This is kind of / sort of interesting: variation in hedging in English

Stefan Th. Gries, University of California, Santa Barbara

Caroline V. David, Université Paul-Valéry, Montpellier III

Abstract

One of the domains where corpus linguistics has been particularly successful is the analysis of variation in the choice of lexical items that is governed by the context around the slot into which one out of several functionally similar lexical items is to be inserted. In this study, we investigate the variation found in two near synonymous hedging expressions - *kind of* and *sort of* - on the basis of data from contemporary British English. We first retrieved all instances of *kind of* and *sort of* from the British National Corpus World edition. As a second step, we annotated each instance for:

i. the lexical item(s) that the hedging expression modified;
ii. the word class(es) instantiated by these expressions;
iii. the medium and the register of the instance.

Finally, we investigated the resulting multidimensional table using distinctive collocate/collexeme analysis (cf. Church et al. 1994, Gries 2003, Gries and Stefanowitsch 2004) and techniques for the analysis of multidimensional contingency tables to determine how and to what extent the two expressions differ. Our discussion of the results focuses on factors that govern the choice of hedge; the factors include (i) factors external to language (viz., the situationally/contextually defined register or text type of the utterance(s) in question) and (ii) factors internal to language (viz., the so far unnoticed preferences of *kind of* and *sort of* to be used together with particular lexical items and semantic fields).

1. Introduction

One particularly interesting kind of pragmatic/discourse phenomenon is the use of hedging (cf. Lakoff 1972:195 for the first mention of the term). The definition proposed by Schröder and Zimmer (2000) is the following (cf. also Markkanen and Schröder 2000):

> A hedge is either defined as one or more lexico-syntactical elements that are used to modify a proposition, or else, as a strategy that modifies a proposition. The term 'hedging' is used to refer to the textual strategies of using linguistic means as hedges in a certain context for specific communicative purposes, such as politeness, vagueness, mitigation, etc.

Well-known English hedges include *technically*, *essentially*, *more or less*, *practically*, *strictly speaking*, *kind of*, *sort of* ...

Hedges are theoretically interesting for various reasons. For example, they pose interesting challenges to logic-based semantic analyses (cf., e.g., Kay 1997 or Denison 2005); research on hedges paved/supported the way to the recognition and investigation of the now widely accepted fact that degrees of category membership can vary considerably. In addition, hedges are practically interesting for the analysis of communicative strategies (cf. Aijmer 1986, Nunberg 2004, Yaguchi et al. 2004) and in contexts of language learning, where second-language learners often face the challenge of having to infer the pragmatically highly loaded meanings and conventions of hedges.

In this paper, we will look at the two English hedges *kind of* and *sort of* to (i) point out a few shortcomings we think some previous works exhibit and (ii) begin a more comprehensive analysis of the actual usage of the two hedges that remedies the above shortcomings.

The findings resulting from previous studies of *kind of* and *sort of* can be summarized in a few groups. For example, pragmatically oriented studies such as Lakoff (1975) or Yaguchi et al. (2004) focused on how conversational contexts and/or speech settings influence the use of hedges. For example, the latter analyze 3,713 and 4,747 instances of *kind of* and *sort of* in the 2m-word Corpus of Spoken Professional American English, finding that speech settings (academic, scientific, reading committee, faculty meeting etc.), social positions other than gender, and variety (BrE vs. AmE) seem to determine the frequency of *kind of* and *sort of* as hedges. Yaguchi et al. consider these hedges as markers of how unassertively the speaker talks and taking into account the relative position or the imbalance position there exists between the speaker and the listener. Also, they discuss the degree to which the casualness of the setting in which these hedges are used correlate with the frequency of *kind of* and *sort of*.

Some other studies are largely concerned with what is at the heart of the present paper, the distribution of *kind of* and *sort of*. For example, Biber et al. (1999:560-561, 870) investigate the Longman Spoken and Written English Corpus (containing 40m words of AmE and BrE) and report that the frequency of *kind of* and *sort of* in conversation is higher than in academic prose or other registers. Also, *kind of* is much more frequent in AmE (1000+ per million) than in BrE (200+ per million) while *sort of* is equally frequent in AmE and BrE (200+ per million); cf. also Quirk et al. (1985:598). However, they also find an interaction such that *sort of* is preferred in conversation in BrE whereas *kind of* is preferred in conversation in AmE. A similar result is reported by Crystal and Davy (1975:29), according to whose analysis *kind of* is twice as frequent in AmE than in BrE.

As to studies that go beyond raw frequencies of occurrence (in varieties or registers), Kay (1997) just states *kind of* and *sort of* may occur directly to the left of any category {N,V, Adj, Adv, S, C}. A more comprehensive study - comprehensive in terms of distributional patterns, that is - is Aijmer (1984); cf. also Aijmer (1986). She investigates the 0.5m-word London-Lund Corpus and reports that *sort of* is often followed by *you know* and "[i]f we look at the distribution of *sort of* before major constituents we find that (a) *sort of* is more common before noun-phrases than before other constituents." (Aijmer 1984:121). She illustrates this distribution by the data in Table 1. [1]

Table 1. The distribution of *sort of* before major constituents (based on Aijmer 1984).

NP	PP	VP	AdjP	AdvP	Total
302	8	145	19	8	482

In addition, she states that "[w]ith *kind of* the proportion of examples modifying a VP is smaller than with *sort of*." (Aijmer 1984:121), using the data given in Table 2.

Table 2. The distribution of *kind of* before major constituents (based on Aijmer 1984).

NP	PP	VP	AdjP	AdvP	Total
73	0	5	3	0	81

In particular, Aijmer (1984:122-123) states that *sort of* tends to collocate with nouns that have "little semantic content" such as *person*, *way*, *place*, *shape*, *area*, *thing*, and *stuff* for nouns as well as "simple non-specific" verbs such as *leap*, *sit*, *look*, *mutter*, *feel*, and *try*. Also, according to Aijmer (1984:124), *sort of* preferably precedes words that are "technical, rare, foreign, formal, vulgar, idiomatic, etc."

