Corpus-based Conceptual Characterisation of Verbal Predicate Structures

Wim Peters
Wim.Peters@ccl.kuleuven.ac.be

Abstract

This paper describes performed research into the feasibility of creating computer-aided methods for knowledge extraction from an English corpus with the aid of existing lexical resources and analysis tools. Starting with subcategorised surface syntactic structures which give valuable information about the semantic and cognitive aspects of a (sub)language, the analysis is transferred to the abstract representation level of deep cases or thematic roles. These reveal information about the conceptual relations between the verbal arguments and the predicate that exist in the semantic domain under consideration. A semantic characterisation of the theta roles themselves will be supplied by using the lexical information of WordNet (Miller and others, 1990).

1 Introduction

The automatic analysis and interpretation of computer-readable texts is a very knowledge-intensive process. The acquisition bottleneck and scarcity of detailed large-scale lexical resources has led to research into the feasibility of obtaining information relating to various levels of linguistic and conceptual analysis from electronic texts. In a number of text interpretation systems semantic templates (Schütz, 1994) or knowledge frames (Sorenson, 1990; Zernik and Jacobs, 1990)) are activated and assigned to certain portions of texts. This must happen accurately and forms a critical part of data extraction (Jacobs and Rau, 1993).

The methods involved vary from purely statistically based information retrieval techniques to detailed linguistic and conceptual preprocessing.

In this paper the emphasis lies on the latter end of the scale. Preprocessing depends heavily on the coverage of the lexicon and the granularity of the lexical information contained in it. Techniques providing verbal case frame templates and case-based acquisition of lexical information control interpretation in combination with the lexicon. In this paper the possibility is investigated of automatically acquiring verbal case frames and semantic constraints on the roles involved.

2 The Source Corpus

The English corpus used for analysis consisted of approximately one million words of morphosyntactically annotated text within the domain of satellite communications (Project, 1993). The morphosyntactic annotation had been added by means of the textscclaws tagger (Leech, Garside, and Sampson, 1987). Furthermore, noun phrase bracketing was applied and for each NP the head noun was isolated. An example is shown below, in which noun phrases are marked by brackets (see Leech, Garside, and Sampson (1987) for tag meanings).

```
[ they_PPHS2 ] receive_VV0 [ transmissions_NN2 ] from_II
[ the_AT Earth_NP1]
```

For this feasibility study four high frequency verbs were chosen from the corpus. They are listed in Table 1.

Verb	Frequency of occurrence
"receive"	2740
"transmit"	1107
"connect"	737
"operate"	599

Table 1: Verb frequency

3 Thematic Roles

The general phenomenon of thematic roles goes under a number of different terms in the linguistic literature, including thematic relations, participant roles, deep cases, semantic cases and theta roles. Though many of these terminological distinctions carry with them particular views of their theoretical status, there is a shared underlying intuition that certain characteristic "modes of participation" in an event can be identified and generalized across a variety of verb or sentence meanings.

Thematic roles have a mediating function in the process of language understanding, where the parser, the discourse model and representation of world knowledge join forces (Tanenhaus and Carlson, 1989). When a verb is encountered, the roles associated with its subcategorised elements become available, and are instantiated by chosen text elements. If the filler is implausible in all these roles, the sentence becomes incongruous. If the thematic slot filling succeeds the text unit receives an interpretation.

Theta-marking (Jackendoff, 1991) establishes a correspondence between syntactic arguments or adjuncts and argument positions in the conceptual structure of a verb. In other words, it involves a mapping procedure between surface syntactic structures and semantic functional primitives like AGENT or DIRECTION (Fillmore, 1968; Steiner, Schmidt, and Zelinsky-Wibbelt, 1988; Halliday, 1985)). In the case of prepositional phrases the functional mapping involves primarily spatial concepts (Bennet, 1975; Rauh, 1993). This intrinsic spatial orientation is expressed by various groupings of prepositions, and can be related to thematic roles like LOCATION, DIRECTION, SOURCE and PATH. Table 2 shows this mapping procedure:

NP	$\longmapsto$	THEME
on/in/at/within	$\longmapsto$	LOCATIVE
over/through	$\longmapsto$	PATH
to/into	$\longmapsto$	DIRECTION/GOAL
from	$\longmapsto$	SOURCE
with	$\longmapsto$	INSTRUMENT/ CONCOMITANT

Table 2: Mapping of PP's onto conceptual roles

Significant co-occurrences of verbs with prepositions such as "through", "at", "over", "with" suggest that the thematic roles involved are typically expressed in the environment of the verb in question, or more specifically, are typical slot fillers within this verb's linguistic context.

