Frank Knowles / Peter Roe
Institute for the Study of Language and Society
Aston University, Birmingham

LSP and the Notion of Distribution as a Basis for Lexicography

Abstract

This research capitalises on the availability of very large-scale text databases and software designed to focus on LSP collocability within distinct discourse communities. A detailed description is given of work involving the statistically-driven detection, identification and extraction of meaningful 'chunks' in the form of multi-word units which unmask themselves as the prominent 'professional' collocations. Novel lexicographical-terminological treatment of this material leads to an analysis and display of the dynamics of LSP term clusters and their associated 'mind maps'. Most of all, the paper demonstrates that the approach used indicates the possibility of shifting the ground of lexicography away from its traditional notional centre of gravity to an AI-oriented strategy for capturing genuine cognitive units without robbing them of their textuality. This leads naturally to a powerful, qua direct, type of lexicographical codification, the primum mobile of which is not definition, but distribution.

1. Introduction

This paper reports on implications for lexicography of the results of research carried out on large LSP (i.e. Language for Specific Purposes) text corpora in English, French and German. In this particular case text selections were made (with the copyright holder's full permission) from the CD-ROM version of the *Financial Times* (FT).

In the context of the Aston LSP Research Sector, LSP is defined as the provision of answers to the following two questions posed by a language learner seeking to enter an expert discourse community (a term discussed more fully in the next section) for which s/he is technically but not linguistically qualified:

- What do I need to 'know' that I do not already 'know', in order to be able to operate effectively in the context of the target discourse community which I wish to join?
- How may I, most efficaciously, effectively and efficiently, overcome my (linguistic) deficiency?

The particular focus taken in this paper is to seek the most appropriate contribution of the lexicographer to the answer to these questions, given

contemporary insights into the nature of language and the current advances in our ability to explore and interrogate large corpora of authentic text. Such LSP investigations (confined, in our case, to written text at this stage) can be classed on a focal cline ranging from the extremely narrow (e.g. Seaspeak[1]) to the extremely broad (e.g. the English novel). On this cline the language of the various domains of the FT ranges from narrow to mid–focus. For those who pursue this line of research, meaning and value are seen as properties of a discourse community negotiating its proper outcomes (Swales[2]; Johns[3]). Thus these meanings/values seek vehicles for their expression and find 'best fits' in shared linguistic experience of other discourse communities. This view contrasts sharply with meaning seen as a pre–determined property of lexis. One conventional approach is the 'fractionating' of lexical items into their world of possible 'meanings' (Bolinger[4]). The less conventional stance adopted here is to conceive of 'meanings' as peculiar to the context in which they arise. Such 'meanings' then innovatively attach themselves to convenient extant lexical items or they realise themselves in neologisms. Thus, whereas the world of the lexicon is relatively stable, the world of meanings is inherently unstable. The mapping of new meanings onto old lexis is inexact and has a considerable aleatory dimension.

The approach taken by this research is not to tread the path of notional definition and categorisation but rather to determine what (relatively arbitrary) decisions a particular discourse community makes and how it 'valorises' the concrete choices which flow from such decisions. All we can be sure of are the lexical selections actually made and their distribution. The rest — that is, the traditional work of lexicographers — must contain an element of guesswork, admittedly often seemingly inspired guesswork as they seek to establish equivalences which at best can only be partial and at worst highly misleading and even dangerous. However, it is only the advent of available and affordable technology, coupled with copious sources of rich data, that enables us to make statistically significant statements of a more objective nature.

Software routines especially constructed for this type of macro–corpus analysis, but not relying on any type of lemmatisation or tagging, are able to identify and extract — with considerable rapidity and flexibility because of an ability to hold millions of running text words in a dynamic index — the essential phraseology and terminology of this global discourse domain, potentially cascading it into appropriate bins representing authentic sub–domains. The actual meaningful 'chunks'[5] identified in this way are clustered and hence made amenable to lexicographical–cum–terminographical treatment. These chunks — either single or multiple orthographic words, flush or discontinuous — may label fully–fledged 'professional' cognitive units, everyday cognitive units, fixed, stable or variable collocations of either a universal or domain–specific nature. Frequency tabulation provides a strong initial basis for classifying them one way or another.

