly, with that procedure subjects are influenced and the results do not reflect the most salient meanings but by chance. The other experiment Roland and Jurafsky use shows even bigger problems: sentence completion tasks do not give unbiased results about the salience of word meanings either.

The studies of Gilquin, instead, are better comparable to our own research. In both cases she uses sentence production tests without a given topic, i.e. she does not guide the subjects' answers into a predefined direction. In the first study she compares the frequency and salience of *take* and *give*, in the second study she broadens the data comparing different languages (English, French, Dutch). She shows that the most frequent senses – found in the corpora – and the most salient senses – found in the elicitation experiments – most of the time do not correspond, but that there are nevertheless some links between salience and frequency. A more detailed comparison of her and our approach will follow in 4.2.

2.2 The frequency and salience problem

As generally assumed, corpus studies and experiments do not seem to lead to the same kind of results: corpus studies tell something about the frequency of meanings and experiments reveal their salience. But how exactly are these two concepts related?

Schmid sets up the so called from-corpus-to-cognition hypothesis. In his opinion the

> "importance of a linguistic phenomenon in a given language can be extrapolated from an analysis of its frequency in a large corpus" (2000: 39). The importance of an underlying cognitive phenomenon in our cognitive system can thus be extrapolated from an analysis of its frequency. In other words: "Frequency in text instantiates entrenchment in the cognitive system" (2000: 39).

Entrenchment, in turn, is a term created by Langacker. His notion of entrenchment is inseparably interwoven with the notion of frequency as well as with the notion of cognitive salience: "Entrenchment pertains to how frequently a structure has been invoked and thus to the thoroughness of its mastery and the ease of its subsequent activation" (1991: 45). As you

[2] E.g. the collaborative research centre SFB 441 (http://www.sfb441.uni-tuebingen.de) and its *Linguistic Evidence* conferences dedicated to data problems.

can see from that quote, for Langacker, frequency is a precondition for entrenchment and ultimately salience (ease of activation). The interdependency of these notions even evokes statements about their being identical (*cf*. 1991: 159).

Gernsbacher (1984), on the contrary, points out that frequency and salience are not necessarily the same, as word frequencies in corpora are only an approximation to "experiential familiarity", i.e. salience. She sustains that corpora might reflect the salience of low frequency words inaccurately, because words can be experientially familiar even if they are not frequent in general language use, a hypothesis which is also supported by De Mauro (1980).

Being confronted with these contrasting hypotheses, we might wonder whether the term *salience* is used in diverging meanings. Asked differently, do ease of activation and experiential familiarity correspond?

Let us have a closer look at what exactly they are about. When talking about salience and frequency Schmid and Langacker are talking about what frequencies found in corpora reveal about the cognitive system, i.e. their notion of salience is nothing but the logical outcome of corpus evidence. When studying experiential familiarity by carrying out off-line rating tasks, on the other hand, Gernsbacher analyzes conscious judgments about language that can not necessarily be put on a par with corpus data.

But nevertheless it is reasonable to claim that the speakers' conscious judgments about language are connected to or even governed by what is also prominent in the unconscious mind. And the latter can be more directly researched by corpus evidence – always keeping in mind that corpus data depend on the representativeness of the corpus.

Summing up the discussion we can say that there are two assumptions that diverge only prima facie: (i) Frequency directly reflects/corresponds to salience. (ii) Frequency and salience are closely connected, but do not correspond directly.

The difference between these two assumptions seems to be due to the difference between the definitions of salience. As we are working – like Gernsbacher – with speaker judgments we would like to part from thesis (ii) and try to check the assumption with our results.

3. Methodology

3.1 Sentence generation (SGT) and definition task (DT)

3.1.1 Presentation of the SGT and DT

The method we chose consists in a sentence generation task combined with a definition task (SGandDT). A SGT, also called sentence production task, is an off-line experiment in which the subjects are asked to produce sentences. The goal of such questionnaires is either to find out which is the first meaning that comes to the subjects' mind (Gilquin, 2006 and 2007) or to get the most important meanings of the stimulus words (Caramazza and Grober, 1976; Colombo and Flores d'Arcais, 1984; Raukko, 2003). Even if the subjects generally formulate sentences that allow the linguist to tease out the intended meanings quite easily, there are always some sentences that stay ambiguous in spite of the subjects' and the linguists' effort. In order to avoid these unfortunate cases, we combined the SGT with a DT. In a DT the subjects are asked to disambiguate the stimulus' meanings by defining or paraphrasing the meanings that come to their mind. As the informants usually are not used to defining the words they use every day and therefore have to perform a quite unnatural task (see Dunbar's critique, 2001: 2-3), the results can be quite obscure, if the DT is applied on its own.

However, in combination, the SGT and the DT allow us to gather quite sound data: as, first, the subjects are given a second chance to explain what they want to express and, second, the linguist can rely on the definition in cases in which the disambiguating sentences are ambiguous and vice versa, the amount of not interpretable data decreases considerably (see Raukko 2003). This can be exemplified by responses to our stimulus *grande*:

(1) It. Quel cantante è **grande** (famoso).
En. This singer is great (famous).
(2) It. Ho preso un televisore **grande** (di grosse dimensioni).
En. I have taken a big television (of big dimensions).

The supposedly disambiguating sentence in (1) on its own is ambiguous (*grande* could both mean 'spatially extended' and 'famous') in comparison with other sentences like e.g. the one in (2), where *grande* cannot mean anything else than 'spatially extended'. If we consider the definition in (1), the sentence becomes unambiguous, too.

Our sample of stimuli consisted of 400 Italian words. Each subject was presented with a questionnaire accessible on internet containing twenty stimuli. The resulting twenty questionnaires were filled out by about thirty informants each. From these data we chose the results for the fifteen most frequent words (according to Juilland 1973) and compared them with corpus occurrences.

3.1.2 The analysis of the sentence generation plus definition task (SGandDT)

As stated in 3.1.1 the linguist has to consider two types of responses when analyzing the SGandDT: (i) hopefully disambiguating example sentences and (ii) definitions or paraphrases. Our hierarchy of analysis was the following:

d. We first looked at the example sentences and the definitions and attributed a meaning to the stimulus.
e. If the definitions contradicted the example sentences, the example sentences were regarded as more important than the definitions.
f. As subjects usually gave more than one example sentence, we had to define the meanings not only with respect to the sentence and the definition, but also with respect to the other sentences the single subjects formulated.

(c) can be exemplified by the stimulus *grande*: 9 percent of the informants of the SGandDT (i.e. three out of twenty-nine) distinguished *grande* in the sense of 'tall, extended in height' like in (3) from *grande* in the sense of 'spatially extended' like in (4).

(3) It. Quel bimbo diventerà **grande** in fretta (crescere).
En. This kid will get big very soon. (to grow).
(4) It. Ha un **grande** appartamento (di vaste dimensioni).
En. He has got a big apartment (of big dimensions).

Taking a look at both of the sentence-definition pairs on their own, we might consider both of the occurrences of *grande* as instances of the meaning 'spatially extended'. However, as 9 percent of the informants distinguished sentences like (3) and (4), we had to define a second, more specific spatial meaning, namely 'tall, extended in height'.

3.2 Corpus study

3.2.1 Corpus presentation

The *Lessico di frequenza dell'italiano parlato* (LIP) was used for the corpus searches. This corpus consisting of transcripts of spoken texts collected between 1990 and 1992 is now accessible online in the *banca dati dell'italiano parlato* (BADIP).[3] It contains 496 transcripts including about 490 000 words from different types of oral conversations. The conversation types comprise bidirectional face-to-face conversations with free turn-taking (e.g. conversations at home, at work), bi-directional non face-to-face conversations with free turn-taking (telephone conversations), bi-directional face-to-face conversations with regulated turn-taking (e.g. legislative assemblies), mono-directional exchange with the addressee being present (e.g. university lectures), and distanced unidirectional exchange (e.g. radio programs), *cf.* e.g. Bellini and Schneider (2006: 15). The recordings were made in four cities: Florence, Rome, Naples and Milan. We have chosen the LIP-corpus as basis for our research, because it is easily accessible on the internet and, what is more, it can be searched for lemmatized stimuli.

Our samples are taken from a subcorpus of the LIP containing all the conversations recorded in Florence. For each of the fifteen stimulus words we chose fifty randomly selected occurrences, ten from each conversation type. The size of the samples for the stimuli equalled the average size of the amount of sentences given in the SGandDT. We sorted out occurrences with non-lexical meanings and occurrences which were part of larger lexical units, i.e. idioms, and substituted them by simple occurrences with lexical meaning.

(5) *volere*
It. * ecco voi capite che cosa **vuol** dire rapporto est ovest *
En. * now you understand what the East West relation means*

This occurrence of *volere* was recognized as a part of the idiom *voler dire* 'to mean, signifiy' and substituted by the next occurrence of *volere*.

3.2.2 Corpus analysis (CA)

The word sense disambiguation was carried out manually. Each occurrence was, if possible, assigned one of the senses established in the analysis of the SGandDT.