In what follows, we would like to point out a few quibbles we have with some of the studies of *kind of* and *sort of*. Again, they come in several groups. One is concerned with the scope of the studies. For example, some previous studies focused on only one of the two hedges as opposed to comparing them. In addition, some - Yaguchi et al. (2004) and Denison's (2005) more than 1,200 examples being the obvious exceptions - were based on rather small corpora and databases. A related point is the resolution or granularity adopted in earlier works. Many if not most previous works investigated the hedges at only one level of granularity (e.g., speaking vs. writing or BrE vs. AmE) and chose few or even just one register for analysis (e.g., academic discourse or spoken conversation).

Another aspect we would like to address is the role played by quantitative analysis. Most previous works did not do by-subjects/by-item analyses (cf. Gries 2006), where by-subjects analysis refers to distinguishing specific speakers' preferences whereas by-item analysis refers to determining if and to what degree the two hedges exhibit lexical co-occurrence preferences to semantically definable groups of words. In addition, some studies were quantitatively less advanced than one would like them to be, not using normalized frequencies. For example, recall the data reported by Aijmer (1984) and their evaluation. From our point of view, the distribution repeated here in Table 3 (with expected frequencies in parentheses) does in fact license conclusions other than the ones proposed by Aijmer. A chi-square test that compares the frequencies of *kind of* and *sort of* with respect to the following constituents shows that the distribution is in fact statistically significant (χ^2=25.43; *df*=4; *p*=4.115e-05, Cramer's *V*=0.213). However, while NPs account for most instances of *sort of*, in comparison with *kind of*, *sort of* actually *dis*prefers NPs.

Table 3. The distribution of *sort of* and *kind of* before major constituents (based on Aijmer 1984).

Hedge	NP	PP	VP	AdjP	AdvP	Totals
sort of	302 (exp.: 322)	8	145 (exp.: 129)	19	8	482
kind of	73 (exp.: 53)	0	5 (exp.: 21)	3	0	81
Totals	375	8	150	22	8	563

The conclusion that *sort of* disprefers NPs may come as a surprise given that (i) *sort of* NP is twice as frequent as *sort of* VP and that (ii) *sort of* NP is about four times as frequent as *kind of*

NP. However, these are not the relevant standards of comparison because comparing only observed frequencies to each other fails to include the baseline frequencies of *kind of* and *sort of* on the one hand and all XPs on the other hand (i.e., the marginal totals in Table 3). Thus, the relevant comparison is the comparison between all cells' observed frequencies and their expected frequencies, which in turn are computed on the basis of exactly the baseline that the former comparison fails to include. This is also graphically represented in the plot in Figure 1. [2]

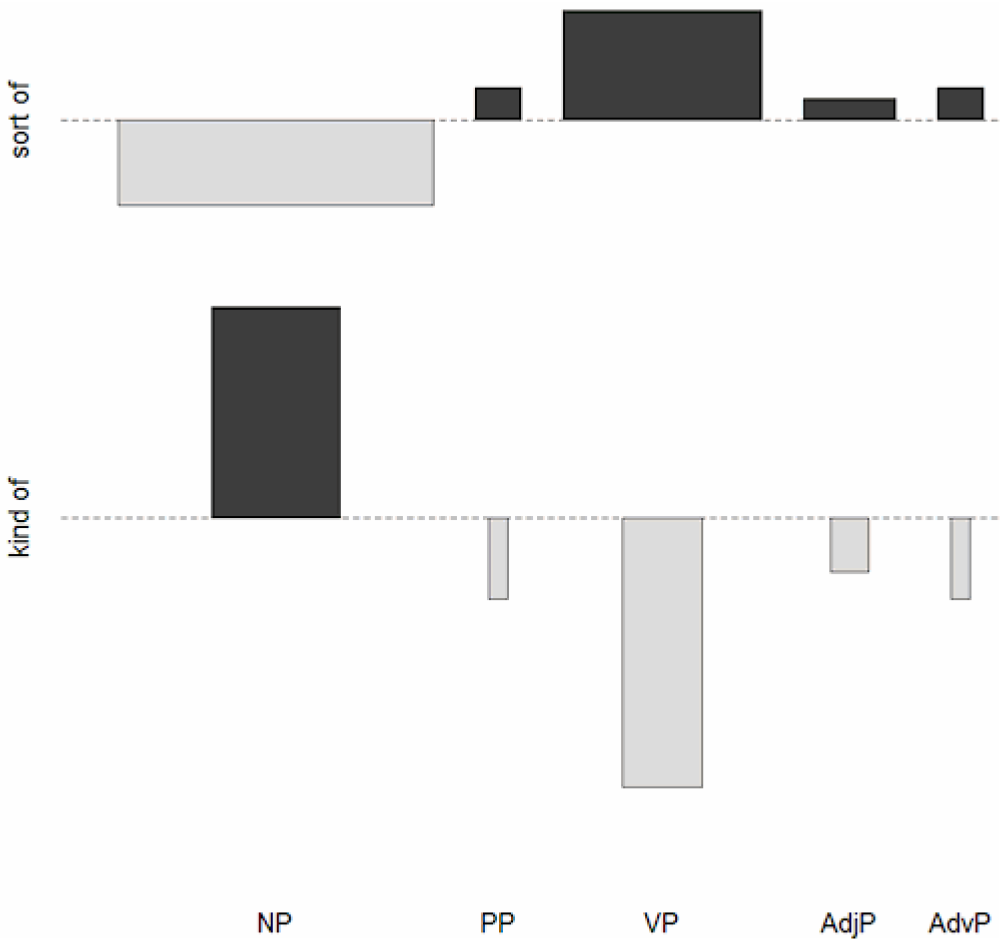

Figure 1. Association plot of *sort of* and *kind of* before major constituents (based on Aijmer 1984).

Finally, previous work sometimes provided descriptively problematic generalizations. For example, Aijmer's (1984) study of *sort of*'s collocates is potentially problematic in at least three respects. First, the proposed categories make up a seemingly unconstrained group, which comprises both "formal" and "vulgar", both "technical" or "foreign" and "idiomatic" etc. Second, these categories do not coincide together well with many collocates that are actually reported: collocates such as *person*, *way*, *place* etc. are none of the above. Lastly, we fail to see in which sense, say, the verbs *leap* and *mutter* are simple and, even more curiously, non-specific since these are certainly fairly specific motion and communication verbs respectively (as opposed to, e.g., *go* and *say*).