In this study, the right context of verbs was analysed as sequences of CLAWS tags. The size of this context was limited to 2–4 elements. NP's were analysed as one element. Of course, the idea of a one-to-one mapping between prepositions and a thematic role is too simplistic a view on the process of interpretation. Prepositions may head constituents that function as prepositional object, which leaves the preposition semantically empty. For instance, in the case of "operate" the literal interpretation of a PP headed by "on" as a LOCATIVE yields wrong information in the case of a sentence such as "The satellite transmitter operates on a separate electric battery".

In this study only those prepositions have been taken into account that, given the technical domain of the text, can be regarded as fairly unambiguous. Only for the hybrid INSTRUMENT/CONCOMITANT role (see below) ambiguity has been allowed, because this can quite easily be resolved by judging the animacy of the entities involved. The position held is that by examining the right contexts of verbs it is possible to get at least a partial semantic profile of verbal predicate structures in terms of syntactico-semantic functions derived from surface syntactic collocates. Table 3 and Table 4 list the extracted theta grids for the verbs "transmit" and "operate" with an indication of the frequency of the found instantiations relative to the frequency of the verbs listed in Table 1.

Table 5 lists the theta grids of all verbs under examination. An important "transmit"

theta role	frequency
THEME	197
LOCATIVE	227
PATH	157
DIRECTION/GOAL	29
INSTRUMENT/CONCOMITANT	43
SOURCE	29

Table 3: Theta grid "transmit"

"operate"

theta role	frequency
THEME	25
LOCATIVE	203
INSTRUMENT/CONCOMITANT	51

Table 4: Theta grid "operate"

	"receive"	"transmit"	"connect"	"operate"
THEME	X	X	X	X
LOCATIVE	X	X		X
PATH		X	X	
DIRECTION/GOAL	X	X	X	
INSTRUMENT/ CONCOMITANT			X	X
SOURCE	X	X		

Table 5: Theta grids of all four verbs

empirical observation is that some roles that are not necessarily intuitively related to the verb (like the LOCATIVE function in the case of "transmit") appear to be rather important in the subdomain of satellite communications. This reflects the fact that the number and diversity of syntactic structures (and the corresponding number of verb readings) is always limited by the subdomain of which the corpus is a representative sample.

4 Semantic Characterisation

In the next stage of the process the head nouns of the phrasal constructions that reflect case frames were isolated in order to obtain a semantic characterisation of the objects entering into case relations with the predicates. A semantic clustering methodology was applied which involves the use of the semantic database WordNet (Miller and others, 1990). Hypernymic semantic information of headwords of NP's and PP's that were judged to reflect case relations was extracted from WordNet, which covered almost 100% of the extracted nouns. Abbreviations, although of considerable conceptual and terminogical importance, were not taken into account. The resulting lists of WordNet labels denote for each thematic role slot the broadest semantic classes to which the head nouns belong. The lists were then divided up by hand into ten broad semantic domains which cover 92% of the extracted WordNet data.

1. **Communication**
   relation, social relation, communication, written communication, message, spoken language, writings, written material, auditory communication, oral communication, interaction

2. **Quantity**
   quantity, definite quantity, amount, quantum, linear measure, unit

3. **Inanimate Object**
   object, physical object, inanimate object, artifact, instrumentality, article, thing, structure, natural object

4. **Concept, Knowledge**
   cognition, psychological feature, content, cognitive content, mental object, conception, concept, knowledge, idea, thought, subject matter, substance, information, higher cognitive process, intellection, thinking

5. **Activity**
   act, activity, human activity, human action, social activity, group action, work

6. **Group, Organisation**
   people, folk, social group,organisation, grouping, group
7. **Property**
   attribute, property, shape, form
8. **Location**
   location, point, workplace, geographic point
9. **Agent**
   causal agency, mortal, man, someone, being, life form, cause, human
10. **Event, Change**
    event, natural event, happening, change,motion, movement

When we compare these semantic domains with the 25 unique beginners mentioned in (Miller and others, 1990) which constitute the top nodes of Wordnet's ontology, we notice that some have been left out because they are too general, such as **Abstraction** and **Entity**. Others simply do not appear to have any instantiations in the text because of the specific lexical properties of the technical sublanguage which forms the corpus domain, e.g. **Animal**, **Fauna**, **Plant**, **Flora**, **Food**, and **Feeling**, **Emotion**. Other top hypernyms have been brought together into one class, such as **Natural object** and **Artifact**.