A natural consequence of the ability to identify the textual labels corresponding to cognitive units constituting the 'business' of a community is the emergence of the possibility and desirability of quasi–lexicographic definition, based on an AI–oriented strategy, which does not rely on any notional basis but proceeds operationally — just as text itself does — to invoke or excite appropriate cognitive structures in readers' minds. This is a type of definition by distribution, notably by co–locational and collocational dynamics. It produces dynamic definitions which are not robbed of their textuality but which are progressively modulated by their environment. Normally, lexicographers provide only 'final', not developing, definitions. Above all, the approach discussed here deals with genuine cognitive units, not with abstractions or fusions of them. In this way a vital bridge is created between static, extra–textual dictionary (sub)–senses and dynamic textual meaning(s). It follows from the above that considerable scope also exists for the automatic disambiguation of homographs and of sub–senses within a lexeme.

What is of particular interest to this Congress is the claim that the approach used in the research described indicates the possibility of shifting the ground of lexicography away from its traditional notional centre of gravity to an AI–oriented strategy for capturing genuine cognitive units without robbing them of their textuality. This leads naturally to a powerful, qua direct, type of lexicographical codification, the primum mobile of which is not definition, but distribution. The next section now takes a closer look at two theoretical notions underlying this position.

2. Theoretical foundations

The starting point for the first of these is that of a writer faced with the task of putting thoughts into words as a communicative act in the context of a highly specialised Discourse Community. During the years of the First World War Wittgenstein made a significant contribution to contemporary philosophy in a series of observations, or "remarks", which came to be known as the *Tractatus Logico–Philosophicus*.[6] He thus faced the universal task of the writer of mapping the conceptual framework of those thoughts onto a chosen subset of elements in the lexical set of a language (in this case German), with all the morphosyntactic possibilities and constraints entailed by such choices. The verbal realisation of the elements of the conceptual framework themselves was clearly constrained by the number of convenient pre–existing lexical entities. The lengths to which Wittgenstein had to go to negotiate the semantic value of the items selected is ample evidence of the imperfect nature of the match between the value of the philosophical construct and the pre–existing meaning potential of the lexeme chosen to carry it. The intrinsic difficulty involved in such linguistic choices is apparent from various of his remarks such as: "Der Sachverhalt ist eine Verbindung von Gegenständen (Sachen, Dingen)." This raises a number of interesting

questions , e.g.

- Do the defining remarks applied explicitly to "Ding" apply equally well implicitly to "Sache" and "Gegenstand"?
- Is the "SACH–" in "Sachen" intended to be congruent with the "SACH–" in "SACHverhalt" and "TatSACHe" (and later "SACH–lage")? We here adopt the convention of representing the formal concept (the exprimendum), as defined, by the notation "SACH".

Now although Wittgenstein himself goes into some detail concerning linguistic ambiguities and tautology in "everyday language" ("Umgangssprache"), this paper is not intended as a contribution to his philosophy. It takes him rather as a particularly interesting example of a writer who was acutely aware of the limitations of language, but was nonetheless ultimately bound by the walls of the prison–house of language. It is this prison–house that this paper will continue to explore. The selection of Bild to represent BILD brought with it a host of options and contraints peculiar to the German language and not mirrored in any other language. Attention is drawn to the following:

- Concepts ('exprimenda') seeking a lexeme to give them flesh may have to select from a number of contending candidates.
- The word chosen to be the vehicle for WORD will come with a (from the writer's point of view, arbitrary) set of pre–established morphological combinations (types and compounds) in which the conventional senses may not all overlap with WORD. (NB This is quite apart from the many senses for "word" listed under a particular dictionary entry.)
- Thus, where WORD is realised by "word", "–word–" fused with a common tied morpheme may no longer carry the value of WORD, unless this is first explicitly negotiated with interlocutors.
- Thus the adjective "word" (bearing the meaning WORD), plus "–ly", may yield an adverb unrelated to WORD.