In the present study, we will present a quantitative corpus-linguistic analysis of *kind of* and *sort of*. We will be concerned with the factors that govern the choice of one hedge over the other in both spoken and written contemporary British English. The factors to be discussed below include (i) factors external to language (viz., the situationally/contextually defined register or text type of the utterance(s) in question) and (ii) factors internal to language (viz., some so far unnoticed preferences of *kind of* and *sort of* to be used together with particular lexical items and semantic fields). The next section will explain the methodology we adopted.

2. Methods

In this section, we will explain how our data were obtained and analyzed.

2.1 Retrieval and annotation

We retrieved all matches of the following search strings from the British National Corpus World edition:

- <w AV0>kind of;
- <w AV0>kinda;
- <w AV0>sort of. [3]

Three different kinds of annotation were performed. First, each of the matches was coded with respect to the mode, the medium (within writing), the domain type, and the genre type (using the categories from David Lee's BNC index (Lee 2001). Second, each instance was coded with respect to the part of speech of the expression modified by the hedge, which usually was the head of the immediately following XP); the inventory of categories was

- adjectives, as in *It was* kind of *expensive though* (BNC WE J1G:3765);
- nouns, as in *This message is* kind of *a test* (BNC WE J1K:210);
- verbs, as in *it just* sort of *drifts up to the roof* (BNC WE A74:2104);
- adverbs/prepositions, as in *Do people* sort of *artificially put on their best behavior* (BNC WE KRH:3673);
- whole propositions, as in *it was an opportunity to a new priest to come in and* sort of *if it had become <unclear>* (BNC WE HUT:67). [4]

Third, each instance was coded with respect to the lemma of the expression modified by the hedge (usually the head of the immediately following XP); for the above examples, these are *expensive*, *test*, *drift up*, *artificially*, and *if*. We omitted cases where the following elements immediately followed *kind of* or *sort of* were immediately followed by pauses, unclear cases, or punctuation marks. The resulting number cases that were largely manually annotated with respect to all three above parameters is 4,825, namely 570 instances of *kind of* and 4,255 of *sort of*.

2.2 Statistical evaluation

The frequency distributions of the above annotation parameters were then evaluated statistically. More specifically, the frequencies of occurrence of each parameter were converted into 2×2 tables. In each of these tables, the columns provide the co-occurrence frequencies with *kind of* and *sort of*. The rows, on the other hand, provide the frequencies of one level as opposed to the combined frequency of all other levels. Since most of the figures that will be reported below are such statistics, it is worth providing one example. Table 4, for instance, looks at the frequency distribution of *kind of* and *sort of* in W_arts as opposed to in all other domains.

Table 4. The distribution of *kind of* and *sort of* across different domains (in the BNC World edition).

	kind of	*sort of*	Totals
W_arts	127 (exp.: 24.8)	83 (exp.: 185.2)	210
all other domains	443	4,172	4,615
Totals	570	4,255	4,825

This kind of table, and all other ones derived analogously for domains and all other text types as well as collocate lemmas, was then evaluated with Coll.analysis 3 (Gries 2004). This is a program for R (for Windows) written by, and freely available from, the first author, which computes distinctive-collexeme-analysis co-occurrence statistics. Frequently used co-occurrence statistics are the chi-square test or *t*-tests or *z*-scores, but we used Coll.analysis 3's default setting, namely the Fisher-Yates exact test (cf. Gries and Stefanowitsch 2004). The reasons for using this procedure are that (i) a chi-square test on one complete table does not straightforwardly allow to determine where exactly the effect comes from, which is why level-wise comparisons of the above kind are useful here, (ii) we wanted to use the same test for all tables rather than using chi-square for some and some other test for others, and (iii) some of the expected frequencies in the tables to follow are very small and, thus, rule out a chi-square test.

The *p*-value of this statistical test is then transformed into a negative logarithm to the base of 10 such that values close to zero indicate a lack of preference of a domain to a hedge. On the other hand, high values indicate a strong preference of a particular domain to that hedge whose observed frequency in one domain is higher than its expected; a value larger than approximately 1.3 indicates a distribution significant at the 5% level. In this case, the log-transformed *p*-value, referred to as *collostruction strength* (or *CollStr* for short) is fairly high, 67.309, and as is indicated in Table 4, the domain W_arts is characterized by a strong preference of *kind of* as opposed to *sort of*. Analogous tests were performed for all annotated parameters and collocate lemmas (within each part-of-speech group), and in the following section we will discuss our results.

3. Results

3.1 Results for corpus parts

The first result pertains to the coarsest level of corpus granularity, namely the distinction between speaking and writing. The distribution we found is summarized in Table 5.

Table 5. The distribution of *kind of* and *sort of* in speaking and writing (in the BNC World edition).

	kind of	*sort of*	Totals
spoken	231 (exp.: 449.2)	3,571 (exp.: 3,352.9)	3,802

written	339 (exp.: 120.9)	684 (exp.: 902.2)	1,023
Totals	570	4,255	4,825

This distribution is highly significant, resulting in a CollStr value of $-\log_{10} p_{\text{Fisher-Yates exact test}}=102.007$. It is easy to see that the effect is that in writing, *kind of* is strongly preferred (in the sense of that *kind of* is about three times more frequent in writing than expected) whereas in speaking, *sort of* is strongly preferred. This finding matches previous results.

While this kind of table is easy to use in the case that the variable cross-tabulated with *kind of* and *sort of* only has two levels - such as here, where we distinguish only speaking vs. writing - this arrangement is less appropriate when more levels come into play. In what follows and in the appendix where we provide some additional results, the representational format is therefore slightly changed to accommodate multiple-level variables. With this format, the information provided in Table 5 would be represented as in Table 6 so that additional variable levels would simply be listed as additional rows together with their CollStr values.

Table 6. The distribution of *kind of* and *sort of* in speaking and writing (in the BNC World edition).