For every thematic role the distribution of the WordNet labels over the 10 Classes was computed by hand (see Table 6). A frequency threshold level for inclusion of 0.5% was used in order to weed the items with very low frequency.

5 Discussion

When examining the distributional characteristics of each thematic role some patterns immediately catch the eye. For instance, in the theta grid associated with "receive", Class 3 covers slightly more than 50% of the THEME instantiations with this verb, 45% of the LOCATIVE, 50% of the GOAL and 49% of the total number of SOURCE instantiations. These figures represent a strong semantic constraint on the possible fillers of this thematic slot. Overall, inanimate objects and, to a lesser degree, activities, communication and concepts play a large role in the realm of satellite communications. As for the SOURCE and GOAL relations, agents seem to have particular importance for the verbs "receive" and "transmit", whereas location is a significant semantic characterisation of the GOAL relation for, again, "transmit" and "connect".

It is obvious that, in order to be able to distinguish valid thematic roles by means of the method used, detailed linguistic knowledge is presupposed "receive" — 1540 WordNet labels in total

	1	2	3	4	5	6	7	8	9	10
Theme	60		396	68	235					
Locative	28	16	87	18	39			20		
Direction/Goal	17		48	16	15					
Source	44		207		44			13	149	

"transmit" — 777 WordNet labels in total

	1	2	3	4	5	6	7	8	9	10
Theme	44		166	10	36	20	10			
Locative	23		116							
Path	14		227	12	15			4		
Direction/Goal	20		147			18		56	104	
Source			14			10			10	

"connect" — 1255 WordNet labels in total

	1	2	3	4	5	6	7	8	9	10
Theme	60		257		102					
Path	9	18	17	30			10			
Direction/Goal	54		496		111			28		
Instrument/ Concomitant	11		70	108	8					

"operate" — 807 WordNet labels in total

	1	2	3	4	5	6	7	8	9	10
Theme		8	181	28	24					
Locative	20		169	59	36	22	13	18		
Instrument/ Concomitant	23		157		48					

Table 6: Distribution of WordNet entities over semantic classes

in the observer, and a general a priori idea about the semantics of the verbs in question. Regarding a right context NP as a direct object is directly linked to the linguistic knowledge of the observer about transitivity. The computer cannot do the task on its own. The human analyst is still the key figure in the whole process. It will be possible, however, after empirical examination of the sublanguage in question, to establish sublanguage dependent correspondences between syntactic patterns and theta grids. This will reduce the burden of manual labour for future analysis of texts belonging to the same sublanguage, and offer good possibilities for formulating generalisation hypotheses on the mapping of syntax and semantics over more than one sublanguage (e.g. texts in a general technological domain).

The methodological path followed consists of a number of steps of which each confronts us with drawbacks. The described method is too coarse-grained to cover all occurring surface-syntactic predicate patterns. With a right context of 2–4 tag elements, there will surely be some loss of information. Also, the theta grids themselves are often incomplete, and the theoretical validity of the assigned roles controversial. Even if all circumstances were ideal, i.e. if patterns were unambiguous and complete, and the validity of theta grids universally agreed upon, the methods involved could not have yielded more information than is actually there in the source corpus. Any computational lexicon derived exclusively by extraction procedures is going to be limited in its ability to cope with productive usages of language. Therefore extraction techniques are best suited for sublanguage analysis and description within well-defined restricted domains such as satellite communications. In sublanguage, as opposed to general language, lexical selection is syntactified (Frawley, 1988). Textual patternings in sublanguage corpora reflect very closely the structuring of the sublanguage's associated conceptual domain. Collocation or the regular co-occurrence of items reveals the cognitive content of the text. These items may be separated by relatively small numbers of words (Phillips, 1985), which makes a 2–4 word context a reasonable option.

