

Matthew Whelpton

Argument Structure – For Mental Dictionaries Only?¹

1 Introduction

There has been a long-standing division of interest between dictionary-makers and researchers in theoretical linguistics. This is reflected in the status of the grammatical information in the “dictionaries” developed by each group. Developments over the last 15 years have however opened up active areas of interaction. In this paper, I would like to look at that division and point to some of the interesting ways in which it has been breaking down.

2 Of lexicographers and linguists

2.1 The lexicon

At its minimum, a traditional dictionary is a list of citation forms and word senses. It may also include a range of other information, typically: broad phonemic transcription; information on inflectional morphology; part of speech; examples of use; and etymologies. The traditional conception of a dictionary is as an encyclopaedia of words.

¹This paper is based on a lecture that I gave at the Orðabók Háskólans on Friday 1st April 2005. I would like to thank the participants in that seminar for their comments and insights. Needless to say, all faults and follies that remain are mine solely!

The core specifies the main semiotic coupling – form and sense. This is enriched with as much general information concerning the word as is possible given restrictions of space (the great proviso!).

Within modern linguistic theory, the role of the lexicon has been very different. In Chomsky's (1957) earliest characterisation, no lexicon was necessary at all. Words were inserted into syntactic trees by rewrite rules, as shown in Figure 1.

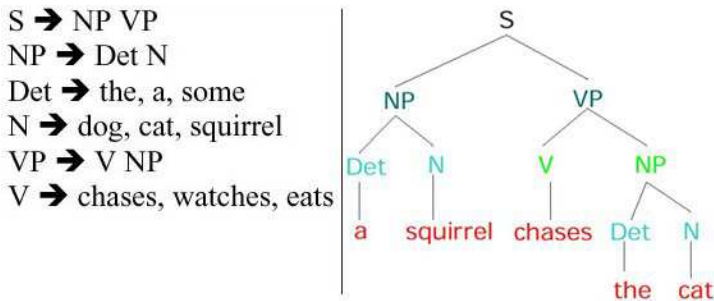


Figure 1: A simple rewrite rule grammar without lexicon

It was soon realised however that making the syntax do everything weakened rather than strengthened it. The role of the syntax was to express general principles of sentence organisation. Many aspects of sentence form, though, are related to the idiosyncratic properties of the words being used. In Chomsky's (1965) account, syntactic rules express exclusively syntactic generalisations (e.g. where do count nouns go) and a lexicon is introduced which lists all the words of the language, tagged with the syntactic information about them that must be learned (and cannot simply be deduced).

The lexicon was thus introduced as a separate module of the grammar, though its function was limited to that of a word store. From this point on within Chomskian work, the lexicon served to list the idiosyncratic grammatical information which the word carried into the sentence. Below are some examples of the kinds of grammatical information that have been suggested.

1. Rely [___ on NP]
 - a. John relied on him.
 - b. *John relied under him.
 - c. *John relied against him.
2. Tell [___ NP_{Acc} SubC_{that}]
 - a. John told him that Mary had left.
 - b. *John told to him that Mary had left.
 - c. *John told that Mary had left.
3. Say [___ (to NP) SubC_{that}]
 - a. *John said him that Mary had left.
 - b. John said to him that Mary had left.
 - c. John said that Mary had left.

This information on the number and type of phrases that must follow the Verb is called a subcategorisation frame.

The kinds of information we see in subcategorisation frames illustrates very clearly what I am calling the division of interests between theoretical linguists and lexicographers. Traditional dictionaries do not require the listing of grammatical information and usually do not list it except for part of speech and examples of use – e.g. Give something to someone. The linguistic lexicon contains little but grammatical information and only those idiosyncratic elements of grammatical behaviour which are not predictable from general syntactic rule.

However, two general developments have changed this significant difference in focus between the two communities of language researchers: first, the rise of corpora and computational modelling in language research, esp. in dictionary development; and second, the rise of semantics, in particular a clear and well-articulated field of formal semantics alongside syntax.