	$\text{obs}_{\textit{kind of}}$	$\text{exp}_{\textit{kind of}}$	$\text{obs}_{\textit{sort of}}$	$\text{exp}_{\textit{sort of}}$	preference	CollStr
spoken	231	449.2	3,571	3,352.9	*sort of*	102.01
written	339	120.9	684	902.2	*kind of*	102.01

It is obvious, however, that speaking vs. writing is only one possible - the coarsest - level of granularity so it is useful to explore whether more fine-grained register/text type distinctions reveal further patterns. The fine-grained resolution of Dave Lee's BNC indexer also allows, for example, to determine whether there are interesting patterns within writing. Table 7 summarizes the distribution of *kind of* and *sort of* in different media within writing.

Table 7. The distribution of *kind of* and *sort of* across different media (in the BNC World edition).

	$\text{obs}_{\textit{kind of}}$	$\text{exp}_{\textit{kind of}}$	$\text{obs}_{\textit{sort of}}$	$\text{exp}_{\textit{sort of}}$	preference	CollStr
periodical	137	82.51	112	166.49	*kind of*	15.86
m_unpubl	16	10.94	17	22.06	*kind of*	1.34
m_pub	2	0.99	1	2.01	*kind of*	0.59
to_be_spoken	1	1.99	5	4.01	*sort of*	0.45
book	183	242.57	549	489.43	*sort of*	17.22

As is obvious, the central three media have no or only a very slightly significant preferences for one hedge, but there are strong effects such that *kind of* is preferred in periodicals whereas *sort of* is preferred in books.

While Table 7 only considers writing, the even more fine-grained resolution in terms of BNC domains, allows us to test for more detailed patterns, which are shown in Table 8.

Table 8. The distribution of *kind of* and *sort of* across different domains (in the BNC World edition).

	obs$_{\textit{kind of}}$	exp$_{\textit{kind of}}$	obs$_{\textit{sort of}}$	exp$_{\textit{sort of}}$	preference	CollStr
W_arts	127	24.8	83	185.2	*kind of*	67.31
W_imaginative	147	65.2	405	486.8	*kind of*	24.17
W_leisure	31	9.9	53	74.1	*kind of*	8.76
W_app_science	9	2.1	9	15.9	*kind of*	4.12
W_soc_science	15	11.2	80	83.8	*kind of*	0.83
S_demog_uncl.	2	0.7	4	5.3	*kind of*	0.82
W_world_affairs	8	5.32	37	39.68	*kind of*	0.81
W_commerce	2	1.42	10	10.58	*kind of*	0.37
W_nat_science	0	0.12	1	0.88	*sort of*	0.06
NA	0	0.12	1	0.88	*sort of*	0.06
W_belief_thought	0	0.71	6	5.29	*sort of*	0.33
S_cg_education	98	103.49	778	772.51	*sort of*	0.55
S_demog_DE	4	11.1	90	82.9	*sort of*	2.01
S_cg_publ_inst	17	32.5	258	242.5	*sort of*	3
S_demog_C1	24	42.3	334	315.7	*sort of*	3.19
S_cg_leisure	40	69.6	549	519.4	*sort of*	4.89
S_demog_ab	25	66	534	493	*sort of*	9.72
S_cg_business	18	63.4	519	473.6	*sort of*	12.68
S_demog_c2	3	59.9	504	447.1	*sort of*	24.46

While it is obvious that there are a variety of very strong preferences of domains to hedges, the most interesting finding from our point of view is that it seems as if this resolution is in fact more fine-grained than is needed. Granted, writing on arts subjects or imaginative writings strongly prefer *kind of* etc., but the strongest pattern is that

- all the written domains prefer *kind of* (usually significantly) and
- all but one spoken domains prefer *sort of* (usually significantly).

However, since this is a finding we already obtained from the much more coarse-grained perspectives of Table 5 and Table 6, this part of the analysis strictly speaking does not seem to provide much new information. This finding is very much in line with the arguments put forward in Gries (2007, to appear), where it is argued that

> [f]irst, there are several levels of hierarchical organization or granularity at which variability might be located: modes, registers, sub-register (see below) or even lexically-defined levels. Second, even within one level of hierarchical granularity there are usually more than two levels between which differences may exist. Thus, for instance, even if differences are located at the level of the register, this need not mean that all registers are (equally) different from each other
>
> (Gries 2007:110)

The simultaneous comparative analysis of several levels of corpus granularity here showed that in spite of the fact that every level of granularity will yield some significant results, the most striking pattern is usually only to be found at one of them, and for the present case we submit that the distinction that accounts for the lion's share of the data most parsimoniously is the coarse-grained one of speaking vs. writing.

In addition to the above data, we also tested each genre's preference of *kind of* and *sort of*. Cf. Table (i) in the appendix for the results; suffice it here to say that again

- all the written domains prefer *kind of* (usually significantly) and
- all but one spoken domains prefer *sort of* (usually significantly).

Thus, tests on all levels of granularity provided by the BNC indexer indicate that most of the variation one obtains is actually exhibited between speech and writing.

3.2 Results for collocates

In this section, we will increase the resolution even more but leave the domain of situationally-defined text types and turn to lexical co-occurrence patterns. In Section 3.2.1, we briefly investigate which parts of speech *kind of* and *sort of* preferably modify, which also gives us an opportunity to determine whether our above methodological criticism of Aijmer's data is in fact warranted. Section 3.2.2 will then look at the hedges' preferences for both specific lexical preferences and semantic fields.

3.2.1 Collocates in terms of parts of speech

The analysis of the hedges' parts of speech preferences is based on the annotation of each instance was coded with respect to the part of speech of the expression modified by the hedge, usually the head of the immediately following XP). A distinctive collexeme analysis with Coll.analysis 3 of the same type as the above yields the results in Table 9, which are again also graphically summarized in Figure 2.

Table 9. The distribution of *kind of* and *sort of* across different parts of speech (in the BNC World edition).

