The expanding lexical universe:

extracting taxonomies from machine-readable dictionaries

Willem Meijs

1. Introduction

In a sense a good monolingual dictionary can be regarded as a linear, alphabetically ordered representation of the passive and active vocabulary of normal, educated speakers of the language. Of course there is ample empirical psycholinguistic evidence to suggest that a language user's mental lexicon is not primarily organized in this way, as a long list of isolated elements. Rather, the mental lexicon forms a coherent, tightly-knit whole, whose elements are somehow intricately related to one another along a number of different dimensions: phonological, morphological, orthographical, etc. One of the most basic organizing dimensions, however, is no doubt the semantic-conceptual one, as witnessed by word-association and semantic priming tests: words activate, «call up», other words that are related to them in meaning (cf. the evidence surveyed in Meijs, 1988, 1989). A dictionary in machine-readable form in principle allows one to waive the alphabetical ordering of the printed book, and study its inherent or implicit semantic-conceptual organization. And this is exactly what we have tried to do at Amsterdam University in a number of related projects: «LINKS», «LEXALIZA» and «ACQUILEX».[1] As a result we are now able to trace this semantic-conceptual organization by means of a dynamic «taxonomy-browser» called «DEVIL» («DEcomposition VIa the Lexicon»).

2. Decomposition and meaning representations

In a normal monolingual dictionary the meanings of the words in the language are explained in the terms of (other) words of the same language. If one could chart out all the connections between the words, the result would be a kind of a huge multi-dimensional (and partly hierarchic) grid or network. The meaning of a word would be a function of the position it occupies in this network, and of its connections with all the other words in it, especially the ones to which it is most closely related. This would be one way, then, of representing the time-honoured view that: «the meaning of a word is its relations to all other words in the language».

1. 'LINKS' (short for 'LINKS in the Lexicon') was a three-year research project funded by the Netherlands Organization for the Advancement of Pure Research (NWO) under project number 300-169-007. 'LEXALIZA' is a four-year project funded by the Amsterdam University Arts Faculty under project number 202.121. 'ACQUILEX' (short for 'Acquisition of Lexical Knowledge for Natural Language Processing Systems') is a Basic Research Action programme funded by ESPRIT (project number BRA 3030; duration two and a half years), in which research teams from the Universities of Amsterdam, Barcelona, Cambridge, Dublin and Pisa collaborate.

The author (research supervisor of the first two projects and of the Amsterdam contribution to the third) would like to thank those who have been working with him on these projects: Inge van den Hurk, Marianne den Broeder, Jeanine Baader, Sylvia Janssen, Iskandar Serail and especially Piek Vossen.

Following Dik (1989) we assume that all of our knowledge is either perceptual or conceptual, and that to the extent in which it is conceptual it is at the same time linguistic. On this view, the combined and interconnected meaning-representations of the words in a dictionary can be regarded as a lexical knowledge-bank which constitutes a specific linguistic representation of (a large part) of our knowledge of the world. Such a view is compatible with the well-documented phenomenon that different cultures cut up reality in different ways, and that such differences are reflected in the meanings of the words in the respective languages. One effect of this is the well-known fact that it is sometimes impossible or very difficult to translate some concepts from one language (culture) to another – «'defining meaning' is a language-internal affair», as Dik (1989:86) puts it. For this reason, among others, Dik's theory of Functional Grammar (henceforth FG), from which we take our theoretical orientation, rejects abstract semantic features to represent word-meanings. Instead, in the FG approach conceptual knowledge is claimed to be stored in the form of basic predicates. A corollary of this is that the same piece of knowledge may in fact be stored in different ways, i.e. in different predicates.

Dictionaries show similar differences: different dictionaries seldom describe the conceptual content of one and the same entry in precisely the same combination of words. Nonetheless, as regards the overall underlying interdependence of the concepts defined, the shared cultural and linguistic anchoring leads to (implicit) classifications and taxonomies which show a great deal of overlap and congruence, even though the predicates actually employed in the wordings of the definitions may vary considerably.