2.2 Semantic Theory

The work of the philosopher Richard Montague (1974) began a tradition of linguistic analysis which has established the importance of compositionality for any serious formal theory of semantics – that is, the property that the semantic interpretation of a sentence can be read systematically off the syntactic structure of the sentence.

Independently, within the Chomskian tradition, researchers were arguing for the importance of general semantic classes of relations between items which helped to establish their syntactic behaviour – so-called thematic roles and related concepts (see Jackendoff 1972 and Fillmore 1966). Thematic roles were adopted in Chomsky's (1982) Government and Binding Theory as a way of regulating well-formedness of grammatical structures: and theta roles, like preposition selection, had to be listed in the lexicon.

Thematic role theory was often driven by discussions of “alternations”, alternative patterns of syntactic realisation.

4. Passive

- a. John hit Bill. Agent - Patient
- b. Bill was hit by John. Patient – Agent

5. Dative alternation.

- c. John gave a book to Bill. Source – Theme – Goal
- d. John gave Bill a book. Source – Goal – Theme

At the crudest level, this meant that the lexicon began to include some form of semantic information for syntactic purposes.

6. Argument structure – *give*

- a. [___ NP_{Acc/Theme} to NP_{Goal}]
- b. [___ NP_{Dat/Goal} NP_{Acc/Theme}]

It also however stimulated a strong tradition of research concerning the regularities expressed by labels such as “Theme” and “Goal” and the patterns of syntactic realisation with which they were associated. And this brings us to one of the most promising areas in which contemporary linguistic theory and lexicographic research can be seen to interact.

An excellent example for our purposes is the work of Beth Levin, as it explicitly relates syntactic behaviour and conceptual structure through the properties of individual verbs: especially her 1993 work on English verb classes. Levin (1993) classifies verbs in terms of the alternations that verbs can and cannot undergo. She then looks for the elements of conceptual meaning that verbs which behave the same way syntactically share and which differ systematically from the verbs which do not share their syntactic behaviour.

Consider for instance Saeed's (1997) excellent introductory discussion of Levin's analysis of the three constructions – Middle, Conative,

Body Part Ascension – as they apply to four verbs – *cut*, *break*, *touch*, *hit*. The results are summarised in Figure 2.

7. Middle

- a. The bread cuts easily.
- b. Crystal vases break easily.
- c. *Cats touch easily.
- d. *Door frames hit easily.

8. Conative

- e. Margaret cut at the bread.
- f. *Janet broke at the vase.
- g. *Terry touched at the cat.
- h. Carla hit at the door.

9. Body part ascension

- i. Margaret cut Bill’s arm // Margaret cut Bill on the arm.
- j. Janet broke Bill’s finger // *Janet broke Bill on the finger.
- k. Terry touched Bill’s shoulder // Terry touched Bill on the shoulder.
- l. Carla hit Bill’s back // Carla hit Bill on the back.

	Touch	Hit	Cut	Break
Conative	No	Yes	Yes	No
Body-part	Yes	Yes	Yes	No
Middle	No	No	Yes	Yes

Figure 2: Verb-Construction Matching

Levin argues that the availability of a particular construction is linked to the presence of a conceptual component in the meaning of the relevant verb, as shown in Figure 3.

touch	Verb of pure contact	CONTACT
hit	Verb of contact by motion	CONTACT, MOTION
cut	Verb of caused change of state by contact in motion	CAUSE, CHANGE, CONTACT, MOTION
break	Verb of caused change of state	CAUSE, CHANGE

Figure 3: Conceptual Analysis of 4 Verbs

The specific component required by each of the constructions is shown in Figure 4:

CONSTRUCTION	COMPONENT REQUIRED
Middle	CAUSE, CHANGE
Conative	MOTION
Body-part Ascension	CONTACT

Figure 4: Conceptual component selections of 3 constructions

Here are the original examples, marked with the relevant conceptual components:

10. Middle = CAUSE, CHANGE
 - a. The bread cuts easily. = CAUSE, CHANGE, CONTACT, MOTION
 - b. Crystal vases break easily. = CAUSE, CHANGE
 - c. *Cats touch easily. = CONTACT
 - d. *Door frames hit easily. = CONTACT, MOTION
11. Conative = MOTION
 - e. Margaret cut at the bread. = CAUSE, CHANGE, CONTACT, MOTION
 - f. *Janet broke at the vase. = CAUSE, CHANGE
 - g. *Terry touched at the cat. = CONTACT
 - h. Carla hit at the door. = CONTACT, MOTION
12. Body part ascension = CONTACT
 - i. Margaret cut Bill's arm // Margaret cut Bill on the arm. = CAUSE, CHANGE, CONTACT, MOTION
 - j. Janet broke Bill's finger // *Janet broke Bill on the finger. = CAUSE, CHANGE
 - k. Terry touched Bill's shoulder // Terry touched Bill on the shoulder. = CONTACT
 - l. Carla hit Bill's back // Carla hit Bill on the back. = CONTACT, MOTION

Levin's work is in effect refining the conceptual elements that might appear in a dictionary entry for a verb – and a dictionary that adopted such “primitive conceptual vocabulary” for its definitions would provide an invaluable resource for linguists following up on Levin's work.

Thus, semantic developments in generative grammar have led to analyses that allow a fruitful interaction between lexicography and

linguistics. Such work however is also related to the other development mentioned earlier – corpora. Levin’s work would not have been possible without the availability of computer corpora. Corpora make it possible to investigate directly the patterns in which particular words appear. A grammatically tagged corpus allows us to investigate the full range of “alternate syntactic environments” in which verbs appear.

2.3 Corpora and lexical analysis

This brings us to another example of work where both lexicographers and linguistics can share and interact – the WordNet project (<http://wordnet.princeton.edu/>).

“WordNet® is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.”

- o The verb cut has 41 senses (first 10 from tagged texts)
- o 1. (58) **cut** -- (separate with or as if with an instrument; "Cut the rope")
- o 2. (18) **reduce, cut down, cut back, trim, trim down, trim back, cut, bring down** -- (cut down on; make a reduction in; "reduce your daily fat intake"; "The employer wants to cut back health benefits")
- o 3. (3) **swerve, sheer, curve, trend, veer, slue, slew, cut** -- (turn sharply; change direction abruptly; "The car cut to the left at the intersection"; "The motorbike veered to the right")
- o 4. (2) **cut** -- (make an incision or separation; "cut along the dotted line")
- o ...

Figure 5 - WordNet entry for verb “cut”

Such work is obviously of direct interest to lexicographers but as the description of WordNet shows, there is a strong linguistic motivation for the project, again in terms of the conceptual structures that underlie lexical organisation.

This development is taken much further by Charles Fillmore (Berkeley) in the FrameNet project (<http://framenet.icsi.berkeley.edu>). This is an attempt to model word properties within a cognitive grammar framework in which syntactic distributions and cognitive frames of reference for word senses come together.

As in WordNet, a verb is listed with a number of senses, though in this case, the senses are not defined in terms of synonym sets but

rather in terms of a much more complex framework of conceptual information, as shown in Figure 6.

- cut.v (Cause change of scalar position) **Created** Lexical entry
- cut.v (Experience bodily harm) **Finished** Initial Lexical entry Annotation
- cut.v (Cause harm) **Finished** Initial Lexical entry Annotation

Figure 6 - FrameNet entries for the verb *cut*

The number of senses of a word is as many as the number of frames of cognition within which it can be cognised or understood. For instance, one sense of the verb *cut* falls within the Cause_harm frame. A frame represents a structured understanding of what is involved in a “harm-causing”, as shown in Figure 7.

Cause_harm

Definition:

The words in this frame describe situations in which an **Agent** or a **Cause** injures a **Victim**. The **Body part** of the **Victim** which is most directly affected may also be mentioned in the place of the **Victim**. In such cases, the **Victim** is often indicated as a genitive modifier of the **Body part**, in which case the **Victim** FE is indicated on a second FE layer.