(6) *volere*
It. (…) gli operai **volevano** un posto nelle fabbriche un salario adeguato (…)
En. The workmen wanted a job in the factories, an adequate salary (…)

The occurrence in (6) would be assigned the meaning 'to want (+to have)', a meaning already defined during the analysis of the results of the SGandDT. If none of the senses fitted we defined new senses. When it was impossible to decide which of two or more senses could be assigned to the occurrence it was marked as ambiguous. Every word was semantically tagged by one tagger and his tagging was reviewed by one of the others in order to obtain a higher grade of objectivity.

[3] http://languageserver.uni-graz.at/badip/.

3.3 Comparison of the two methods of data interpretation

Comparing our data and the way in which they had to be interpreted, we can make some general observations:

(i) There are more cases of not interpretable data in the CA than in the sentence SGandDT.

There is a twofold advantage in the SGandDT: (a) As it is combined with a definition task, unclear sentences become quite clear. (b) If the definition does not help us to understand the intended meaning, there still is the opposition to the other sentences formulated by the same informant that may point to a sentence's meaning. In the CA, on the other hand, an unclear sentence usually remains unclear. Sometimes the textual context gives us a cue to the meaning of the word we study, but most often even the context is ambiguous or simply too vague. Apart from that there are numerous gaps, incomplete words or sentences, ungrammatical constructions and self-corrections of the speakers which hinder the semantic interpretation of the corpus occurrences.

This is why there are, e.g., 22 percent of not interpretable pieces of data in the CA for *dovere* opposed to only 6 percent of such data in the SGandDT, see Figure 1.

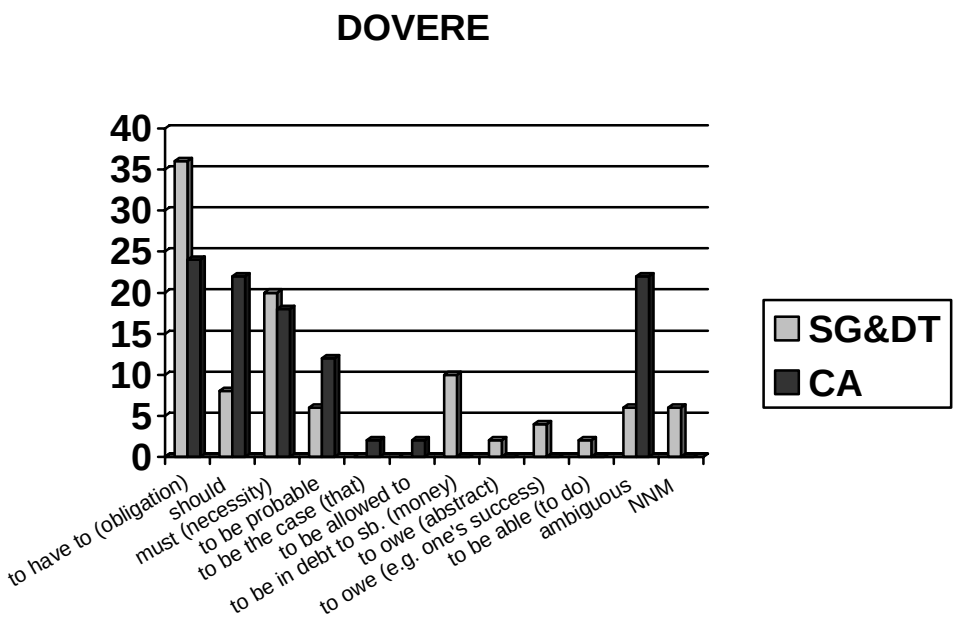

Figure 1: Results for It. *dovere*: 22 percent of ambiguous pieces of data in the CA vs. 6 percent in the SGandDT (NNM stands for 'no new meaning' which means that an informant produced different sentences with the same meaning).

(ii) The CA does not lead to the same meaning clusters as the SGandDT.

On the one hand, this is due to the fact that, as explained above, the informants distinguished senses which they obviously considered as different but which could not be distinguished in the corpus, where meanings often had to be defined in a more general manner. Thus, an analysis as fine-grained as in the SGandDT was often not possible.

On the other hand, several new meanings which did not arise in the SGandDT had to be defined for the CA. The reason for this is the composition of the corpus: The corpus contains quite specific subjects, e.g. sale offers for different articles, where both *fare* and *essere* were used in the sense of 'to cost' – a quite specific meaning speakers were not aware of in the SGandDT.

This discrepancy between the senses defined in the CA and the SGandDT is only natural, as we could not, in each case, apply the meaning distinctions already obtained in the SGandDT directly to the corpus, but unfortunately it makes the comparison of the results more difficult. In order to get more user-friendly results, we thus decided to merge together meanings originally distinguished either in the SGandDT or in the CA. For *essere*, e.g., in the SGandDT speakers distinguished the sense of 'to be (+permanent characteristic)' from 'to be (+temporary characteristic)', while this distinction did not hold in the CA, where only a general meaning 'to be (+characteristic)' could be defined. However, the sum of the two "characteristic"-meanings in the SGandDT directly corresponds to the more general "characteristic"-meaning in the CA, which was the most important sense for both samples. Consequently, in Table 1 below, the data for *essere* for both AGandDT and CA are summed up under the meaning 'to be (+characteristic)'. At the same time, we are perfectly aware of the fact that in the SGandDT alone, this meaning has to be further distinguished.

Similar merges were done for the most important meanings of *fare, potere* and *dovere* which were further distinguished either in the corpus or in the SGandDT.

4. Results of the data comparison

4.1 General observations

4.1.1 Are the most salient meanings of the SGandDT and the most frequent meanings of the CA the same?

In the light of a closer look at the data from our two studies, we can draw two conclusions:

(i) The most salient meanings in the SGandDT in the majority of the cases (nine out of fifteen) correspond to the most frequent meanings of the CA.

This is true for *venire* 'to come (to speaker)' which covers almost 29 percent of the occurrences in the SGandDT (followed by 'to go somewhere' with nearly 24 percent and 58 percent of the occurrences in the CA (followed by 'to cost (price)' with 10 percent). The same holds for *andare*, *dire*, *dovere*, *essere*, *fare*, *potere*, *sapere*, *vedere*, as you can see in Table 1. Among these *essere* is slightly different from the others, because the CA led to two equally frequent most frequent meanings. This can be explained by the composition of the corpus itself: as there are many telephone conversations in it, one of the two tied most frequent meanings is 'to be (+identity)' (40 percent), because *essere* is used in this meaning when presenting oneself on the phone. Probably, this sense would have been much less frequent in differently composed corpora. Consequently, it is reasonable to assume that the equally frequent meaning 'to be (+characteristic)' is more important and that therefore (i) can be said to hold for *essere*, too.

	most important meaning SGandDT	**%**	**2nd most important meaning SGandDT**	**%**	**most important meaning CA**	**%**	**2nd most important meaningCA**	**%**
andare	to go (+ concrete destination)	54	to walk	7	to go (+ concrete destination)	57	to feel (+bad, good)	12
			to move (+ means of motion)	7				
			to proceed	7				
avere	to have (materially)	22	to have (non-materially)	15	to have (non-materially)	54	to feel (psychic impression)	16
			to feel (physical impression)	15				
cosa	concrete object	39	thing (no concrete object)	20	thing (no concrete object)	84	concrete object	10
dare	to hand over (+ concrete object)	32	to give sth. (+abstract object)	21	to give sth. (+abstract object)	48	to hand over (+ concrete object)	30
dire	to say	68	to ask sb. to do sth.	13	to say	56	to think	8
dovere	to have to (obligation)	36	must (necessity)	20	to have to (obligation)	24	must (necessity)	16
essere	to be (+ characteristic)	41	to be (somewhere)	18	to be (+ characteristic)	40	to be (somewhere)	4
					to be (identity)	40	to be located (+abstract object)	4
fare	to do	42	to prepare (+ food)	13	to do	36	to make sb. do sth.	20
grande	spatially extended	33	important/ admirable	18	big, (abstract entity, intensitv)	38	spatially extended	32
potere	to be able to do sth.	49	to be allowed to do sth.	16	to be able to do sth.	60	to be allowed to do sth.	26
sapere	to know sth.	50	to be able to do sth.	15	to know sth.	70	to learn	16
stare	to be (somewhere)	24	to feel (+bad, good)	22	to stay (+condition)	26	to be (somewhere)	18
vedere	to see sth.	23	to perceive (with all senses)	21	to see sth.	20	to meet sb.	14
							to look at	14
venire	to come (to speaker)	29	to go somewhere	24	to come (to speaker)	58	to cost (price)	10
							ambiguous	10

volere	want (+to have)	61	to intend	16	to intend	50	want (+to have)	46

Table 1: The most and second most important meanings in the SGandDT and the CA

(ii) If the most important meaning in one data type does not correspond to the most important meaning in the other data type, it is likely to correspond to the second most important meaning. Sometimes this works in both directions.

This is the case with *cosa* in the sense of 'abstract thing' and 'concrete thing'. 'Abstract thing' is the most frequent meaning in the CA (84 percent), whereas it is the second most salient meaning in the SGandDT (20 percent). 'Concrete thing' on the other hand is the second most frequent meaning in the CA (10 percent) and the most salient meaning in the SGT (39 percent). The same holds for *avere*, *dare*, *grande*, *stare* and *volere*. Interestingly, the most frequent and the most salient meanings correspond to each other crosswise for *cosa*, *dare* and *volere*, but *avere*, *stare* and *grande* do not fulfil the condition in both directions, as the most salient meaning of the SGandDT of *stare*, e.g., i.e. 'to be (somewhere)' corresponds to the second most frequent meaning in the CA, but the most frequent meaning in the CA 'to stay (+ condition)' is not the same as the second most salient meaning of the SGandDT, which is 'to feel (+ bad, good)'.