In the LINKS project we have tried to trace the inherent semantic-conceptual organization of the ***Longman Dictionary of Contemporary English*** (LDOCE, for short - Procter, 1978). The basic methodology for the LINKS project was as follows: first an appropriate grammatical coding was applied to the words of the restricted vocabulary and their inflected forms. This coding was then automatically inserted in all of the meaning descriptions, the outcome being a grammatically coded 'corpus of meaning descriptions'. Subsequently a syntactic typology was developed for the structures of the meaning descriptions of each of the major parts of speech, i.e. nouns, verbs and adjectives, resulting in parser-grammars for each of them. Applying these grammars to the corpus we generated syntactically-analyzed meaning descriptions in which it is possible to systematically identify premodifiers, kernels, postmodifiers, etc. (for details see Vossen *et al.*, 1988, 1989).

The corpus of fully analyzed meaning-definitions was then subjected to thorough scrutiny to establish a typology of typical, recurrent patterns. In standard dictionary practice a prototypical definition consists of a definition kernel (the genus term), which puts the concept denoted by the word defined (the entry word) in a wider class, and of pre- and post-modifiers (the differentiae) which serve to set off the concept defined from other concepts in the same general class. In other words, the relation between entry word and definition kernel would be an IS-A, or hyponymy relation. If the word functioning as definition kernel itself corresponds to a dictionary entry, the process would repeat itself, and this might go on a number of times, leading to ever more general genus terms.

Obviously, after a limited number of cycles such chains of definitions would have to come to an end, somewhere, somehow. And this is indeed what we found: the definition-chains would either just peter out with some rather general kernel-word

that was not further defined (i.e. did not have a corresponding entry-word in the dictionary), or, more often, end up in circularity. Either way, one might regard the ends of such chains as corresponding to the 'primitives' of similar theoretically-based taxonomies. An example of such a prototypical definition-chain ending in circularity is given in Fig. 1:

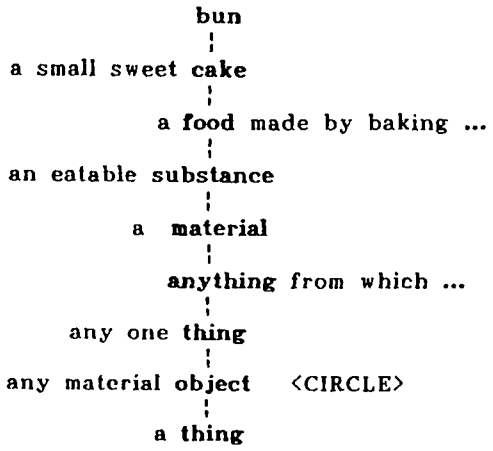

Fig. 1. Hyponymus definition-chain ending in circularity

3. Characteristic structures in meaning definitions

While most meaning-definitions indeed adhere to the prototypical homonymous pattern discussed in the preceding section, we found that a substantial number of them exhibit more complex patterns, which any attempt to automatically create taxonomies from related definitions would have to take into account. For nouns, for instance, our investigation of dominant definition-patterns yielded the following typology:

1. SYNONYMS
2. LINKS
3. LINKERS
4. SHUNTERS
5. SHIFTERS

- SYNONYMS are typical 'equality' cases: thus *abattoir* is 'defined' as '**slaughterhouse**', without any pre- or postmodifications; the two words (or rather their meanings) are simply equated, and one has to look at the definition of the definiens for further elucidation ('a **building** where animals are killed for meat').

- LINKS are the characteristic hyponomy cases: a *watchdog* is defined as 'a fierce **dog** used to guard property'.

- LINKERS are definitions in which the syntactic kernel is taken from a restricted set of 'relator words' mediating and labelling the relation between the

entry-word and the semantic head of the definition (often contained in a post-modifying '**of**-phrase'). Thus *candy* is defined as '**a type of** *sweet*', *stomach* as '**a part of the** *body*', *waste* as '**a stretch of** *land*', *band* as '**a group of** *musicians*', etc.

– SHUNTERS are cases in which the syntactic head of the definition is again a general 'relator-word', and in which the semantic burden is carried by an element from a different part of speech (Verb or Adjective in the case of nominal meaning definitions). Thus *actor* is defined as 'a **person** who *acts...*'; *gasp* as 'an **act** of *gasping*'; *ambiguity* as 'a **condition** of being *ambiguous*', etc.

– SHIFTERS, finally, are definitions in which the syntactic head of the definition is itself a transform of some kind of a word from a different part of speech. Thus *advent* is defined as 'the **coming** of Christ', *apiculture* as 'the keeping of bees', etc.