	$\text{obs}_{\textit{kind of}}$	$\text{exp}_{\textit{kind of}}$	$\text{obs}_{\textit{sort of}}$	$\text{exp}_{\textit{sort of}}$	preference	CollStr
N	110	67.1	458	500.9	*kind of*	7.76
Adj	113	73.6	510	549.4	*kind of*	6.35
Prop	10	16.66	131	124.34	*sort of*	1.35
Adv	39	60.48	473	451.52	*sort of*	3.14
V	285	340.82	2,600	2,544.18	*sort of*	6.52

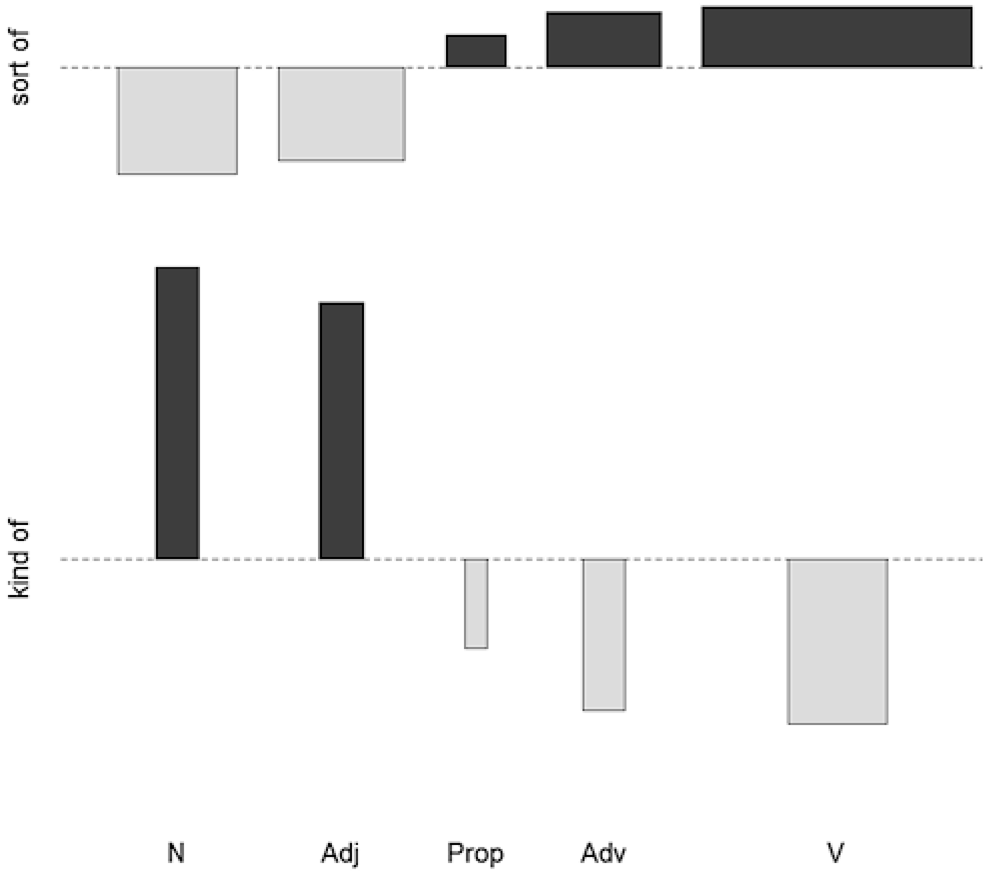

Figure 2. Association plot of *sort of* and *kind of* before major constituents (the present data set).

As is obvious, *kind of* strongly prefers to modify nouns and adjectives while *sort of* strongly prefers to modify whole propositions, adverbs and verbs. It is interesting to note in passing that this division perfectly fits the distinction of stative relations - nouns and adjectives - versus dynamic relations involving verbal predication - propositions, adverbs and verbs. Also, note that the present result for NPs now matches our corrected evaluation of Aijmer's (1984) data, which lends further support to our above critique of the initial interpretation of that data set.

3.2.2 Collocates in terms of lemmas and semantic fields

The detailed annotation of the lemmas of the heads of the XPs modified by *kind of* and *sort of* allows us to not only observe part-of-speech specific effects, but also lexically specific effects. We restrict our attention to the three most frequent parts of speech, adjectives, nouns, and verbs. Within each part of speech, we did a distinctive collexeme analysis to determine each adjective's/noun's/verb's co-occurrence preference of *kind of* or *sort of*. Given the large number of different lemmas - the table for the adjectives alone has more than 450 rows - we can not provide all the results here and discuss the strongest preferences only summarily. [5], [6]

Within the class of adjectives, some clear patterns were obtained. For example, color adjectives exhibit a strong preference to be modified by *sort of*: the only color adjective significantly attracted by *kind of* is *bunched-white*, but *sort of* attracts many different color terms including *black*, *dark*, *grey*, *blue*, *burnt-orange*, *dark-green*. On the other hand, *kind of* has a strong preference for adjectives having something to do with what can be broadly classed as emotional states: *fun*, *cool*, *calm*, *eerie*, *funny*, *bluesy*, *depressing*, *dramatic*, *emotional*, *helpless*. By contrast, *sort of* has few of these as (significantly) attracted adjectives: *happy*, *enthusiastic*. Then, *kind of* and *sort of* both take a lot of inherently 'positive' adjectives:

- *kind of*: *fun*, *cool*, *calm*, *cute*, *funny*, *authentic*, *exceptional*, *famous*, *graceful*, *exciting*;
- *sort of*: *happy*, *personal*, *good*, *comforting*, *enthusiastic*, *forceful*, *friendly*, *holy*, *humane.*

Interestingly, though, *kind of* modifies even more 'negative' adjectives while *sort of* does not:

- *kind of*: *eerie, depressing, disoriented, dramatic, expensive, fishy, helpless, weird, awkward, boring, childlike, doubtful, scary*;
- *sort of*: *dark*, *difficult*, *clumsy*, *formal*, *fuzzy*, *odd*, *rough.*

Finally let us mention as an aside that the adjectives modified by *kind of* are on average significantly shorter than those modified by *sort of* (W=13570; $p_{2\text{-}tailed}$=0.0454). This is peculiar because we would have expected the reverse effect given that it is *sort of* which is preferred in speaking, where we would expect shorter and more frequent words. Be that as it may, the effect is small (mean difference: 0.7 letters) and we can so far not relate it to anything more substantial.