Put another way, individual members of the set of permitted variants of "do" may each contain one or more "DO"s not found in any of the other members.

The second theoretical notion, to which reference has already been made above, is that of "Discourse Community" discussed at length in Swales (1990).[7] For him a Discourse Community is a body of people united by a common purpose, pursuing its business through its established mechanisms, generating thereby a discourse proper to its nature, constitutions, membership, values and intended outcomes, which will manifest itself in a variety of linguistic genres. Now in any mature, dynamic community these parameters will change with time and progress, and such changes will be

reflected in the discourse. The *Tractatus* is an exponent of one such genre, and a living witness to attempts to adapt language to the changing views and values of the Discourse Community. It is a chain of struggles to obtain a best fit and to negotiate the new WORD/word value–bondings demanded by those changes. It is the contention of this paper that the major WORD/word realignments are enacted, always at least imperceptibly, often very visibly, in every Discourse Community, including those whose members only communicate across distances. Furthermore, it is contended that these adjustments to the discourse tend to operate, initially at least, possibly permanently, at the level of the type, as well as at the level of the 'chunk'. Thus a type containing "–word–" may be found to realise "WORD", whereas the headword "Word" does not. Types (and 'chunks') thus lead a life of semantic independence.

By way of illustrative example, John Sinclair has pointed out[8] to us that the COBUILD corpus exhibits a strong tendency for the word "torrent" to relate to noise, commonly verbal, commonly abusive, whereas "torrential" relates almost exclusively to water. His suggestion that the modern reader might, on encountering "a torrent of water" view it as a creative metaphor derived from "a torrent of abuse", no longer sounds so far–fetched.

3. Conclusion

When selecting a word to serve as a vehicle for WORD, the Discourse Community member does not seek a match by reference to the semantic range of a headword, but independently in the fields of the particular morphological variants. Semantic development takes place at the level of type, and may or not spread to the lemma represented — almost in splendid isolation — by the headword. It is further considered that WORD–values are in constant flux, that WORD–word bondings are slow to gel, and that nonce–bondings, illucidated and disambiguated by context and explicit definition sufficiently for the purposes of the business of the Discourse Community, are omnipresent. This is why dictionaries are inevitably out of date as soon as they are published. New lexicographic and publishing solutions are needed to overcome this problem, as well as the problem of the type freeing itself from the lemma's umbilical cord.

4. Illustrative and supporting data

Concrete evidence to support and illustrate the above theoretical position was sought in a well–defined corpus of written English, namely the *Financial Times* for the year 1992. This is considered an adequate compromise between a general English corpus such as COBUILD, which would call for extremely large samples of language, and very narrow corpora such as a professional journal, which might throw up highly idiosyncratic data. Selections were made from this corpus of 25 million words, as follows:

- one full day from each month of the year, taking each day Monday to Saturday twice, total 996,000 tokens (Corpus A);
- the full annual coverage of a number of specific financial sectors (e.g. the bonds market, stock market reports etc., totalling over six million tokens (Corpus B).

Searches were made using two different platforms, UNIX–based routines developed by us in Aston and operating on a Sun workstation (details are available in Roe (1994)[9]) and ATA, the Aston Text Analyser currently being developed with our industrial partner MS Technology A/S Copenhagen. This constitutes a new set of procedures capable of handling large files. These generate a database by means of which large concordance listings (in excess of 2000 lines) can be generated in no more than a few seconds, and right or left ordered in well under a second. Apart from conventional frequency lists, concordances and substring searches, the –3/+3 windows were found to be particularly insightful. These are illustrated below. The lemma "sure" and its supposed type "surely" were investigated to determine whether "surely" = SURE+ly. A search of Corpus A revealed the data shown in Figure 1. Figure 2 shows all contexts for "surely". As can be readily seen, "surely" cannot contain the sememe SURE as evinced in "sure +that/whether". One must therefore posit a new sememe SURELY. And returning to our primum mobile of the two learner questions, one must now ask how the lexicographer can best make this particular knowledge available. The gloss under the type "surely" in Chambers (1993):[10] "as it would seem (often *ironic*)" would appear to offer little insight.