With respect to the extraction of semantic information only one instance of each head noun encountered in the right context has been fed through WordNet. This means that the frequency of the corpus words has not been taken into account while establishing the importance of the semantic classes for each theta role. A comparison of the two distributional sets of WordNet labels for the Direction/Goal role associated with "transmit", one where word frequency has not been taken into account, and one where it has, yields no extra significant class, only an increase of Class 9. Table 7 below shows the differences in distribution. The general conclusion I venture to draw here is that using unique instances of right context elements does not affect the semantic characterisation of this role in terms of distributional patterns over semantic classes.

Another conceivable problem is that many of these nouns have multiple

	1	2	3	4	5	6	7	8	9	10
GOAL +freq	89		260	24	10	28		65	394	
GOAL −freq	36		190			18		56	104	

Table 7: Word-frequency based and non-word-frequency based differences in distribution of WordNet entities over semantic classes

readings in WordNet, being a general language semantic database. Instead of choosing the appropriate WordNet reading in each instance, all readings have been taken together, on the assumption of an effective dispersal of inappropriate senses. This is in line with Resnik (1993, pp. 28–29). The most frequent WordNet hypernyms yield the relevant semantic classes selected by the verbs.

Concluding, the results do seem to indicate some significant clustering in one ore more semantic classes, and suggest that automatic acquisition of coarse semantic constraints on fillers of verbal case frame slots is possible using existing resources. Information on semantic subcategorisation preference will facilitate the automatic filling of the roles or slots of these templates, and thus guide the process of text understanding.

References

Bennet, D. C. 1975. *Spatial and Temporal Uses of English Prepositions. An Essay in Stratificational Semantics.* Longman, London.

Fillmore, Charles J. 1968. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory.* New York.

Frawley, William. 1988. Relational models and metascience. In M. Evens, editor, *Relational Models of the Lexicon.* C.U.P., Cambridge.

Halliday, M. A. K. 1985. *Introduction to Functional Grammar.* Edward Arnold, London.

Jackendoff, R. 1991. *Semantic Structures.* Number 18 in Current Studies in Linguistics. MIT Press, Cambridge/London, second edition.

Jacobs, P. S. and L. F. Rau. 1993. Innovations in text interpretation. In Fernando Pereira and Barbara Grosz, editors, *Natural Language Processing.* MIT Press, London.

Leech, G., R. Garside, and J. Sampson. 1987. *The Computational Analysis of English. A Corpus-Based Approach.* Longman, London.

Miller, George. A. et al. 1990. Five papers on wordnet. CSL report, Princeton University.

Phillips, M. 1985. *Aspects of Text Structure: an Investigation of the Lexical Organisation of Text.* Elsevier, Amsterdam.

Project, ET10/63. 1993. Probabilistic and corpus-based methods in EUROTRA: Terminology, lexicon and preference. Final report, Paris.

Rauh, G. 1993. On the grammar of lexical and non-lexical prepositions in english. In C. Zelinsky-Wibbelt, editor, *The Semantics of Prepositions.* Mouton de Gruyter, Berlin/New York.

Resnik, P. S. 1993. *Selection and Information: A Class-Based Approach to Lexical Relations.* Ph.D. thesis, University of Pennsylvania.

Schütz, J. 1994. *Terminological Knowledge in Multilingual Language Processing.* Number 5 in Studies in Machine Translation and Natural Language Processing. E. C., Luxembourg.

Sorenson, H. S. 1990. The use of knowledge-based frames for terms in eurotra. Proceedings of TKE, Frankfurt. Indeks Verlag.

Steiner, E., P. Schmidt, and C. Zelinsky-Wibbelt, editors. 1988. *From Syntax to Semantics. Insights from Machine Translation.* Pinter Publishers, London.

Tanenhaus, M. K. and G. N. Carlson. 1989. Lexical structure and language comprehension. In W. Marlsen-Wilson, editor, *Lexical Representation and Process.* MIT Press, London.

Zernik, U. and P. Jacobs. 1990. Tagging for learning: Collecting thematic roles from corpus. volume 1 of *Proceedings of COLING*, pages 34–39.