Notice that a unified treatment of these different types of definitions is possible once we realize that in all five of them in fact a **relation** is involved between the entry-word and the semantic head of the definition. In types (3), (4) and (5) this relation is made explicit in the relator-word which functions as the syntactic head of the definition. In the 'default' case (type (2)), the relation is not explicitly expressed (and hence the syntactic kernel coincides with the semantic kernel of the definition), but the implied relation in most cases is clearly either 'TYPE' or 'INSTANCE'. And in the case of SYNONYMS (type (1)), finally, the 'equality' relation is likewise implicit and therefore, as in type (2), the semantic and syntactic head of the definition once again coincide.

In Vossen *et al.* (1989) we pointed out that many noun definitions have such an explicit relator-word in (syntactic) kernel position, followed by a postmodifying construction (usually an *of-* phrase) containing the semantic genus term, and we gave some examples of possible candidates. Vossen (1990) provides a more complete list of relators, and discusses some characteristics of relators. Essentially relators serve to specify the relation of the entry-word concept to the genus-term concept in terms of notions such as 'quantity', '(non-)countability', 'collectivity', 'composition', 'structure', 'membership', 'instantiation', etc. There is a roughly inverse relationship between the degree of generality and semantic content of these relators on the one hand and their relative frequency in the definitions on the other. Thus *type* occurs 1308 times as a relator, *piece* 766 times, *part* 679, *group* 437, *kind* 257, etc., and in this way the scale goes down gradually to words like *pile* (19), *ball* (19), *bunch* (18), *block* (17), *lump* (16), *bar* (15), *circle* (13), etc.

Like the words at the top of the frequency scale these relators towards the bottom end denote instances, parts, collections, etc., but they do so in more specific, often collocationally determined ways, involving in addition notions such as shape, size, arrangement, texture, and so on. In a sense there is a kind of slot/filler relationship between the words towards the top and those lower down on the relator-scale. Thus, while a *wreath* is no doubt a 'collection' of flowers, the definition in terms of a 'circle' of flowers further specifies this collection as involving a particular arrangement of the items concerned. Similarly, while *ingot* (in one of its senses) is certainly a 'piece' of some precious metal, the definition in terms of a 'bar' of gold or silver literally gives that piece of precious metal a particular shape.

4. Hierarchies and taxonomies

In the context of the ACQUILEX project a program for automatic taxonomy building-and-browsing has now been developed, called 'DEVIL' (DEcomposition VIa the Lexicon - Vossen and Serail, 1990). This program allows one to systematically trace the taxonomic relationships in a database of analyzed definitions like the one developed for LDOCE in the LINKS project, taking due account of the presence of Linkers, Shunters, Shifters, etc. The system, which uses a special data-structure called 'L-tree' (Skolnik, 1980) that allows fast access and retrieval, can be used interactively or in batch mode and in either bottom-up or top-down mode. The prototype version of DEVIL produces output like that in Fig. 2:

```
dandelion (bottom up):
     (00.00) flower <TYPE>
          (01.01) plant <PART>
               (AA.AA) thing <LINK> ATOM

line (bottom up):
     (02.18) soldier <ARRANGEMENT>
          (01.01) army <MEMBER>
               (00.01) force <LINK>
                    (01.06) people <GROUP>
                         (01.01) person <LINK> ATOM

line (topdown):
     zigzag 01.00   wrinkle 01.01  worm 01.04     waistline 00.01
     vertical 02.01 verse 00.02b   vein 00.03     trunk_line00.01
     trawl 02.02    tramline 00.01 track 01.05    track 01.04
     touchline 00.00      tie 01.07 thread 01.04  (...)
```

Fig. 2. Some examples of DEVIL output.

Thus the bottom-up search from *dandelion* shows that it is a type of *flower*, which is itself defined as part of a *plant*, and the chain ends at the 'meaning-atom' *thing*. One of the many senses of *line* leads up to an arrangement of *soldiers*, who are defined as members of an *army* in the sense of *force*, which means that it is a group of *people*, and the chain ends at the meaning-atom *person(s)*. The last example shows a few of the many (more than 150) descendants of *line* in top-down mode.