(iii) The less frequent meanings in the CA and the less salient meanings in the SGT usually diverge.

This is, to a more or lesser extent, true for all of the stimuli and is exemplified by *essere* in Figure 2. We can explain this phenomenon simply by the fact that the frequency of these meanings is extremely low.

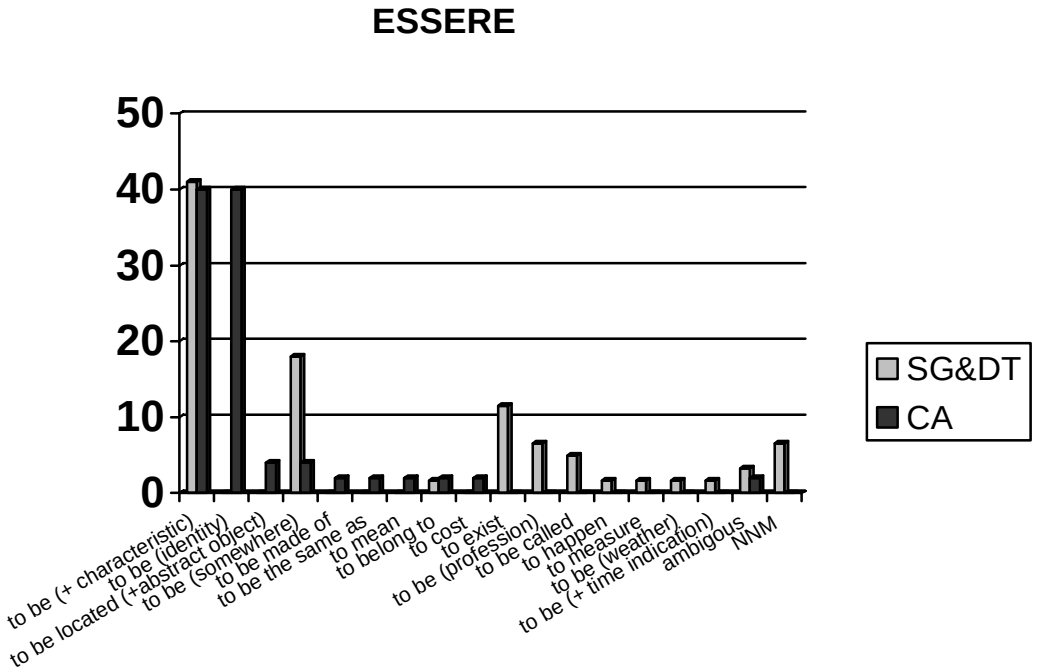

Figure 2: It. *essere* as an example for extremely diverging low-frequency meanings in SGandDT vs. CA

4.1.2 Concrete and abstract meanings

As we have seen in 4.1.1. the majority of the results show that the most prominent meaning in the CA and in the SGandDT are identical. Nevertheless there are other cases in which the most prominent meanings do not coincide. For these cases another interesting result can be shown: The most important meaning in the corpus is an abstract meaning and the most important meaning in the SGandDT is a concrete meaning. Having a look at Table 1 one can see e.g. that the most frequent meaning for *cosa* with 84 percent in the corpus is the 'abstract thing' meaning, in the SGandDT, instead, we have a not so big but still evident preponderance of the 'concrete object' meaning. This also holds for *avere*, *stare*, *dare* and *grande*, i.e. for five of the six stimuli that have a concrete meaning as the most important meaning in the SGandDT and an abstract meaning as the most important meaning in the CA.

The only stimulus where the situation is not so easy to determine because of the general non-concreteness of its meanings is *volere*. Nevertheless even in the case of *volere* one could maybe say that the meaning 'to want (+ to have)' – the meaning which wins in the SGandDT – is still more concrete than the meaning 'to intend'. This result seems to show that concrete meanings are very prominent in the consciousness of the speaker – at least in those cases in which the words do have concrete and abstract meanings and in which (i) in 4.1.1 does not hold - and that this fact seems to inhibit the total congruence of the results of the both studies.

4.2 Comparing our results with Gilquin's findings

As stated above (2.1) Gilquin (2006 and 2007) also compares the notion of frequency on the basis of corpus data with the notion of salience on the basis of data gathered in an elicitation test.

Discussing her results for the two high-frequency verbs En. *take* (and Fr. *prendre,* Dt. *nemen)* and En. *give* (and Fr. *donner,* Dt. *geven*) she concludes that salience does not reflect frequency directly, but that there are some weak links between the two: the most salient sense of En. *take* 'to move', e.g., is at least fourth in the corpus. If phrasal and collocational meanings (which are usually said to be holistically stored in the mental lexicon, see Wray (2002) are excluded from the analysis of the corpus, the 'move' sense is even the most prominent.

Considering our own results, which are comparable to the ones Gilquin obtained despite some minor methodical differences (*cf.* 2.1 and 2.2), we agree with her in that frequency does not reflect salience directly: Our SGandDT-data diverge too much from the corpus-data as to claim a one-to-one-correspondence. However, our results reflect stronger links between the two notions than those claimed by Gilquin (2007):

First of all, we can confirm her observation about the relation between concrete and abstract senses and the data types: As presented in 4.1.2, the primacy of the concrete over the abstract only holds for the SGandDT, whereas in the CA abstract meanings predominate. Thus, as far as the relation between concrete and abstract meanings is concerned, our data do not indicate any link between frequency and salience.

A factor that might distort the direct comparability of Gilquin's and our results is that we do not consider collocations as an extra-category, because our SGandDT-results have shown that, in contrast to what Wray (2002) claims, speakers do not seem to store these kind of multi-word expressions holistically, as they exemplify meanings quite often with the help of (non-idiomatic) collocations. This means that the cases Gilquin considered as collocations

are regular lexical meanings for us and consequently increase the amount of occurrences subsumed under one (mostly abstract) meaning.

However, there are two additional points that signal a relation between frequency and salience: The first one is the identity of the two most important meanings for nine out of fifteen words we studied, the second one is the fact that if the first observation does not match, frequently the most salient meaning corresponds to the second most frequent one (and/or vice versa, *cf.* 4.1.1). While the first case constitutes a rather important congruence, the second case shows that there is a certain convergence between the most important and the second most important meanings of both data samples.

5. Conclusion

Coming back to the two hypotheses cited in chapter 2.2 we can say that our findings support the second hypothesis: Frequency and salience are not the same, but are closely connected. We have tried to show above that there are particularly two important convergences between the corpus and the experiment results which go beyond the link found by Gilquin (*cf.* 4.2):

(i) In almost two thirds of the data the most important meanings in CA and in SGandDT are identical.
(ii) If (i) is not matched there are still correspondences between the two most important meanings (*cf.* 4.1).

This interesting result shows the need for further studies on the relation between corpus and experimental data. Maybe frequency and salience are closer connected than predecessor studies (Roland and Jurafsky, Gilquin) have shown?

References

Bellini, D. and St. Schneider (2006) Spoken Italian Online: The *Banca dati dell'italiano parlato* (BADIP), in B. Kettemann and G. Marko (eds) Planning, Gluing and Painting Corpora. Inside the Applied Corpus Linguist's Workshop, pp. 13–26. Frankfurt am Main: Peter Lang.

Caramazza, A. and E. Grober (1976) Polysemy and the structure of the subjective lexicon, in C. Rameh (ed.) Semantics: Theory and Application. Georgetown University Round Table on Linguistics 1976, pp. 181–206. Washington D.C.: Georgetown University Press.

Colombo, L. and G. B. Flores d'Arcais (1984) 'The meaning of Dutch prepositions: Psycholinguistic study of polysemy'. *Linguistics 22*, 51–98.

De Mauro, T. (1980) Guida all'uso delle parole. Come parlare e scrivere semplice e preciso. Uno stile italiano per capire e farsi capire. Roma: Editori Riuniti.

De Mauro, T. and F. Mancini (1993) Lessico di frequenza dell'italiano parlato. Milano: Etas Libri.

Dunbar, G. (2001) 'Towards a cognitive analysis of polysemy, ambiguity, and vagueness'. *Cognitive Linguistics 12(1)*, 1–14.

Gernsbacher, M.A. (1984) 'Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy'. *Journal of Experimental Psychology 113*, 256-281.

Gilquin, G. (2006) Towards an empirically grounded definition of prototypes, Poster Presentation at *Linguistic Evidence II*, Tübingen, 2–4 February 2006.

Gilquin, G. (2007) Universality and language specificity in prototypicality. Paper presented at *the Second AFLiCo Conference*, Lille, 10-12 May 2007.

Gries, St. Th., B. Hampe and D. Schönefeld (to appear) Converging evidence II: more on the association of verbs and constructions, in J. Newman and S. Rice (eds). Experimental and empirical methods in the study of conceptual structure, discourse, and language. Stanford, CA: CSLI. Available online from http://www.linguistics.ucsb.edu/faculty/stgries/research/ Coll_vs_Freq_2.pdf (accessed 25 June 2007).