Figure 7 – Frame for Causing_harm

The frame is effectively a constellation of elements related in a systematic way which provides the context of understanding of particular instances. Here, within the Cause_harm frame, we have an Agent/Cause and a Victim as the primary elements of the frame with provision made for the affect on the Victim being realised on a Body Part of the Victim rather than on the Victim holistically. Notice that there is a mix here of conceptual elements (Cause) and formal elements (“genitive modifier”). In this sense the frame looks like an elaborated thematic description and the frame elements look rather like traditional thematic labels, relating elements of conceptual structure to particular formal realisations. Within this framework however there is no attempt to limit the relation between a frame element and its syntactic realisations. Such relation simply are as they are.

An entry for the verb *cut* in this particular sense will then appear as in Figure 8.

Frame: Cause_harm

Definition

COD: make an opening, incision, or wound in (something) with a sharp tool or object.

Frame Elements and Their Syntactic Realizations

The Frame elements for this word sense are (with realizations):

Frame Element	Number Annotated	Realizations(s)
Agent	(40)	CNI.-- 5exx DNI.-- 1exx NP.Ext 34exx
Body_part	(36)	NP.Ext 4exx PP[on].Dep 1exx NP.Obj 31exx
Victim	(39)	2nd.-- 32exx NP.Ext 1exx INI.-- 2exx DNI.-- 1exx NP.Obj 3exx

Figure 8 – FrameNet entry for “cut.v”

This entry contains a specification of the items’s part of speech (*v.*), the frame within which it should be interpreted (which will also supply the frame elements to be mentioned in a moment), a sense which defines the specific conceptual delineation of this activity, and a grid relating the elements of the frame within which this word sense is embedded to the specific syntactic realisations of that element in uses of this particular lexical item. Notice that there are therefore two “sense” elements in this “definition” – the specific “micro” details of the activity itself (very reminiscent of a traditional dictionary definition) and the “macro” context of interpretation by which the individual words sense is related to others by a shared context of understanding.

The syntactic realisation information from Figure 8 shows the range of realisations for a particular frame element. However, the information can also be represented in a way more reminiscent of the kind of subcategorisation frame discussed earlier. Such labelled valence patterns are shown in Figure 9:

cut.v

Valence Patterns:

These frame elements occur in the following syntactic patterns:

Number Annotated	Patterns		
1 TOTAL	Agent	Body_part	
1 exx	NP Ext	NP Obj	
35 TOTAL	Agent	Body_part	Victim
4 exx	CNI --	NP Ext	2nd --
1 exx	CNI --	PP[on] Dep	NP Ext
1 exx	DNI --	NP Obj	2nd --
27 exx	NP Ext	NP Obj	2nd --
2 exx	NP Ext	NP Obj	INI --
4 TOTAL	Agent	Victim	
1 exx	NP Ext	DNI --	
3 exx	NP Ext	NP Obj	

Figure 9 – Valence patterns for “cut.v”

So far we have discussed the frame as the overarching interpretative umbrella under which particular word senses reside. But one frame is itself of course just one structured cluster of conceptual information which is related to others in a variety of ways. The frame Cause_harm is thus embedded within a network of other frames, stretching out into a vast network of “world knowledge”, as shown in Figure 10.