To create reliable networks of the semantic relations hidden in the interconnections between word-senses in dictionaries, disambiguation of the words used in the definitions is clearly essential, i.e. it must be unambiguously clear in which sense any word is being used in a dictionary definition. For LDOCE various researchers are working on this kind of disambiguation (cf. Wilks *et al.*, 1989; Vossen, 1990). Given the daunting amounts of data to be dealt with, such disambiguation is a formidable job. The task is alleviated somewhat by the fact that LDOCE uses a controlled vocabulary for its definitions and examples. However, since these are all common words, they are also ones that tend to have relatively many senses. The job is further complicated by the fact that the defining vocabulary also contains many complex words

derived from the controlled vocabulary. However, as Vossen (1990) points out, a number of heuristic strategies may be employed to do this semi-automatically. This disambiguation has now been completed semi-automatically for the most central and most frequently used words, and the results have been incorporated in the DEVIL system. Thus the numbers between brackets in Fig. 2 are homograph and sense numbers. The marking (02.18) for *line*, for instance, means that we are dealing with the second homograph and the eighteenth sense of that homography entry. The marking (00.00) means that there is only one homograph, with only one sense, and (AA.AA) denotes an element which is atomic to the system.

The DEVIL system provides us with a powerful tool to study the taxonomies and hierarchies inherent in dictionary-definitions, in terms of breadth, depth, consistency etc., giving us a detailed overview of the taxonomic relationships embodied in the definitions in LDOCE (and in other dictionaries, such as the Van Dale Dutch-Dutch and Dutch-English dictionaries that are going to be 'DEVILized' in the ACQUILEX context).

By way of illustration Fig. 3 (from Vossen 1990) shows how a large portion of the abstract nouns in LDOCE 'hang together' taxonomically, via their definitions:

Fig. 3. Part of the top of the abstract-noun taxonomy in LDOCE

This tree diagram shows part of the top of the abstract noun taxonomy in terms of the number of times a particular word functions as the (syntactic) kernel of the definition of an abstract noun entry, plus some examples of branches further down. Notice that towards the top most of these are typically what we call 'Shunters', i.e. they are themselves general, semantically relatively empty words, which function as pointers to the verbs and adjectives in the post-modifying phrase (usually a nominalized predicate of some kind: 'the condition of being ambiguous', 'an act of gasping', 'something which annoys', etc.).

On the basis of the LDOCE material we have come to distinguish three basic levels in the hierarchies:

- bottom-level: words that do not occur as heads of definitions of other words;
- core-level: words which occur very frequently as difinition-kernels, and
- top-level: a small set of very general circularly-defined or 'dangling' words in which all chains end.

In our dictionary material the above three-level picture is based on a computationally verifiable distinction. Taxonomies with a depth of more than three result from 'recycling' within the core-level. Interestingly, the depth and overall characteristics of the taxonomies that emerge are in many ways quite similar to findings in empirical and theoretical discussions of taxonomies (cf. Berlin *et al.* 1973; Rosch, 1976; Rhodes, 1985). Thus folk-taxonomies typically have a depth of three or four, while scientific taxonomies tend to have a depth of five or more. For a more detailed discussion of these aspects see Vossen (1990).

5. Expanding DEVIL's lexical universe

So far our work on the DEVIL system has been mainly concerned with Nouns. There is a prototype version for Verbs, but the chains that are created at this stage are not very revealing. Adjectives have not yet been 'DEVILized', but they will be in the near future, for there are now plans to expand and refine the DEVIL system in such a way that it can constitute the central lexicon in a knowledge-engineering context (the 'LIKE' framework: Linguistic Instruments in Knowledge Engineering' - cf. Weigand, 1990). Expanding DEVIL's 'lexical universe' in such a way that the three major parts of speech are integrated to allow effective knowledge-handling constitutes the major new challenge in the version of the DEVIL system now being developed, in which we are charting out new trails, allowing shifts back and forth between nominal, verbal and adjectival taxonomies as required, and catering for a wide range of query-types relevant to inferential logic and knowledge-handling generally, such as:

- which category of items (activities, events, states, etc.) is X an instance of?
- which items (events, states) exhaustively or typically constitute exemplifications of category X?
- what kind of item (etc.) is X a part of?
- what are typically parts of which X is composed?
- what are the activities (functions, states, etc.) typically associated with items of category X, specified for case role (Agent, Instrument, etc.)?