Juilland, A. and V. Traversa (1973) Frequency dictionary of Italian words. The Hague: Mouton de Gruyter.

Langacker, R. W. (1991) Foundations of Cognitive Grammar. Vol. II. Stanford: Stanford University Press.

Merlo, P. (1994) 'A Corpus-Based Analysis of Verb Continuation Frequencies for Syntactic Processing'. *Journal of Psycholinguistic Research 23(6)*, 435–57.

Raukko, J. (2003) Polysemy as flexible meaning: experiments with English *get* and Finnish *pitää*, in B. Nerlich *et al.* (eds) Polysemy: Flexible Patterns of Meaning in Mind and Language, pp. 161–93. Berlin: Mouton de Gruyter.

Roland, R. and D. Jurafsky (2002) Verb sense and verb subcategorization probabilities, in Stevenson, S. and P. Merlo (eds) The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues, pp. 325–46. Amsterdam: John Benjamins Publishing Company.

Schmid, H. J. (2000) English Abstract Nouns as Conceptual Shells. From Corpus to Cognition. Berlin: Mouton de Gruyter.

Wray, A. (2002) Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Collocations in Elicitation and Corpora
Predictable Divergence

Dawn Nordquist[1]

Abstract

While the divergence of elicited data from corpora presents considerable methodological problems, the discrepancy between these data types has also been recognized as a theoretical concern within usage-based linguistics (e.g. Barlow 1996): if frequency of usage, for example, is an important determinant of mental linguistic constructs, elicited data, which are presumably a product of the grammar, should perhaps better mirror corpus data. Yet, when prompted with a word in an elicitation task, speakers do not typically respond with that word's most frequent corpus collocation. This paper argues that such behavior in elicitation with respect to collocations is actually consistent with the usage-based principles of holistic and autonomous storage of frequent multimorphemic units (e.g. Bybee 1985). The paper also considers how an Interactive-Activation model (McClelland and Rumelhart 1981) can capture this behavior, further demonstrating the value of analyzing elicited data from a psycholinguistic perspective to predict divergence between the two data types with respect to collocations.

1. Introduction

The lack of correspondence between speaker intuitions and corpus data is not new to linguistics. Even before the advent of today's powerful computer technology and large corpora, Labov (1972: 103), for instance, noted a "disjunction between norms and the patterns of everyday speech". However, with the availability of large corpora, this phenomenon is now receiving more attention as the patterns of everyday speech are brought into relief against a backdrop of isolated examples. In fact, Hopper (1997: 234) has coined the term "source conflicts" to refer to those instances when isolated sentences and data derived from textual sources do not match.

As it turns out, source conflicts are especially manifest with respect to frequency. For example, while the Japanese suffix *–teiru* is polysemous, marking both Resultative and Progressive meanings, Shirai (1997) reports that conversational *–teiru* dominates as a Resultative marker sixty-five percent of the time and signals Progressive only twenty-one percent of the time. This contrasts with the distribution of the suffix in a written elicitation task where native Japanese speakers preferred to use the suffix to mark Progressive eighty percent of the time and Resultative only nineteen percent of the time (*ibid.*). (See also Du Bois 1985; Fox 1987; Gilquin 2003, 2005; McGee, this volume; Nordquist 2004; Shirai 1990; Tao 2001; Thompson and Hopper 2001).

While source conflicts have been upheld as evidence of the inadequacy of elicited/isolated data for linguistic theorizing (e.g. Hopper 1997; Laury and Ono 2005; Sinclair 1991), Barlow (1996: 5) asks whether poor native-speaker intuitions are of any theoretical consequence. He responds affirmatively, noting that "[usage-based] models

[1] Department of Linguistics, University of New Mexico
e-mail: nordquis@unm.edu

assume that grammar formation is inductive to a large extent and *that frequency of linguistic usage events has a direct effect on the form of the grammar*" (*ibid*. emphasis added). The implication is that intuitive data are of theoretical concern because, given the role frequency is afforded in the models, intuitive data should mirror more closely the frequent structures found in usage data.

Usage-based psycholinguistic research also implies that the frequency of patterns in the two types of data ought to be more in line with one another. For example, Jurafsky (2003) summarizes a great deal of experimental data which indicate that "given multiple possible structures a speaker might say, probability may play a role in choosing among them" (40). Although Jurafsky is not addressing source conflicts, his statement nonetheless suggests, at least as a plausible hypothesis, that speakers would rely on their experience, as encoded in their grammar, to inform their choices in non-communicative contexts such as elicitation tasks. Bybee and Hopper (2001: 19) also allude to this possibility, commenting that "intuitions could be based on the user's experience with language rather than on an abstract grammar autonomous from language function and use".

The position taken in this paper, then, is that source conflicts, rather than merely a methodological dilemma, are in fact a theoretically-pressing phenomenon for usage-based linguistics: if frequency of use, as identified in corpus studies, is assumed to be indicative of storage units, the absence of such units in elicitation demands explanation.

When viewed from this perspective, the discrepancy between the two data types presents an opportunity to formulate explicit predictions about the relationship between frequency, storage, and access, offering a new testing ground for usage-based theories.

2. Research methodology and theoretical assumptions

While there are many aspects of language which could serve as a resource for probing native-speaker knowledge vis-à-vis corpus data, this study focuses on examining frequent American English collocations in a spoken corpus and the source conflicts that appear in an elicitation task with respect to these collocations.

2.1. Definition of 'collocation'

Gries (to appear) suggests six criteria be considered when identifying phraseological units, collocations being one such example. Using those six criteria, the term *collocation* is defined here as:

i) *nature of elements*: a collocation is the co-occurrence of a key word and any other linguistic element. In some cases, an element is an open slot. For example, *I'm looking forward to X* is counted as a collocation for the key word *forward*.

ii) *number of elements*: a collocation is the co-occurrence of a key word with one or more other linguistic elements (lexical and/or grammatical).

iii) *frequency*: a collocation is the most frequently repeated co-occurrence string for a key word. This criterion will be discussed further in Section 2.3.1.

iv) *permissible distance between elements*: a collocation is the most frequent pattern identified in an approximate 4:4 word span of a key word's concordance

lines (Jones and Sinclair, 1974). However, because speech fillers, interjections, and transcription practices can easily take up much of that window, this span was used as a guide.

v) *degree of flexibility of elements*: a collocation may exhibit morphological flexibility. For example, *the thing that bothers me* and *the things that bother me* were counted as the same collocation.

vi) *semantic unity and semantic non-compositionality*: a collocation represents a semantic unit. For example, while the key word *bucks* collocates more often with *to* than with *big*, *bucks to* is not a semantic unit whereas *big bucks* is. A collocation does not have to exhibit non-compositionality.

2.2 Data collection

Per the criteria outlined in the last section, a collocation is the most frequent expression that contains a key word in co-occurrence with at least one other linguistic element in a restricted span, allowing for morphological flexibility and representing a semantic unit whose meaning is not necessarily non-compositional. This definition guided the corpus data collection phase of the study.

Using the Switchboard Corpus of Recorded Telephone Conversations (Godfrey and Holliman 1997), collocations were identified for twelve different key words. These words were then used as stimuli prompts in the elicitation phase of the study. Table 1 provides the list of word prompts and their associated target corpus collocations.

Prompt	**Corpus Collocation**
necessarily	*that's not necessarily true*
plenty	*plenty of time*
perfectly	*perfectly happy*
bucks	*big bucks*
end up	*END up*[2] *with X*
glad	*I'm glad; glad to see*[3]
basis	*on a regular basis*
boring	*kind of boring*
begin	*to begin with*
someplace	*someplace else*
forward	*I'm looking forward to X*
bother	*THING that BOTHER me*

Table 1: List of Prompts and Target Corpus Collocations

Fifty-four native speakers enrolled in two Introductory Linguistics courses at the University of New Mexico participated in the elicitation experiment. Each participant was asked to read through a list of twelve index cards, randomly sorted, and to use the word on each index card as quickly as possible in an example utterance.

It is important to note that this methodology attempts to elicit procedural linguistic knowledge rather than propositional or factual knowledge about language use since usage-

[2] Capitalization is used to indicate a lemma.

[3] During the corpus analysis phase, two different collocations emerged for *glad*—one in which the material before *glad* was counted and another in which the material following *glad* was counted.

based linguists believe much of language production and comprehension is part of implicit memory (e.g. Ellis 2002; see also Bybee 1998; Boyland 1996). Furthermore, it seems that propositionally-oriented tasks may encourage source conflicts (Fox 1987: 144; Biber *et al.* 1998: 41; De Beaugrande 1999: 247; Kennedy 1991: 95-6; Stubbs 1995: 219). Therefore, speakers in this study were *not* instructed to provide what they believed was the most frequent collocation for a prompt, but merely to use the prompt with the first thing that came to mind. This was done with the intent of maximizing responses that tap procedural knowledge rather than propositional knowledge about language.