Within the class of nouns, the only major finding is that *sort of* has an extremely large number of cases in which it modifies quantified expressions, i.e., expressions involving numbers:

- numbers: NUM, *one*, NUM *perc*, NUM *month*(*s*), NUM *week*(*s*), *number*, NUM *day*(*s*), NUM *people*, NUM *year*;
- time(s): *time*, NUM *month*(*s*), NUM *week*(*s*), NUM *day*(*s*), NUM *year*(*s*).

In addition, there is a tendency for *sort of* to modify nouns that refer to people and their inalienable parts: *people*, *hair*, *boy*, *leg* - *kind of* takes no such nouns in our sample. [7]

Finally, within the class of verbs, there are also a few noteworthy tendencies. The first of these is that *kind of* does not significantly modify communication verbs at all while *sort of* does (e.g., *say*, *talk*, *ask*, *call*, *explain*). Secondly, *kind of* does not significantly modify perception verbs, whereas *sort of* has several perception and causation-of-perception verbs (e.g., *see*, *look*, *find*, *show*, *watch*). Interestingly, with mental activity verbs, it is the other way round: *kind of* takes *know*, *wish*, *want*, *like*, *hope*, *admire*, *anticipate*, and *arouse*, whereas *sort of* only takes *think* and *learn*. Finally, as to motion verbs, both *kind of* and *sort of* take many different motion verbs, but there is no particularly clear pattern to discern:

- *kind of*: manner-of-motion verbs and/or end-point verbs such as *jump*, *get off*, *stroll*, *dance*, *descend*, *break in*, *break up*, *bring out*;
- *sort of*: more basic (causation of) motion verbs such as *go*, *sit*, *walk*, *push*, *put*, *move*, *stay*, *pull*, *come*.

It becomes relatively obvious that apart from text-type specific patterns, there is also clear evidence for lexically determined variation. This variation is found both on the level of the part of speech which is modified by a hedge but also directly on the level of the lemma itself and the semantic and evaluative domains to which lemmas belong. The patterns noted here have not been observed in previous work so far, but they often strongly exceed standard levels of significance and, thus, have to be part of a comprehensive characterization of the hedges' variation in contextually and socially different utterance situations.

3.3 A question of synonymy

So far the scope of this paper has been (i) descriptive in the sense of providing a comprehensive characterization of when one hedge is preferred over the other and (ii) methodological in the sense of exemplifying the multiple-levels-of-granularity issue discussed at length in Gries (2007, to appear). However, the results are also interesting from a different methodological angle, namely when they are compared to the findings of other related approaches to near synonymy. For example, previous work has shown that intuitions or lexicographic analysis concerning (degrees of) synonymy correlate strongly with corpus data reflecting collocational overlap or in fact even benefit in terms of explanatory power; cf. Gries (2001, 2003) for studies investigating adjective pairs ending in *-ic* and *-ical*. Gries (2001, 2003) finds that (i) the significantly distinctive nominal R1-collocates of these adjectives allow for a sometimes surprisingly precise characterization of their semantic differences and that (ii) the degree of significant collocate overlap is a good indicator of the degree of semantic similarity.

It is intuitively relatively obvious that *kind of* and *sort of* are also very close in meaning, are in fact near synonyms. It would therefore only be natural to expect that this would also be reflected in a strong degree of collocational overlap. However, the results show that this is not uniformly the case: *kind of* and *sort of* exhibit relatively small collocational overlap of adjectives, but score high with nouns and verbs; cf. Table 10: [8]

Table 10. Overall and significant collocate overlap of *kind of* and *sort of* for adjectives, nouns, and verbs.

	Adj	N	V
collocate overlap	17.5% (452 types)	10.5% (378 types)	31.4% (1,133 types)
significant collocate overlap	10.7% (28 types)	50% (5 types)	69.2% (13 types)

While the high percentages for nouns and verbs are certainly in part due to the small number of types reaching the standard level of significance of 5%, there is a marked difference between the adjectives on the one hand and the nouns and verbs on the other hand. In Gries (2003), percentages of the (small) size obtained here for the adjectives usually correlated with marked and fairly obvious semantic differences whereas the uncontroversially good cases of nearly synonymous adjectives regularly scored percentages of 40%, 50%, and higher, i.e., the values we found for nouns and verbs. At present, we are unsure what to make of this difference ... does it point to a problem in the method of collocational overlap? Or, does it indicate that the performance of collocational overlap as a diagnostic is contingent on the kind of node word and/or the part of speech of the collocate? Is the method of collocational overlap so much interrelated with text-type distributional findings that it may have to be adjusted? While we cannot address all these questions in the present paper, this shows that further work on this issue is certainly called for.

4. Discussion

We argued that previous studies leave open a variety of questions in terms of the distribution of the two hedges *kind of* and *sort of*. We hope to have shown that, first, the re-analysis of previous data at times reveals 'unnoticed' trends regarding the part-of-speech specific distribution of the hedges, which is then even supported by the present data set. Second, the application of the distinctive collexeme analysis method revealed quite a variety of so far unnoticed usage preferences of the hedges in terms of register / text type distribution. In this connection, the data discussed here also strongly underscore the necessity to perform analyses of the same set of corpus data on multiple levels of corpus organization. Without such comparisons, a corpus linguist may arbitrarily pick any level of hierarchical organization and just report these findings without ever having tested whether more fine- or more coarse-grained perspectives provide more revealing results; cf. Gries (2007, to appear for a variety of case studies). Third, methodological refinements in terms of, say, the corpus-linguistic analogon to by-items analyses (cf. Gries 2006) reveal strong semantic preferences both on the level of the individual lemma and that of semantic fields that go beyond the usually discourse-pragmatic studies concerning these hedges.

There are a variety of steps that could be undertaken next. On the one hand, the most obvious continuation would be to combine the two kinds of parameters we have investigated separately, thus effectively doing a multidimensional analysis to analyze to what degree lexical preferences are register/...-dependent. Hopefully, this would not only provide a more complete descriptive characterization of the use of the two hedges, but also allow for a more sophisticated explanation - and ultimately perhaps even prediction - of which hedge will be used in which communicative situation, which we would consider to be the best indicator of the reliability of the analysis.