The question of whether a plural form must share a sememe with the singular has been raised before, e.g. as long ago as 1973 by ERA 56,[11] who went so far as to claim: "La confusion du singulier et du pluriel sous la même forme canonique est néfaste du point de vue statistique." (op. cit. p. 21). Their sample corpus was however small, and their methods necessarily more laborious than need be the case today. Corpus A yields the data shown in Figure 3 for "future" and "futures". At first glance it appears that there is a measure of overlap between the two in the case of "gilt future(s)". Closer examination, however, shows that the 16 examples of "gilt future" can be accounted for by a fixed phrase "Liffe long gilt future" used as a heading, or variants of that phrase. But little casuistry is required to show that FUTURE $\neq$ FUTURES.

The argument can now be considerably extended by a consideration of a single lemma headword which has spawned many types and combinations, namely "employ". In Corpus A this occurs as shown in Figure 4.

These types and combinations of *–employ–* are seen as types generated by the language system, available as potential vehicles generated by the 'business' and values of Discourse Communities, selected on a 'best fit' basis measured in terms of values already associated with the components, and the assumption that context and common sense will clarify what new value is

intended by the use of an old shell (cf. Eliot's "husk of meaning"). But that best fit may be nevertheless a long conceptual jump, and the process may be repeated within the community any number of times. Thus the underlying concepts *UNEMPLOYMENT* and *EMPLOYER* may be quite distinct to expert members of the community. And the same may apply even to *EMPLOYER* and *EMPLOYERS*, as in the ERA 56 argument, and also suggested by the *unions/employers* juxtaposition in Figure 4. For further evidence one need only consider that these types need not, and often do not, share a morpheme when translated into other languages.

Finally we turn to Corpus B for an example of the characteristic chunkings which typify the language of specialist communities. The Government Bonds section of Corpus B covers 222 articles, 137,058 tokens, 5,318 types and 2,035 hapax legomena. The type *benchmark* occurs a total of 487 times. The profile for *benchmark* is shown in Figure 5. Most of these data can be reduced to the narrow set of options shown in Figure 6.

Benchmark can be seen to function here exclusively as a modifier, not as a noun. Language usage at morphosyntactic level in specialised communities can thus be shown to be just as idiosyncratic as at word level. How does one initiate the neophyte into this linguistic culture? And what role does the lexicographer have to play at this level?

5. Discussion and implications

The preceding material has demonstrated that some serious implications for lexicographical practice (as well as theory) emerge from the phenomena dwelt on. How does the above approach to lexical dynamics fit in with what is known and felt about dictionaries? It is firstly contended that the recent plea made by Knowles[12] is reinforced even further by the powerful and attractive functionality of innovative "dictionary software" such as ATA and analogous products.

One frequently encountered objection to dictionaries is that they are always (well) out–of–date by the time they are published. This claim normally focuses on the lack of current, up–to–the–minute neologisms in dictionaries but it also has validity with respect to allegedly 'static' definitions and 'frozen' collocational dynamics. Genuine argument has not really been possible hitherto because of difficulties in providing water–tight 'proof'. However, the position is now different: a genuinely up–to–the–minute generator corpus allied to a comprehensive and efficient delivery system can indeed provide 'chapter and verse' about such matters, as well as rendering complaints of the above sort null and void!