- what are the categories of items typically associated with activities (etc.) of category X, specified for case role?

In this expanded version it will be possible to trace genus and differentiae information via a generalized entry format in which the scope of pre- and post-modifying elements is taken into account. In building up this new DEVIL version for LDOCE we will use other kinds of semantic information available in the machine-readable version as a means of checking and refining the reliability of the data. In particular we will look at the so-called 'box-codes' and 'subject-field codes' preceding many of the definitions.

'Box-Codes'. Most important here are the codes that occur in positions 5, 8 and 10. Essentially these constitute a branching set of hierarchically related classificatory categories, with a first division into 'abstract' and 'concrete'; 'concrete' branching into 'animate' and 'inanimate', 'animate' into 'plant', 'animal', 'human', etc. In combination with the syntactic part-of-speech labels elsewhere in the entry these codes give important information both about the items themselves and about the way they combine with other elements in utterances in which they can occur. For nouns the codes indicate to which (sub)branch of the classificatory hierarchy they belong, for adjectives they point to the semantic category to which the noun which they modify (attributively or predicatively) may or should belong, and for verbs the codes refer to the categories to which their arguments (subject, direct and indirect object) are supposed to belong. Box codes refer to specific senses of lexical items; i.e. a verb may require a human agent in one sense and an abstract one in another sense.

LDOCE's box-code hierarchy is rather lopsided, however: it is fairly detailed and refined on the 'Concrete' side, but under-developed on the 'Abstract' side. The 'Abstract' branch of the tree can probably be refined semi-automatically, however, by reference to information in the definitions. Thus it will be possible to subdivide 'Abstract' for most lexical verbs and many nouns (especially morphologically derived ones) into 'Act(ion)', 'State', 'Process', etc., thanks to the fact that the definitions often systematically make use of phrases such as 'the act/state/process of X-ing...' etc. (cf. Fig. 3). One important distinction that is conspicuously absent from the LDOCE hierarchy is the notion 'Artefact'. Here again, it may be possible to enhance the hierarchy with this notion (or some related label like 'Purposeful'), by making use of systematically-occurring expressions like 'made from X', 'used to X', 'used for X-ing...', etc., in the definitions.

'Subject Field Codes'. These constitute a large set of markers indicating the semantic fields to which a word (once again in a particular sense) refers/belongs, ranging from 'basketball' and 'entertainment' to 'dentistry', 'music', etc. Some fields are rather wide ('economics', for instance), others quite narrow ('cricket'). Many fields are further divided into subfields (for instance 'accounting', 'banking', 'taxation' as subfields of 'economics').

The subject field divisions vary widely and in a rather unsystematic fashion in their degree of specificity and organization. Thus there is a field labelled 'sp' for 'sports', which covers a few subfields like 'archery', 'mountaineering', etc., but the bulk of what people normally take to be sports have separate field-labels of their own (some 15 of them, with over 30 subfields). Similar observations can be made about other areas such as games, crafts, branches of science, etc. We intend to do a few major reshuffles here,

grouping the various fields with their subfields into a limited number of broad, general areas such as GAMES, ARTS, SPORTS, NATURE, TRANSPORT, INFORMATION, etc. Some fields or subfields may in fact cross-classify with respect to these wider areas. Thus it is clear that the 'hunting & fishing' field (hf) with subfields like 'fisheries', 'falconry', etc., while primarily classified in the SPORTS area, is also closely connected with the NATURE area.

It is obvious that the information in the box-codes is more general and 'higher' than that in the subject-field codes. The relation between the two kinds of information is roughly similar to the distinction between HPRIM and LPRIM features in the Aarts and Calbert (1979). One could say that the box-code information '(sub)categorizes' the lexical items concerned, while the subject-field codes '(sub)classify' them. The arrangement of the box-codes is clearly hierarchic, while the subject-field codes involve only a marginal amount of hierarchic structure. The two systems are essentially independent, i.e. they do not form a continuous hierarchy. Thus a word like *chef* links up with 'M' (Male, implying Human, Animate and Concrete) in the box-code hierarchy, and extending to 'fo' (food) with subclass 'c' (cookery) in the subject-field hierarchy, while a word like *soup* also extends to 'fo' and 'c' in the box-code hierarchy, but this time attaching to 'L' (Liquid, implying Inanimate and Concrete) in the box-code hierarchy.