2.3 Theoretical assumptions

2.3.1 Repetition and holistic storage

To explain source conflicts with respect to collocations, this study assumes that repetition is one of the mechanisms which shapes linguistic storage units, with the frequent collocations listed in Table 1 arguably having separate representation in the mental lexicon. In other words, collocations are like frequent morphologically-complex words which have been argued to be holistically represented (e.g. Alegre and Gordon 1999; Bybee 1985; Stemberger and MacWhinney 1988). As such, frequent collocations are believed to constitute a single choice in processing (i.e. Sinclair's [1991] Idiom Principle) despite outward appearances of having been generated according to abstract rules of morpho-syntax (Barlow 2000; Bybee 1995; Lamb 2000; Langacker 2000; *inter alia*). Therefore, as shown in Figure 1, the collocation *big bucks* has its own mental representation stored alongside its constituent words.

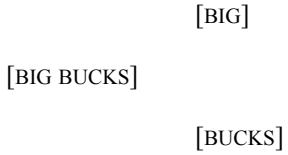

Figure 1: Holistic Storage of Collocations

Although frequency was used to identify the target corpus collocations, it should be noted that an absolute frequency measure was not a prerequisite for unit status in this study. This was decided for two reasons. First, a frequency threshold for positing holistic storage for collocations remains to be empirically determined (see also the discussion in Section 4.1 which makes it clear that frequency is not the only mechanism by which collocations come to be holistically stored in a usage-based model). Secondly, and perhaps more importantly, the research hypothesis simply asks whether speakers draw upon their most frequent collocational experience with a word, making a token frequency threshold unwarranted at this early stage of investigation.

2.3.2 Lexical strength and autonomy

Another theoretical assumption adopted to explain source conflicts with respect to collocations is *lexical strength.* As Bybee (1985) explains, the more often a unit is used, the stronger its mental representation, or the greater its lexical strength. This storage feature alone suggests that a word's most frequently stored collocation should be produced fairly often in

an elicitation task. That is, a representation with greater lexical strength has a higher resting activation level than other less frequent representations with weaker lexical strength.

However, representations that have greater lexical strength are also likely to undergo automatization effects (Bybee 1985; or *entrenchment*, e.g. Langacker 2000) such that semantic and/or phonological connections to other representations are weakened or lost. As a result, multimorphemic units with greater lexical strength are prone to losing their internal structure (e.g. Bybee and Thompson 2000) since the constituents contained within the larger autonomously-represented unit are not strengthened or preserved via feedback from connections to other instances. In turn, a type of "creeping double articulation" may result (Haiman 1998) such that the constituent parts of a collocation no longer represent fully independent words, even though they may continue to be orthographically represented as independent units (Sinclair 1987: 321).

Consequently, because of their relatively high frequency, the target collocations in this study are generally assumed to be autonomous (or semi-autonomous) and therefore have weaker and/or fewer connections to the individual representations for the constituent words that make them up (see also the discussion in Section 4.1 which provides additional evidence for autonomy). For example, in Figure 2, the collocation *big bucks* is represented with weak connections (i.e., dashed lines) to the individually-stored representations for the adjective *big* and for the noun *bucks*.

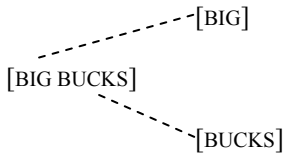

Figure 2: Autonomous Representation of Collocations

2.3.3 Hypothesis

Autonomy, therefore, has profound implications for what we predict about the elicitation of target collocations: individual constituent words of an autonomously-stored collocation do not provide the best access to the mental representation of the target collocation. Although the target collocation may have greater lexical strength than other stored collocations, and therefore be primed for quick access, higher frequency units are also more likely to have autonomous representation. In terms of eliciting for collocations via their constituent words, then, elicitation is predicted to contain few target collocation responses. As we will see in the next section, the results of the elicitation task are consistent with this hypothesis.

3. Data and results

Table 2 shows that elicited responses adhere overall to the hypothesis in that the most frequent collocation for a prompt is not typically used to complete the experimental task. In fact, for eleven of the twelve prompts, the target collocations represent approximately twenty percent or less of the elicited data (i.e. the first three rows of Table 2). For instance, only twelve out of fifty-four people (twenty-two percent) produced *big bucks* when prompted with the word *bucks*, even though *big bucks* is the most frequent collocation in the corpus.

Perhaps even more striking are the data for the prompts *forward* and *basis*. For example, only three speakers uttered *I'm looking forward to X* (representing approximately

six percent of the elicited data for *forward*), and only one speaker produced *on a regular basis* (representing two percent of the elicited data for *basis*). This almost complete absence of target collocations in elicitation is consistent with the hypothesis that the target collocations are autonomously stored and that isolated presentation of the prompt does not strongly activate these collocations despite the fact that *I'm looking forward to X* and *on a regular basis* represent respectively twenty-eight percent and twenty percent of the *forward* and *basis* tokens in the Switchboard corpus.

Nonetheless, as Table 2 shows, the elicited data for *glad* do not conform to the hypothesis since the target collocation *I'm glad* accounts for sixty-five percent of the elicited data, representing a clear outlier. Before turning to an analysis of *glad*, however, it is first necessary to consider two other possible storage hypotheses that also predict a general absence of target collocations in elicitation.

Number of Prompts	Target Collocations	Percentage of Elicited Data
4	*kind of boring; to begin with; thing that bother me; perfectly happy;*	0 percent
5	*end up with; on a regular basis; I'm looking forward to X; glad to see; not necessarily true*	2 percent – 10 percent
3	*someplace else; big bucks; (plenty of time)* [4]	≈ 20 percent
1	*I'm glad*	65 percent

Table 2: Results of Elicitation Task

4. Psycholinguistic explanations

Three storage hypotheses are listed in (1) – (3). The Autonomy hypothesis listed in (3), of course, is the hypothesis currently being investigated. The alternative hypotheses assume respectively that 1) the target collocations are not frequent enough to have been stored in the lexicon at all; or, 2) if stored, the target collocations are not frequent enough to have been activated during the elicitation task.

(1) No Mental Representation Hypothesis: Collocations may be stored in the lexicon, but the target collocations in this study are not frequent enough to have separate lexical representation and therefore were not actual contenders in the activation phase of this task.

(2) Infrequency Hypothesis: Collocations of varying frequencies are in fact stored in the lexicon, but the target collocations investigated in this study are not frequent enough to have significant lexical strength; as a result, these collocations' representations have resting activation rates below threshold, making them unlikely to be easily or quickly activated during elicitation.

(3) Autonomy Hypothesis: Frequent collocations are stored as separate units in the lexicon, and undergo automatization effects. As a result, the constituent words lose their independent status in the collocation,

[4] The chi-square value for the distribution of collocation responses in elicitation is significant for ten of these prompts; while the collocation data for *plenty* is in the predicted direction, it does not reach significance.

and the individual words of the target collocation are not necessarily activated when a word prompt is presented in elicitation.

Given that all three hypotheses predict that the target collocations will not generally be used in elicitation, what evidence is there that the Autonomy Hypothesis should be adopted since the elicited data are predicted by all three?

4.1 No mental representation hypothesis

Because the proportional frequencies for some of the target collocations represent under ten percent of the prompts' total *corpus* tokens, it could be argued that these collocations are not really stored in the lexicon. In other words, there may be some frequency threshold which must be surpassed before a collocation can be stored in the lexicon. Alegre and Gordon (1999) report six occurrences per million as a threshold for lexical storage. Based on such a threshold, only six of the collocations in this study (those for *BOTHER, basis, BEGIN, END up, forward* and *glad[preceding]*) would be considered frequent enough to have lexical storage in memory. However, Alegre and Gordon studied regularly inflected verb forms, not phraseological units, such as collocations, which often exhibit delexicalization and/or unproductive grammatical properties, hence requiring rote learning and storage. For instance, idioms are stored in memory because they often retain archaic lexical items or obsolete grammar which obviously must be memorized as part of the expression (e.g. Bybee 1998). Therefore, if the collocations considered in this current work exhibit semantic or grammatical characteristics that are similar to those characteristics we observe in idioms, or other prefabricated language, we can assume their holistic storage as well.

Because space limitations prevent an exhaustive analysis of each collocation, only a few examples are presented here for illustration purposes (see Nordquist 2006: Chapter 6 for a fuller discussion). We can start by first noting that many of the collocations in this study have meanings which are not derivable from the sum of their parts and would thus "appear to be processed without recourse to their lowest level of composition" (Wray 2002: 4). For instance, the collocation *I'm looking forward to X* does not have a compositional meaning of 'forward-facing eye gaze', but rather a metaphorical meaning which signals anticipation of a future event. While non-compositional semantics was not a criterion for inclusion in the study (see Section 2.1), many of the collocations nonetheless exhibit non-compositionality as predicted by the autonomy argument. In light of this, it seems unlikely that these target collocations are generated anew for each usage event but are in fact stored holistically in the lexicon.

In addition to displaying non-compositional semantics, idioms and formulaic language often exhibit frozen syntax and do not participate in productive grammatical processes (e.g. Erman and Warren 2000). Again, this is generally true of the collocations in this study. For example, the collocation *THING that BOTHER me* performs a topic-introducing function which is tied to the construction's syntax. That is, the collocation cannot be passivized and still introduce a topic: *?I am bothered by the thing*. Perkins (1994: 328) makes a similar observation for the formula *What's the use of Xing?* which does not have a corresponding declarative form that retains the same semantics of despondency (e.g., *the use of Xing is…*), thereby necessitating its holistic storage.