Another substantive — hitherto one might even have said intractable — problem is neutralised. Traditional dictionaries never could and cannot even now give any undertakings at all that a 'complete' list of sememes has been treated on a literally exhaustive basis. Computer–implemented lexical databases (LDB's) can, however, present for inspection arbitrarily large

numbers of in vivo citations. These, in their turn, can prime an extensive and intensive analysis or 'capture' of fine semantic differentiation or shading which is much more authentic because no fragmentation or rupture of context is necessary. If the lexical material relates to the use of language for specific purposes (LSP), this process is even less fraught than normal because of the reduction in 'scatter' observable in the discourse of professional communities. In these particular circumstances the yield of re–utilisable information — semanto–sociological, onomasiological and pragmatic — is notably high–grade. It is not pushing things too far at all to say that the general facilities under discussion — or to state it clearly: a writer's dictionary/LDB — work to best effect in the context of LSP/L2SP activity, most of all, it is submitted, with respect to the encoding / production of professional documents.

One of the more interesting such practical implications for lexicography, mentioned above, is the question of placement. If it is demonstrable that a particular word–form — or 'type' — carries, in terms of occurrence frequency in text, the greatest proportion of its associated lemma's functional load, then why should that word–form not itself be in a dictionary's headword list and, a fortiori, why should it not be the main locus of information about the lemma concerned, even to the point of clearly indicating, say, the predominantly adjectival function of a 'formal' noun in the majority of contexts? The corollary of this, of course, is that the actual canonical form of the lemma would appear in a headword list organised on this principle merely as a navigation beacon pointing elsewhere, to something functionally more important. This point, it is submitted, has particular force in the context of pedagogical lexicography and learner's dictionaries, in particular.

One justification, on the level of lexicographical theory, for such an approach to dictionary–making and –presentation would be the claim that the approach offers a method of reducing the gulf between the textual sense(s) of words in vivo and the in vitro meaning(s) , i.e. potential senses, of lexemes outside text, bereft of contextual linkages and — in the case under discussion — listed according to the formal and content–free mechanism of alphabetisation. In complete and encouraging contrast to that, what the freely configurable LDB offers is no less than a type of non–notional definition. The delimitation process occurs in situ and is primed in the simple case by encyclopaedic stimuli and in the more complex, qua general, instance by coherential pointers and clues.

The type of dictionary display needed for the above technique to carry complete conviction is that of an efficiently indexed LDB. This is the only working environment in which those consulting a dictionary so configured that is able to offer all the unconstrained facilities needed and to prevent such a system from being compromised by handling inadequacies: speed of retrieval, especially of citations; the formulation and execution of browse requests; cross–checking and grouping options. It is probably more accurate

and expedient to talk in terms not of a dictionary but rather of a 'dictionary shell' which can be flexibly configured to suit a whole range of purposes, languages and situations.

Another significant implication for lexicography appears to be that evolution of the sort described here makes it even more necessary for actual lexicographers to continue to deliver the results of their expertise in full whilst almost concealing the nature of the expertise itself! According to such an analysis, lexicographical expertise becomes, as it were, a somewhat different commodity which is delivered covertly rather than overtly. No longer would dictionary users, in these circumstances, have to go to the same lengths — that is to say, by entering the microcosm of lexicographers — to find out about words by having to find out, on a meta level, about the 'special' words and conventions used by lexicographers to communicate appropriate semantic and procedural information about the 'ordinary' words of everyday discourse and intercourse. The numerous citations banked in a lexical database could surely be largely left to do their own talking to interested enquirers instead of being defined and 'explicated' by third parties.