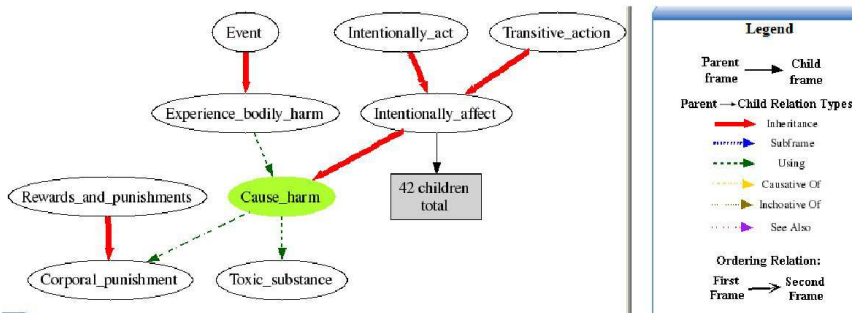


Figure 10 – Position of Cause_harm frame in knowledge network

The parent frame of *Cause_harm* is thus the frame *Intentionally_affect* and inherits from it a particular set of conceptual properties. This frame can also bear other relations than inclusion/inheritance. For instance, it will be linked to particular structured frames which describe possible Causes or Agents of the harm – in Figure 10 for instance we see the frame *Toxic_substance* listed. In this way, the representation attempts to address the problem faced by all who study lexical meaning that the sense of particular words in context seems ultimately to pull in or play on our total knowledge of how the world (and our culture) works.

The FrameNet project thus aims at the ultimate ambition of providing a computer-tractible model of human conceptual structure, within which language is embedded. I said earlier that traditional dictionaries are essentially encyclopedias of words – the FrameNet project effectively claims that an encyclopedia of words, properly pursued as a part of cognitive science, will be nothing less than an encyclopedia of the mind. . .

3 Conclusion

So where does that leave us? The approaches of Levin, the WordNet project and Fillmore's FrameNet project show a range of possible ways that theoretical linguists and dictionary makers can interact. Certainly, linguists may still pursue explicit models of syntactic organisation and lexicographers may still pursue an encyclopaedic listing of the diverse properties of words. But linguists can pursue their work more effectively with the resources that a corpora-driven grammatically-sensitive dictionary can now offer. And lexicographers can expand and enrich the sophistication of the information they provide concerning a word's syntactic behaviour through the research efforts of linguists.

What kinds of information might a linguist find useful? A list of the different phrases that form a word's complementation; a list of the case assignments associated with these complementation patterns; a classification of verbs in terms of "conceptual components" or "sense type", e.g. Transfer, psychological experience, internal change of state, external change of state etc. A linguist can more easily investigate theoretical questions using data organised in this way – and I would

hope that in pursuing theoretical questions of this kind, the linguist will refine and throw further light on the kinds of conceptual characterisations and syntactic constructs that are needed in characterising the properties of words effectively – which is presumably what lexicographers want.

Certainly, recent developments should act as a spur for lexicographers, linguists, psychologists and computer scientists to work more effectively together!

References

- Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton & Co.
- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. 1982. *Some concepts and consequences of the theory of government and binding*. Cambridge, MA: MIT Press.
- Fillmore, C. 1966. The case for case. *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston.
- Jackendoff, R. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.
- Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: University of Chicago Press.
- Montague, R. 1974. *Formal Philosophy*. New Haven: Yale University Press.
- Saeed, J. I. 1997. *Semantics*. Oxford: Blackwell.
- Websites accessed 31st March 2005

Útdráttur

Greinin fjallar um stöðu rökliðagerðar (snertiflatar setningar- og merkingarfræði) í þróun orðabóka af ýmsum gerðum. Byrjað er á því að bera það gjörólíka viðhorf til merkingar og setningafræði sem kemur fram í eldri orðabókum og í fyrstu ritum málkunnáttufræðinga (generatífista). Síðan er talað um vaxandi skörun þessara tveggja sviða og möguleika á frjórri samvinnu milli orðabókafræði og málvísinda, með hliðsjón af kenningum Beth Levin á sviði málkunnáttufræði, af vinnu við gerð orðasafnslegra málheilda innan WordNet verkefnisins og af vinnu Fillmore á sviði hugrænna (kognítífra) málvísinda innan

FrameNet-verkefnisins. Niðurstaðan er sú að aldrei hafa möguleikarnir á samvinnu orðabókafræði og hinna ýmsu málvísindastefna til skýringa á rökliðagerð verið meiri.

Keywords:

argument structure, lexical sense, semantic frame, semantics, valency

Dr. Matthew Whelpton

Senior Lecturer (Associate Professor) in English Linguistics

Department of English – Faculty of Humanities

University of Iceland

IS-101 Reykjavík, Iceland.

whelpton@hi.is