Grammatical restrictions may also be more local in nature as we see in (4) for the collocation *on a regular basis*.

(4) a. *on ?these regular ?bases*
b. *on ?my regular basis*
c. *on a regular basis ?to which we agree*

The general constraint against manipulating *basis* morpho-syntactically should come as no surprise, however, since *basis* is a largely *dependent* form within the holistically-stored collocation. In fact, *basis* is perhaps more adverbial than nominal in the expression, rarely functioning as a discourse manipulable entity (Hopper and Thompson 1984), even though it fills a traditionally recognized nominal slot.

By and large, then, the collocations considered in this study have features which suggest that they are separately stored units in the lexicon. The non-compositionality and grammatical invariance of the collocations serve as an independent measure of their holistic storage since it is largely agreed that we wholly store units that are very infrequent in the language but are idiomatic in some way. Crucially, without such storage, the idiosyncrasies discussed above could not have accrued to the representation (Bybee 2006). We can therefore dismiss the argument in (1) that these collocations are not stored in the lexicon even if an objection about their textual frequency is made (i.e. the No Mental Representation Hypothesis).

4.2 Lack of activation hypotheses

Having argued for the mental storage of these collocations, we are left with determining whether the collocations are too infrequent to be reasonably activated in the elicitation task (i.e. the Infrequency Hypothesis listed in [2]); or, if the collocations are in fact proportionally frequent enough to be activated but are not good competitors due to other storage factors (i.e. the Autonomy Hypothesis listed in [3]). According to the Infrequency Hypothesis, these collocations are in fact stored in the lexicon, much like infrequent idioms (e.g. *let the cat out of the bag*), but they are too infrequent to compete with other representations that are also stored in the lexicon. In the alternative scenario, the Autonomy Hypothesis put forth here, the most frequent collocation for a given prompt *should* be frequent enough to be activated in an elicitation task; however, it is not successfully retrieved because of autonomous storage. In order to evaluate these two claims, we need to consider more specifically how the hypotheses differ.

First, both hypotheses must account for why some speakers do in fact use the most frequent collocation in elicitation. Why, for example, did twelve speakers utter *big bucks* when prompted with *bucks*? Under the Infrequency Hypothesis, the holistic storage of the collocation is accepted, but its lexical strength is argued to be too weak for activation. As a result, we are forced to assume that twelve speakers accessed an appropriate construction (e.g. [PREMODIFIER *bucks*]) and then coincidentally filled the open slot with *big*. However, this is not intuitively convincing, especially if we recognize that *big bucks* is a recurring unit for speakers. In other words, if speakers produce *big bucks* in elicitation, it is probably because they accessed it holistically.

The larger issue, however, is that an empirically-determined threshold—below which a collocation could be argued to be too weak to effectively compete—has not been established. While this issue cannot be resolved here, it is clear that elicitation is not completely devoid of collocations or prefabricated language. Thus, other collocations which are less frequent than the target collocations did surface in the elicited data.

For example, in the elicited data for the prompt *forward*, ten utterances contained *fast forward*. This represents a considerable proportion of the elicited data for a collocation which

is not relatively frequent in our collective conversational experience.[5] Additionally, one participant uttered *they were forward in their thinking* and another replied with *I will forward the information to you*. *Fast forward, forward in their thinking* and *forward the information to you* are all recognizable collocations in English which are sensitive to the kinds of tests discussed in Section 4.1. However, none of these collocations were more frequent in the Switchboard corpus than *I am looking forward to X*, which represents twenty-eight percent of the *forward* corpus tokens. If the Infrequency Hypothesis maintains that the most frequent collocation is too infrequent for activation, then it should also predict that these even less frequent corpus collocations should not appear in elicited data. Furthermore, it seems reasonable to expect that a collocation which represents more than a quarter of speakers' experiences with the word *forward* is in fact frequent enough to be activated for competition in the elicitation task. The fact that *I'm looking forward to X* is not produced needs explanation, but that explanation will not be found in the Infrequency Hypothesis.

A similar situation holds for the elicited data for *basis*. The most frequent corpus collocation (*on a regular basis*) was used only once in elicitation, but other collocations, which are not at all frequent in the corpus, were also elicited (e.g. *on the basis of their color* or *on a need-to-know basis*). The fact that these collocations were available to speakers during the elicitation task makes it difficult to accept that *on a regular basis* was too infrequent to compete with other representations. And, if twenty-eight percent strikes us as a large enough proportion of a word's corpus tokens to influence elicitation, we might expect that when a fifth (twenty percent) of a word's tokens is represented by a single collocation (e.g. *on a regular basis*), that collocation, too, is frequent enough to be activated in elicitation.

Of course, it may be that there is in fact some threshold below which the relative frequency of a collocation for a particular word is too low for the collocation to be realistically called upon in elicitation even if the collocation does represent the most frequent usage for a prompt. Given that it is not clear where to draw that line with respect to these data, the Autonomy Hypothesis may fade into the Infrequency Hypothesis. Because the predicted empirical results would largely be the same, it is difficult to tease these apart. For now, though, it would appear that while the Infrequency Hypothesis cannot be completely discounted, it has been brought into question as a viable explanation for the lack of the most frequent collocations in these elicited data. In the next section, the Autonomy Hypothesis finds even more support in that its effects are easily captured within an independently developed model.

4.3 Modelling source conflicts in elicitation

The Interactive-Activation Connectionist model developed by McClelland and Rumelhart (1981) offers an independent way of explaining source conflicts in elicited data with respect to collocations. Although the model's architecture contains nodes for specific words, it can easily be expanded to contain nodes for stored collocations. The model also has nodes for morpho-syntactic and semantic features, and the model allows connections between nodes. Activation proceeds through excitatory and inhibitory messages sent along links between various nodes. These connections can be excitatory if two nodes suggest each other's existence, or the connections can be inhibitory if two nodes are inconsistent with one another. Excitatory messages increase the activation of a node while inhibitory ones decrease the level

[5] This is not to say that the collocation is not salient in our culture. Indeed, especially today, technology has made *fast forward* a unit. According to the corpus data, however, speakers have experienced *I'm looking forward to X* much more often than *fast forward,* the latter of which represents only four percent of the corpus tokens.

of activation. Because processing occurs in a parallel fashion, the sum of all excitatory and inhibitory connections for a specific node indicates whether that node is activated or not.

Figure 3 is a simplified adaptation of this model, using the eight most frequent corpus collocations for *bucks* as likely competitors in the elicitation task.[6] The diagram also includes two semantic feature nodes (i.e. NUMERAL and OTHER MODIFIER) that are mutually exclusive. The green arrows from the collocation nodes to the semantic feature nodes represent excitatory connections while the orange arrows represent inhibitory connections. The bigger and bolder node for the *big bucks* collocation represents its greater textual frequency and hence higher initial resting activation level vis-à-vis the other stored collocations.

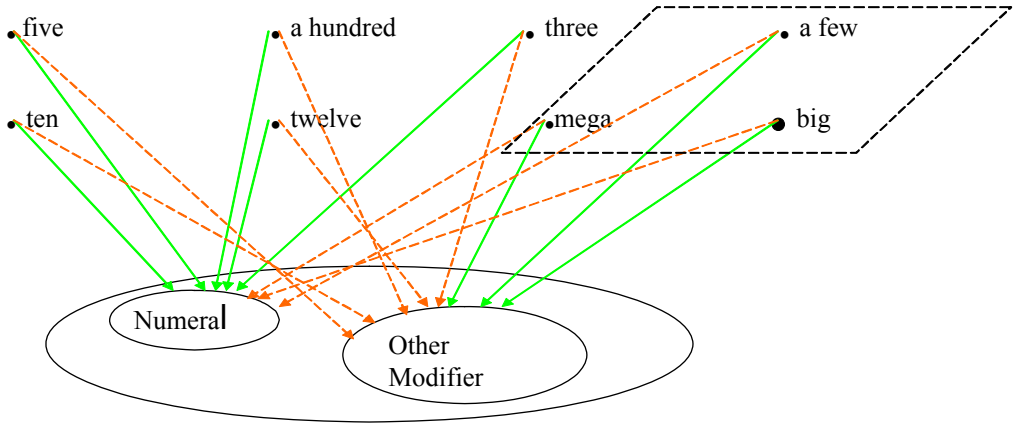

Figure 3: Modelling the Interactive-Activation Connectionist Model for *bucks* Collocations

The distribution of these connections reveals two groups of collocation nodes with *big bucks, mega bucks* and *a few bucks* forming one group, and *five bucks, ten bucks, a hundred bucks, twelve bucks*, and *three bucks* forming another. However, the distribution of the connections across the linguistic features is not equal but rather skewed. While the feature node for OTHER MODIFIER has three excitatory connections, it also has five inhibitory connections. In contrast, the NUMERAL node has five excitatory connections and only three inhibitory connections.

McClelland and Rumelhart (1981: 395) explain that when this kind of skewed distribution of excitatory and inhibitory connections occurs, gang effects are possible where a set of mutually excitatory nodes work together to reinforce feature nodes and also produce much stronger reinforcement for themselves.