The fully-analyzed meaning-definitions can be used to improve the highest-level categorizing hierarchy. They are also invaluable in disentangling various other knotty problems. Thus quite often bracketed portions can be linked up systematically with different kinds of syntactic and semantic information. For instance, often the only way to 'disambiguate' a verb which may syntactically have both a transitive and an intransitive reading (and hence semantically different argument-frames) is to hunt for a bracketed portion containing the object (often 'something' or 'someone'). Similarly, in some definitions noun and adjective meanings are conflated, as in '(a supporter of) the system of government introduced in the USSR in 1917', where the inclusion or exclusion of the bracketed portion corresponds to the [Human] Noun versus [Abstract] Adjective reading.

Often the differentiae reinforce the other kinds of semantic information. Thus the presence of the word *aircraft* in the first definition of *pilot* obviously goes with the subject-field label 'ae' ('aeronautics') for that first meaning, while the presence of words like *water, ships* and *harbour* of course underlines the label 'na' going with the second meaning of *pilot*. Such correspondences can in principle be used to 'double check' the consistency of the hierarchies, thus enhancing their reliability for inferencing and other kinds of knowledge-handling.

References

AARTS, J. and CALBERT, J. P. 1979: *Metaphor and Non-Metaphor*. Tübingen: Niemeyer Verlag.

BERLIN, B., BREEDLOVE, D.E. and RAVEN, P.H. 1973: «General principles of classification and nomenclature in folk biology». *American Anthropologist*, 75:214-42.

DIK, S.C. 1989: *The Theory of Functional Grammar. Part I: The Structure of the Clause*. Dordrecht: Foris.

MEIJS, W. J. 1988: «Knowledge-activation in a large lexical data-base: problems and prospects in the LINKS-project». In *Amsterdam Papers in English (APE) No I*, Amsterdam: Amsterdam University English Department, 101-23.

- 1989: «Spreading the word: knowledge-activation in a functional perspective». In J. Connolly & S. Dik (eds.) *Functional Grammar and the Computer*, Dordrecht: Foris, 201-15.

PROCTER, P. (ed.) 1978: *Longman Dictionary of Contemporary English*. Harlow: Longman.

RHODES, R. 1985: «Lexical taxonomies». In G.A.J. Hoppenbrouwers, P.A.M. Seuren and A.J.M.M. Weijters (eds.), *Meaning and the Lexicon*, Dordrecht: Foris, 458-70.

ROSCH, E. 1978: «Principles of categorization». In E. Rosch and B.B. Lloyd (eds.) *Cognition and Categorization*, Erlbaum: Hillsdale.

SKOLNIK, J. 1980: *L-trees*, Technical Report. Amsterdam: Arts Faculty Computer Department, University of Amsterdam.

VOSSEN, P. 1990 (forthcoming): «The end of the chain: Where does decomposition of lexical knowledge lead us eventually?» To appear in: *Proceedings of the 4th Conference on Functional Grammar, June 1990, Kopenhagen* (provisional title).

VOSSEN, P., DEN BROEDER, M. and MEIJS, W.J. 1988: «The LINKS project: building a semantic database for linguistic applications». In M. Kytö, O. Ihalainen and M. Rissanen (eds.) *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora*, Amsterdam: Rodopi, 279-93.

VOSSEN, P., MEIJS, W.J. and DEN BROEDER, M. 1989: «Meaning and structure in dictionary definitions». In B. Boguraev and E. Briscoe (eds.), *Computational Lexicography for Natural Language Processing*, London: Longman, 171-92.

VOSSEN, P. and SERAIL, I. 1990: *Devil: A taxonomy-browser for decomposition via the lexicon*, ACQUILEX Working Paper, Amsterdam University.

WEIGAND, H. 1990: *LIKE: Linguistic Instruments in Knowledge Engineering. A Framework for Research* (restricted draft version), Tilburg: Infolab KUB.

WILKS, Y., FASS, D., GUO, C., MCDONALD, J., PLATE, T. and SLATOR, B. 1989: «A tractable machine dictionary as a resource for computational semantics». In B. Boguraev and E. Briscoe (eds.), *Computational Lexicography for Natural Language Processing*, London/New York: Longman, 193-228.