Notes

1 Robertson FA (1987), *Airspeak: Radiotelephony communication for pilots,* Prentice Hall.
2 Swales J (1990), *Genre analysis: English in academic and research settings,* CUP. See also in this connection: Nystrand M (1986), *The structure of written communication: studies in reciprocity between writers and readers,* Academic Press.
3 Johns AM, "L1 Composition theories: implications for developing theories of L2 composition" — in: Kroll B (1990), *Second language writing: research insights for the classroom,* CUP.
4 Bolinger D, "The atomisation of meaning" — in: Jakobovits LA / Miron MS (1967), *Readings in the psychology of language,* Prentice Hall.
5 On the notion of "chunk" as used in this context, see Skehan P (1992), "Second Language Acquisition Strategies and Task–Based Learning", in: *Thames Valley Working Papers in English Language Teaching",* Vol. I, Spring 1992.
6 The edition referred to passim is: "Prototractatus: An early version of Tractatus Logico–Philosophicus" edited by McGuinness, Nyberg and von Wright, with an interleaved English translation by Pears and McGuinness, published in 1971. This includes a facsimile of the author's manuscript, and all variations from the 1921 edition.
7 Swales J (1990, *Genre Analysis,* CUP.
8 During and after a lecture in Madrid in1987.
9 Roe P, "Astec: User's Guide to the Aston Corpus of Scientific and Technical English", Language Studies Unit, revised 1994.
10 *The Chambers Dictionary,* Chambers Harrap, 1993.
11 Geffroy A, Lafond P & Tournier M, ERA 56 au CNRS, ENS de Saint–Cloud, 1973.
12 Knowles F, "Dictionaries for advanced learners and users of foreign languages", in: *Verbatim* Vol. XIX/iii, 1993.

Contexts for *sure* (f = 68)

5 I	13 to	17 make	21 that	10 the	3 will
3 you	5 am	14 not	6 the	3 it	2 to
3 was	4 can	11 be	3 whether	3 are	2 they
3 he	4 and	4 making	3 it	2 you	2 of
2 will	3 not	3 made	3 I	2 we	2 have
2 we	3 is	3 am	2 you	2 was	2 any
2 wants	2 are	3 a	2 to	2 trust	2 always
2 want	2 I'm	2 was	2 they	2 that	
2 the	2 I	2 quite		2 none	
2 proposals	2 'I'm	2 for			

Contexts for *surely* (f = 27)

3 the	2 and	3 must	3 a	3 to	3 a
2 Mr		3 is	2 have	3 the	2 one
		2 would	2 be	2 time	2 of
		2 this	2 as		2 for
		2 as			
		2 are			

Figure 1. Frequency lists for types occurring within the range –/+3 of 68 occurrences of *sure* and 27 occurrences of *surely* for f>1.

unds have to be granted there is	surely a case for trying to remove some of
sixth of the earth's surface is	surely a more satisfactory way of priming
the bursting of the asset bubble	surely a transitory phenomenon it has beco
h sustained Durante's reading as	surely as did his partnering The dance bla
ernative interpretations just as	surely as the Liszt sonata or Beethoven's
ment which understood this would	surely be demanding a container load of de
ce the productivity damage would	surely be for me to go on holiday again un
d agents and direct sales forces	surely bear the lion's share of responsibi
o that before long somebody will	surely begin to do a regular survey of med
HUGH JONES Sir We hacks must	surely bow to Howard Davies's idea
nges in relevant technology' But	surely every business has to start modestl
from the concert suite Koechlin	surely had a point Les Biches so successfu
le of a recession the government	surely have higher priorities than aboliti
chairman Walls's reputation must	surely have suffered more than it has from
en in the first half of 1992 but	surely in time for a May 1992 election The
arketplace You are also and this	surely is the key much more limited in the
JUREK MARTIN For the fourth and	surely last time on President George Bush'
escaping Wassall's clutches must	surely lie in a white knight Logic and the
's Slater Walker connections are	surely more worthy of note True former
le serious social questions this	surely must be one of them Even in France
about it However the reasons are	surely pretty clear First politicians and
impetus will come from Wimpey is	surely right to dispose of the sort of low
ry official aka Mr David Mulford	surely simply rings one of many friends in
ongside snaps of Mr Smith waving	surely some mistake and sundry hard up
lable from any supplier and were	surely studied by a company where electric
cle Killings at Bisho SLOWLY BUT	surely the prospect of a stable democratic
urban event perhaps but hardly	surely unworthy of comment Calle's grim

Figure 2. All 27 contexts for *surely* in Corpus A.