For example, in their model of orthographic symbol recognition, there is a [MA_E] gang (*make, male, etc.*) and an [_AVE] gang (*save, have, etc.*). While these two gangs contain six members each, another [M_VE] gang has only one member, *move*. Consequently, when the probe *mave* is presented for word recognition, the textually less frequent *save* node ultimately receives more activation than the textually more frequent *move* node; *save* pulls ahead in the cohort, so to speak, due to a gang effect where gang words "work together to reinforce the [letter] nodes, thereby producing much stronger reinforcement for themselves"

[6] The cohort is undoubtedly much larger upon initial activation of *bucks*. For example, *buckskin* or *Starbucks* would presumably be initially activated just as *big bucks* or *five bucks* is. For the purposes of illustration, though, only a few nodes for collocations have been selected. Of course, the actual composition of the activated cohort is unknown, and the items listed in Figure 3 are motivated by textual frequency counts.

(McClelland and Rumelhart 1981: 395; see also McClelland 1981 and who simulates the same effect).

Stated another way, *save*'s activation level is bolstered by the fact that it has letter nodes which are highly activated by the other members of its gang: the letter nodes for *a, v* and *e* receive activation from other gang members, and because these letters are mutually excitatory with *save*, the node for *save* receives a type of feedback activation. On the other hand, while *move* is initially activated within the cohort upon presentation of *mave*, it belongs to a gang of one, so it does not receive additional feedback to maintain its initial high resting activation level.

Therefore, in terms of activation during elicitation, the collocation *ten bucks*, although textually less frequent than *big bucks*, may benefit from membership in a [NUMERAL *bucks*] gang whose members jointly excite the same features which then serve to reinforce the collocation nodes themselves. However, *big bucks*, by virtue of not being a member of this gang, does not benefit from the same kind of effect. In fact, despite an initial advantage due to an above-threshold resting activation state, *big bucks* is actually suppressed during the elicitation task just like *move* is in the orthographic recognition simulation. A reflex of this proposed gang effect is witnessed in elicitation where infrequent *ten bucks* fares about as well as *big bucks*, occurring eight times in the elicited data (versus twelve tokens of *big bucks*), suggesting that competition gang effects may level frequency effects in elicitation.

Moreover, as discussed in Section 2.3.2, higher frequency items will typically be autonomous or semi-autonomous and will therefore often exhibit non-compositional semantics and/or unproductive grammar. As a result of having such idiosyncratic linguistic features, autonomous representations will crucially have excitatory connections to feature nodes which are not strongly activated alongside a cohort of other activated collocations. This means that the most frequent collocations for a given word will typically be excluded from the more fully activated set of collocations. A gang effect will then typically suppress the high resting activation level for a word's most frequent collocation and make the target collocation an unlikely choice in elicitation.

This model, then, provides a plausible architecture for capturing the autonomous nature of target collocations and hence the somewhat surprising finding that speakers do not typically use a word's most frequent collocation in completing the elicitation task. However, as it turns out, the notion of gang effects also captures the unexpected results for the data on *glad* where speakers used the target collocation *I'm glad* sixty-five percent of the time in elicitation.

Figure 4 shows the partial spreading activation assumed during elicitation for the top eight *glad* corpus collocations. Again, green lines indicate excitatory connections to feature nodes, and the lexical strength of *I'm glad* is represented with a larger, darker node.

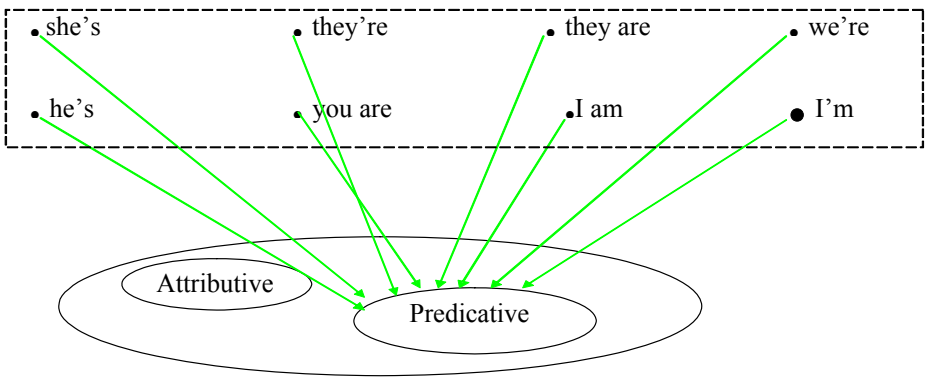

Figure 4: Modelling the Interactive-Activation Connectionist Model for *glad* Collocations

When the pattern of excitatory links shown in Figure 4 is taken into account, we find that the instance node for *I'm glad* is contained within, rather than excluded from, the gang that wins. Thus, the highly activated PREDICATIVE gang (eight excitatory versus zero inhibitory links) contains the instance node for *I'm glad*. This means that *I'm glad*, unlike *big bucks,* may benefit from a gang effect, making it a strong competitor in the elicitation task. Furthermore, the lexical strength of *I'm glad* will make it even more accessible due to a higher resting activation rate. As a result, we should perhaps anticipate *I'm glad* to dominate in the elicitation task, as it does, occurring in sixty-five percent of the elicited utterances.[7]

5. Conclusions

Based on the results of the elicitation experiment, we find fairly strong support for the claims made in this study: the tendency to not produce a word's most frequent collocation is consistent with the usage-based principles of holistic storage of frequent multimorphemic units and their autonomous representation.

This finding is perhaps strengthened by the fact that a variety of target collocations were included in the experiment. For example, *bucks* targets *big bucks,* a simple noun phrase, while *BOTHER* targets *THING that BOTHER me*, a complex noun phrase that serves a discourse function. Also, per Erman and Warren's (2000) semantic classification of prefabricated language, some target collocations evoke properties and states (e.g. *kind of boring*) while others refer to situations and events (e.g. *I'm looking forward to X*), and others yet indicate places and positions (e.g. *someplace else*), or a period or point in time (e.g. *on a regular basis)*. That the pattern of source conflicts held up across various types of target collocations is suggestive of their shared storage properties since they clearly do not share grammatical or semantico-pragmatic features.

However, caution must be exercised. Corpus data may not always be indicative of mental units. For example, Schmitt, Grandage, and Adolphs (2004) report that speakers who were asked to repeat frequent corpus clusters in a dictation task displayed disfluencies in their target productions, suggesting that the clusters are not holistically stored despite their high frequency in corpora; or if they are stored, they are stored on an individual by individual basis. Indeed, Barlow (2005) has argued that corpus data can not be used to posit production routines since amalgamated frequency data wash out individual speaker production preferences.

Finally, there is little question that the mechanisms which drive holistic storage are multi-faceted. Frequency is just one mechanism by which collocations come to be holistically-stored, and there are undoubtedly other factors, in addition to frequency, which may have an obscuring effect and lead to source conflicts.

For example, McGee (this volume) suggests that the structure of the target collocation influences elicitation responses, reporting that an adjective stimulus is more likely to evoke its most frequent noun collocate if the adjective and noun form an independent ADJECTIVE-NOUN dyad (e.g. *good idea*). On the other hand, an adjective stimulus is not likely to evoke its most frequent noun collocate if the adjective and noun are part of a larger unit or frame-like structure (e.g. ADJECTIVE NOUN *of* NP: *<u>different</u> types of* ...). While it is true that many of the target collocations in this study represent frame-like expressions rather than stand-alone dyadic collocations, it is also the case that the current study's elicitation methodology did not discourage frame-like responses (see Section 2.3). Therefore, it is difficult to ascertain whether speakers avoided these collocations because the stimulus was contained within a

[7] See Nordquist (2006: Chapter 6) for a fuller analysis of the data on *glad* and for an analysis of *plenty* within this connectionist model.

frame-like collocation or because the collocation is autonomous. We can note, however, that the infrequency of the dyad *big bucks* in elicitation suggests that an autonomy obscuring effect is at least partially involved.[8]

The post-hoc analysis of *I'm glad* predicts that elicitation is also influenced by competing representations and overall activation patterns within the lexicon, so that ultimate retrieval of a unit is dependent upon the nature of other activated representations as well. This latter point is worth investigating since other independent research suggests that the composition of an activated cohort has a role to play in understanding how stored representations compete during access (e.g. Luce, Pisoni and Goldinger's 1990 review).
As a result, stored collocations, which are posited to be word-like, might also be subject to neighborhood density effects.

Future work will need to address these other potentially influential factors as we continue to investigate elicited data and what its patterns reveal about the mental lexicon, storage, and processing.

References

Alegre, M. and P. Gordon (1999) 'Frequency effects and the representational status of regular inflections'. *Journal of Memory and Language 40*, 41–61.

Barlow, M. (1996) 'Corpora for theory and practice'. *International Journal of Corpus Linguistics 1*, 1-37.

Barlow, M. (2000). Usage, Blends and Grammar, in M. Barlow and S. Kemmer (eds) Usage-Based Models of Language, pp. 315–45. Stanford: CSLI.