In addition, an analysis of sociolinguistic determinants of hedge choice as well as speaker-specific preferences could reveal interesting distributions that shed further light on patterns of variation in the contexts of social interaction. Also, we have begun to explore whether the identification of the semantic fields that are probabilistically associated with the hedges can be made more objective and/or maybe even more comprehensive by using WordNet, but the results are as yet inconclusive. However, it should have become clear that more rigorous statistical treatment and by-item analyses along the lines of Gries (2006) as well as more fine-grained text type analysis still have a lot to offer to the corpus-linguistic tool box, and we hope that the recognition of this fact will result in a wealth of more precise and more comprehensive corpus-linguistic findings.

Notes

[1] It is not clear to us what to make of the difference between the 585 occurrences Aijmer (1984:118) claims to have found in the corpus and the 482 discussed in her (and our) Table 1. The only possibility we can think of is that the remaining 103 instances did not occur before "major constituents" and therefore did not make it into Table 1.

[2] Figure 1 is a so-called association plot, the from our point of view best way to summarize two-dimensional frequency tables. The (dark) rectangles above the dotted lines represent co-occurrence frequencies that are larger than expected; the (lighter) rectangles below the dotted lines represent co-occurrence frequencies that are smaller than expected. The rectangle's height represents the cell's contribution to Pearson's χ^2, the width represents the square root of the frequency expected by chance so that "the area of the box is proportional to the difference in observed and expected frequencies" (cf. R help, s.v. *assocplot*).

[3] We also retrieved all matches for "<w NN1>kind <w PRF>of" and "<w NN2>kinds <w PRF>of" as well as "<w NN1>sort <w PRF>of" and "<w NN2>sorts <w PRF>of". However, we will not be concerned with these here because these are part of regular NPs rather than hedges as in, for example, the sentence *This is the kind of expression we did not include in our analysis*; the distinction is not always easy to make, however (cf. Manning and Schütze 2000:13-14).

[4] From these examples, the difficulty to categorize accurately already emerges clearly. Without hearing the utterance, it is not fully clear whether there is in fact a break around *sort of* that indicates that *sort of* is just a disfluency marker. The coding was undertaken such that it stuck to the material as closely as possible and since there was no "<pause>" annotation here, the sentence was coded as indicated above. However, many difficult cases remained and we are far from certain to always have made the optimal decision.

[5] The results tables are available upon request from the primary author.

[6] The interpretations of the patterns are by necessity somewhat subjective. The classifications were first made by the primary author, then checked for consistency with the second author. While we are confident that most readers will agree with the classification we arrived at, we hope that at some later point of time there will be an opportunity to arrive at an even more objective way of classification (cf. also Section 4).

[7] As a matter of fact, there is one other small finding to be mentioned, which is the large number of tagging errors that occurred with expressions such as *This is very kind of you*. The vast majority of these are tagged incorrectly as *<w AV0>kind of <w PNP>you* rather than *<w AJ0>kind <w PRF>of <w PNP>you*. We thank David Denison for pointing out this possibility to us at the ICAME conference where this paper was first presented.

[8] The results for nouns are identical if the coding errors for the collocate *you* are discarded.

Sources

The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk

Gries, Stefan Th. 2004. *Coll.analysis 3. A program for R for Windows*. http://www.linguistics.ucsb.edu/faculty/stgries/

Lee, David. 2001. The BNC index. Available from http://clix.to/davidlee00; last access December 2003.

R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org

References

Aijmer, Karin. 1984. "*Sort of* and *kind of* in English conversation". *Studia Linguistica* 38: 118-128.

Aijmer, Karin. 1986. "Discourse variation and hedging". In *Corpus linguistics II: [...]*, Jan Aarts and Willem Meijs (eds.), 1-18. Amsterdam: Rodopi.

Biber, Douglas, Stig Johansson, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Crystal, David & Derek Davy. 1975. *Advanced Conversational English*. London: Longman.

Church, Kenneth W., William Gale, Patrick Hanks, Donald Hindle & Rosamund Moon. 1994. "Lexical substitutability". *Computational Approaches to the Lexicon*, ed. by Beryl T. Sue Atkins & Antonio Zampolli, 153-177. Oxford: O.U.P.

Denison, David. 2005. "The grammaticalisations of *sort of*, *kind of* and *type of* in English". Paper presented at New Reflections on Grammaticalization (NRG) 3, University of Santiago de Compostela.

Gries, Stefan Th. 2001. "A corpus-linguistic analysis of -*ic* and -*ical* adjectives". *ICAME Journal* 25: 65-108.

Gries, Stefan Th. 2003. "Testing the sub-test: A collocational-overlap analysis of English -*ic* and -*ical* adjectives". *International Journal of Corpus Linguistics* 8(1): 31-61.

Gries, Stefan Th. 2006. "Some proposals towards more rigorous corpus linguistics". *Zeitschrift für Anglistik und Amerikanistik* 54(2): 191-202.

Gries, Stefan Th. 2007. "Exploring variability within and between corpora: some methodological considerations". *Corpora* 1(2): 109-52.

Gries, Stefan Th. to appear. "Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions?" *Expanding Cognitive Linguistic Horizons*, ed. by Mario Brda & Milena Žic Fuchs. Amsterdam/Philadelphia: John Benjamins.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. "Extending collostructional analysis: A corpus-based perspectives on 'alternations'". *International Journal of Corpus Linguistics* 9(1): 97-129.

Kay, Paul. 1997. "The *kind of/sort of* constructions". *Words and the Grammar of Context*, ed. by Paul Kay, 145-58. Stanford, Calif.: CSLI Publications.

Lakoff, George. 1972. "Hedges: a study in meaning criteria and the logic of fuzzy concepts". *Papers from the 8th Regional Meeting*, ed. by Paul M. Peranteau, Judith N. Levi & Gloria C. Phares, 183-228. Chicago, Ill.: Chicago Linguistics Society.