Contexts for future (f = 399)

90 -	60 -	114 the	79 -	120 -	158 -
19 the	38 the	36 -	51 of	32 the	11 the
19 in	26 in	21 in	8 in	16 of	5 to
10 Liffe	19 for	16 gilt	8 The	9 a	5 be
8 to	16 about	15 near	7 and	7 in	5 and
8 for	14 long	15 for	5 the	6 will	4 on
6 of	13 of	10 a	4 growth	6 to	4 of
6 and	11 to	9 on	4 for	5 would	4 is
5 a	11 on	9 its	4 economi	5 with	4 in
4 that	9 over	8 of	4 as	3 that	3 time
4 confidence	9 and	8 foreseeable	4 EC	3 for	3 homelands

Contexts for futures (f =159

49 -	29 -	19 index	21 and	35 -	41 -
9 the	14 stock	13 -	18 contract	14 the	10 trading
7 in	12 the	12 the	17 -	13 options	10 at
7 Stock	8 of	7 and	16 market	6 and	8 the
4 of	7 Exchange	7 US	6 markets	5 The	4 to
4 The	4 in	7 Equity	6 exchanges	4 in	4 in
3 a	4 Liffe	6 sterling	5 trading	4 at	4 from
2 trading	3 natural	5 gas	3 were	3 to	4 a
2 introducti	2 second	4 gilt	3 opened	2 which	3 markets
2 for	2 options	3 property	3 contracts	2 were	3 September
2 by	2 cash	3 oil	2 scandal	2 was	2 is
2 and	2 arabica	3 by	2 rise	2 opened	2 future
	2 The	3 bund	2 remained	2 on	2 below
	2 Nymex	3 PLATINUM	2 reflected	2 lower	
	2 French	3 COFFEE	2 prices	2 hit	

Figure 3. Frequency lists for types occurring within the range –/+3 of 399 occurrences of *future* and 159 occurrences of *futures* for higher frequencies only.

EMPLOY (f =869)

196	unemployment	19	employs	2	employee's
185	employment	15	employ	1	unemploy-ment
149	employees	10	non-employment	1	un-employment
116	employers	9	self-employed	1	umemployment
36	unemployed	3	unemployable	1	self-employment
36	employed	3	non-employed	1	employer-provided
32	employee	3	employer-related	1	employer-led
24	employing	2	re-employed	1	employable
22	employer				

unions v. employers

by unions over the behaviour of	employers in the 1989 dockworkers'
the government, trade unions and	employers to scrap wage indexation,
Neither the unions nor the	employers totally reject reform of the curr
trade unions have united with	employers' organisations in

Figure 4. All occurrences of the string *–employ–* in Corpus A, including uncorrected literals. The four co–occurrences of *employers* and *unions* are also shown.

In [late (afternoon) trading/the cash market], (the yield on) the [treasury/x-year] **benchmark** [x-year/x-per cent/No. x] government [bond/issue/gilt] [due/maturing 19XX] [was up/was down/rose/ fell/opened (at/with)] ...

Figure 5. Syntactic profile for benchmark where "/" represents a choice commonly made and "()" a less frequent option. The frequencies of constituent items can be found in Figure 6.

152 late	148 trading	315 the	174 30-year	152 government	134 bond
96 yield	98 on	128 The	103 No	88 129	66 per
20 the	18 the	12 Treasury	54 11	57 3/4	63 issue
11 cash	17 with	12 10-year	29 bond	40 per	40 cent
9 trading	16 market	3 new	23 10-year	40 bond	25 was
8 on	8 contracts		21 9	22 145	23 opened
6 interest	8 afternoon		19 8	9 No	19 JGB
6 -	7 The		7 government	6 1/2	11 129
5 per	6 rates		6 gilt	5 the	7 moved
5 market	5 while		5 issue	4 maturing	6 the
4 and	5 day		4 no	4 issue	5 government
4 afternoon	5 cent		4 85	3 no	5 No

Figure 6. Higher frequencies of the profile for 487 occurrences of benchmark in the Government Bonds section of Corpus B. (NB: *No = Number.*)