Barlow, M. (2005) Input grammars and output grammars: Investigating the language of individual speakers. Paper presented at the Third Corpus Linguistics Conference. University of Birmingham, 14–17 July 2005.

Biber, D., S. Conrad, and R. Reppen (1998) Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Boyland, J. T. (1996) Morphosyntactic Change in Progress: A Psycholinguistic Treatment. Doctoral Dissertation. Berkeley, CA: University of California at Berkeley.

Bybee, J. L. (1985) Morphology: A Study of the Relation between Meaning and Form. Amsterdam: John Benjamins Publishing Company.

Bybee, J. L. (1995) 'Regular morphology and the lexicon'. *Language and Cognitive Processes 10*, 425–55.

Bybee, J. L. (1998) The Emergent Lexicon, in M. C. Gruber, D. Higgins, K. S. Olson, and T. Wysocki (eds) Papers from the 34th Annual Meeting of the Chicago Linguistics Society, Part 2: The Panels, pp. 421–39. Chicago: Chicago Linguistics Society.

Bybee, J. L. (2006) 'From usage to grammar: The mind's response to repetition'. *Language 82*, 529–51.

Bybee, J. L., and P. Hopper (2001) Introduction to Frequency and the Emergence of Linguistic Structure, in J. Bybee and P. Hopper (eds) Frequency and the Emergence of Linguistic Structure, pp. 1–24. Amsterdam: John Benjamins Publishing Company.

Bybee, J. and S. Thompson (2000) 'Three frequency effects in syntax'. *Berkeley Linguistics Society 23*, 378–88.

De Beaugrande, R. (1999) 'Reconnecting real language with real texts: Text linguistics and corpus linguistics'. *International Journal of Corpus Linguistics 4*, 243–59.

[8] McGee (this volume) notes, however, that adjectives more reliably elicit nouns than vice-versa in free association tasks.

Du Bois, J. W. (1985) Competing Motivations, in J. Haiman (ed.) Iconicity in Syntax, pp. 343–65. Amsterdam: John Benjamins.

Ellis, N. C. (2002) 'Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition'. *Studies in Second Language Acquisition 24*, 143–88.

Erman, B. and B. Warren (2000) 'The idiom principle and the open choice principle'. *Text 20*, 29–62.

Fox, G. (1987) The Case for Examples, in J. M. Sinclair (ed.) Looking Up, pp. 137–49. London: Collins ELT.

Gilquin, G. (2003) Prototypicality: Corpus vs. elicitation. Paper presented at ICLC–8. University of La Rioja, Spain, 20–25 July 2003.

Gilquin, G. (2005) What you think ain't what you get: Highly polysemous verbs in grammar and mind. Paper presented at From Gram to Mind Conference. Université de Bordeaux, 19–21 May 2005.

Godfrey, J. J. and E. Holliman (1997) Switchboard-1 Release 2. Philadelphia: Linguistic Data Consortium.

Gries, S. Th. (to appear) Corpus Data in Usage-Based Linguistics: What's the Right Degree of Granularity for the Analysis of Argument Structure, in M. Brda and M. Žic Fuchs (eds) Expanding Cognitive Linguistic Horizons. Amsterdam: John Benjamins Publishing Company.

Haiman, J. (1998) Talk is Cheap: Sarcasm, Alienation, and the Evolution of Language. Oxford: Oxford University Press.

Hopper, P. (1997) When 'Grammar' and Discourse Clash: The Problem of Source Conflicts, in J. Bybee, J. Haiman, and S. Thompson (eds) Essays on Language Function and Language Type, pp. 231–47. Amsterdam: John Benjamins Publishing Company.

Hopper, P. J. and S. A. Thompson (1984) 'The discourse basis for lexical categories in universal grammar'. *Language 60*, 703–52.

Jones, S. and J. M. Sinclair (1974) 'English lexical collocations'. *Cahiers de Lexicologie* 24, 5-61.

Jurafsky, D. (2003) Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production, in R. Bod, J. Hay, and S. Jannedy (eds) Probabilistic Linguistics, pp. 39–95. Cambridge, Mass: MIT Press.

Kennedy, G. (1991) *Between* and *Through*: The Company They Keep and the Functions They Serve, in K. Aijmer and B. Altenberg (eds) English Corpus Linguistics: Studies in Honour of Jan Svartvik, pp. 95-110. London: Longman.

Labov, W. (1972) 'Some principles of linguistic methodology'. *Language in Society 1*, 97–120.

Lamb, S. (2000) Bidirectional Processing in Language and Related Cognitive Systems, in M. Barlow and S. Kemmer (eds) Usage-Based Models of Language, pp. 87–119. Stanford: CSLI.

Langacker, R. (2000) A Dynamic Usage-Based Model, in M. Barlow and S. Kemmer (eds) Usage-Based Models of Language, pp. 1–63. Stanford: CSLI.

Laury, R., and T. Ono (2005) 'Data is data and model is model: You don't discard the data that doesn't fit your model!' *Language 81*, 218–25.

Luce, P. A., D. B. Pisoni, and S. D. Goldinger (1990) Similarity neighborhoods of spoken words, in G. T. M. Altmann (ed.) Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives, pp. 122–47. Cambridge, MA: MIT Press.

McClelland, J. L. (1981) Retrieving General and Specific Information from Stored Knowledge of Specifics, in Proceedings of the Third Annual Conference of the Cognitive Science Society, pp. 170–72. Berkeley, CA.

McClelland, J. L. and D. E. Rumelhart (1981) 'An interactive activation model of context effects in letter perception: Part I. An account of basic findings'. *Psychological Review 88*, 375–407.

McGee, I. (this volume) Teachers' Lexical Intuitions versus Corpus Data: Differences, Similarities and Explanations.

Nordquist, D. (2004) Comparing Elicited Data and Corpora, in M. Achard and S. Kemmer (eds) Language Culture and Mind, pp. 211–23. Stanford: CSLI Publications.

Nordquist, D. (2006) Corpus Patterns and Elicited Language: Implications for Language Storage and Processing. PhD Dissertation. Albuquerque, NM: University of New Mexico.

Perkins, M. (1994) 'Repetitiveness in language disorders: A new analytical procedure'. *Clinical Linguistics and Phonetics 8*, 321–36.

Schmitt, N., S. Grandage, and S. Adolphs (2004) Are Corpus-Derived Recurrent Clusters Psycholinguistically Valid?, in N. Schmitt (ed.) Formulaic Sequences: Acquisition Processing and Use, pp. 127–51. Amsterdam: John Benjamins Publishing Company.

Shirai, Y. (1990) 'Putting PUT to use: Prototype and metaphorical extension'. *Issues in Applied Linguistics 1*, 78–97.

Shirai, Y. (1997) 'On the primacy of progressive over resultative state: The case of Japanese –*teiru*'. *Japanese/Korean Linguistics 6*, 512–24.

Sinclair, J. McH. (1987). Collocation: A Progress Report, in R. Steele and T. Threadgold (eds) Language Topics: Essays in Honour of Michael Halliday, Volume II, pp. 319–31. Amsterdam: John Benjamins Publishing Company.

Sinclair, J. McH. (1991) Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Stemberger, J. P., and B. MacWhinney (1988) Are Inflected Forms Stored in the Lexicon?, in M. Hammond and M. Noonan (eds) Theoretical Morphology: Approaches in Modern Linguistics, pp. 101–16. San Diego: Academic Press, Inc.

Stubbs, M. (1995) 'Collocations and semantic profiles: On the cause of the trouble with quantitative studies'. *Functions of Language 2*, 23–55.

Tao, H. (2001) Discovering the Usual with Corpora: The Case of *Remember*, in R. C. Simpson and J. M. Swales (eds) Corpus Linguistics in North America: Selections from the 1999 Symposium, pp. 116–44. Ann Arbor: The University of Michigan Press.

Thompson, S. A., and P. J. Hopper (2001) Transitivity, Clause Structure, and Argument Structure: Evidence from Conversation, in J. Bybee and P. Hopper (eds) Frequency and the Emergence of Linguistic Structure, pp. 27–60. Amsterdam: John Benjamins Publishing Company.

Wray, A. (2002) Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Combining Corpus and Experimental Data to Capture Idiomaticity

Stefanie Wulff[1]

Abstract

It is a by now established fact that idiomaticity cannot be equated with non-compositionality alone, but is a complex concept that is also associated with various aspects of formal flexibility. This raises the question to what extent speakers call up these different factors when judging the overall idiomaticity of a phrase.

In the present paper, experimental and corpus-linguistic methodologies are combined to address this question. For a total of thirty-nine V NP-idioms of the kind *make a point* or *take the plunge*, comprising more than 13,000 tokens overall, their compositionality, syntactic, lexico-syntactic, and morphological flexibility were assessed corpus-linguistically. The corpus-based results thereby obtained were then correlated with native speakers' overall idiomaticity judgments in a multiple regression analysis to determine each factor's impact on the overall judgments. The results indicate that speakers indeed rely on multiple factors simultaneously, with lexico-syntactic and morphological factors being even more important than compositionality, and verb-related being more important than NP-related information. Overall, the results back up the theoretical concept of a collocation-idiom continuum, and demonstrate how various, and sometimes competing, motivations determine a phrase's position on this continuum.

[1] University of Bremen
e-mail: stefaniewulff@gmail.com