Lakoff, Robin. 1975. *Language and women's place*. New York: Harper and Row.

Manning, Christopher D. & Hinrich Schütze. 2000. *Foundations of statistical natural language processing*. 2nd printing with corrections. Cambridge, Mass. and London: The MIT Press.

Markkanen, Raija & Hartmut Schröder. 2000. "Hedging: A challenge for pragmatics and discourse analysis". http://www.sw2.euv-frankfurt-o.de/Publikationen/Hedging/markkane/markkane.html

Nunberg, Geoffrey. 2004. "It's sort of like a, you know, verbal Rorschach Test". http://people.ischool.berkeley.edu/~nunberg/sortof.html

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Schröder, Hartmut & Dagmar Zimmer. 2000. "Hedging research in pragmatics: a bibliographical research guide to hedging". http://www.sw2.euv-frankfurt-o.de/Publikationen/Hedging/zimmer/zimmer.html

Yaguchi, Michiko, Yoko Iyeiri & Hiroko Okabe. 2004. "Style and gender differences in formal contexts: An analysis of *sort of* and *kind of* appearing in the CSPAE". Paper presented at the 5th Annual Wenshan Conference on ELT, Literature, and Linguistics.

Appendix

Table (i). The distribution of *kind of* and *sort of* across different genres (in the BNC World edition).

words	obs$_{\textit{kind of}}$	exp$_{\textit{kind of}}$	obs$_{\textit{sort of}}$	exp$_{\textit{sort of}}$	preference	CollStr
W_pop_lore	115	21.74	69	162.26	*kind of*	62.84
W_fict_prose	141	64.03	401	477.97	*kind of*	21.97
S_lect_soc_science	68	25.16	145	187.84	*kind of*	14.99
W_biography	18	6.14	34	45.86	*kind of*	4.87
W_email	10	2.48	11	18.52	*kind of*	4.29
W_misc	9	2.13	9	15.87	*kind of*	4.12
W_non_ac_tech_engin	5	1.06	4	7.94	*kind of*	2.72
W_non_ac_humanities_arts	4	0.95	4	7.05	*kind of*	2.04
W_newsp_brdsht_nat_arts	4	1.06	5	7.94	*kind of*	1.83
W_ac_polit_law_edu	4	1.3	7	9.7	*kind of*	1.49
W_newsp_tabloid	2	0.35	1	2.65	*kind of*	1.41
W_newsp_other_report	3	1.18	7	8.82	*kind of*	0.98
W_newsp_other_sports	3	1.18	7	8.82	*kind of*	0.98
W_newsp_other_arts	1	0.12	0	0.88	*kind of*	0.93
W_non_ac_soc_science	7	4.37	30	32.63	*kind of*	0.86
W_ac_humanities_arts	2	0.83	5	6.17	*kind of*	0.71
S_brdcast_news	5	3.19	22	23.81	*kind of*	0.68
W_newsp_brdsht_nat_report	1	0.24	1	1.76	*kind of*	0.65
W_non_ac_medicine	1	0.24	1	1.76	*kind of*	0.65
S_lect_humanities_arts	3	1.65	11	12.35	*kind of*	0.65
W_essay_school	1	0.35	2	2.65	*kind of*	0.5
W_fict_poetry	1	0.47	3	3.53	*kind of*	0.4
W_non_ac_nat_science	1	0.47	3	3.53	*kind of*	0.4
S_brdcast_documentary	1	0.59	4	4.41	*kind of*	0.33
W_newsp_other_social	1	0.59	4	4.41	*kind of*	0.33
W_news_script	1	0.71	5	5.29	*kind of*	0.28
S_conv	58	180.04	1466	1343.96	*sort of*	36.61
S_unclassified	2	18.78	157	140.22	*sort of*	6.41
S_meeting	24	50.92	407	380.08	*sort of*	5.44

S_speech_unscripted	7	24.69	202	184.31	*sort of*	5.05
S_classroom	15	36.15	291	269.85	*sort of*	4.79
S_tutorial	4	17.01	140	126.99	*sort of*	4.1
S_brdcast_discussn	7	21.85	178	163.15	*sort of*	4.05
S_interview_oral_history	30	45.01	351	335.99	*sort of*	2.21
S_consult	5	10.4	83	77.6	*sort of*	1.37
S_pub_debate	0	2.48	21	18.52	*sort of*	1.15
S_interview	2	4.73	38	35.27	*sort of*	0.88
W_ac_soc_science	2	4.73	38	35.27	*sort of*	0.88
S_speech_scripted	0	1.54	13	11.46	*sort of*	0.71
S_demonstratn	0	1.3	11	9.7	*sort of*	0.6
S_sportslive	0	1.06	9	7.94	*sort of*	0.49
S_lect_commerce	0	0.59	5	4.41	*sort of*	0.27
S_lect_polit_law_edu	0	0.59	5	4.41	*sort of*	0.27
W_commerce	0	0.59	5	4.41	*sort of*	0.27
S_lect_nat_science	0	0.47	4	3.53	*sort of*	0.22
S_parliament	0	0.47	4	3.53	*sort of*	0.22
W_non_ac_polit_law_edu	2	2.24	17	16.76	*sort of*	0.22
W_newsp_brdsht_nat_misc	0	0.35	3	2.65	*sort of*	0.16
S_courtroom	0	0.24	2	1.76	*sort of*	0.11
W_ac_nat_science	0	0.24	2	1.76	*sort of*	0.11
W_instructional	0	0.24	2	1.76	*sort of*	0.11
W_religion	0	0.24	2	1.76	*sort of*	0.11
NA	0	0.12	1	0.88	*sort of*	0.06
S_sermon	0	0.12	1	0.88	*sort of*	0.06
W_advert	0	0.12	1	0.88	*sort of*	0.06
W_newsp_brdsht_nat_science	0	0.12	1	0.88	*sort of*	0.06

Studies in Variation, Contacts and Change in English 2: Towards Multimedia in Corpus Studies
Article © 2007 Stefan Th. Gries & Caroline V. David; series © 2007 VARIENG
Last updated 2007-12-17 by Tanja Säily
Links updated 2008-12-11 by Tanja Säily