Natural Language Engineering **3** (4): 279–315 Printed in the United Kingdom

© 1997 Cambridge University Press

Systematic construction of a versatile case system

KEN BARKER, TERRY COPECK AND STAN SZPAKOWICZ

School of Information Technology and Engineering, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5*
e-mail: {kbarker, terry, szpak}@csi.uottawa.ca

SYLVAIN DELISLE

Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada, G9A 5H7
e-mail: Sylvain_Delisle@uqtr.uquebec.ca

(*Received 9 October 1996; Revised 24 January 1997*)

Abstract

Case systems abound in natural language processing. Almost any attempt to recognize and uniformly represent relationships within a clause – a unit at the centre of any linguistic system that goes beyond word level statistics – must be based on semantic roles drawn from a small, closed set. The set of roles describing relationships between a verb and its arguments within a clause is a case system. What is required of such a case system? How does a natural language practitioner build a system that is complete and detailed yet practical and natural? This paper chronicles the construction of a case system from its origin in English marker words to its successful application in the analysis of English text.

1 Motivation

1.1 Case systems

Cases are generalizations of the semantic relationships between an activity (a process or event) or state and the participants in and circumstances around the activity or state.[1] The activity or state is usually represented by a verb and the participants and circumstances by the verb's syntactic arguments. Natural language processing systems that use cases reduce much of the meaning of a sentence to instances of case relationships.[2] Despite this reliance on cases, researchers building such systems seldom devote much effort to developing a set of cases. Often the use of cases is not even explicitly acknowledged. Consider the description from Grishman (1995) of one of the tasks in the popular MUC competition: 'The template-filling task for MUC-

* Formerly of the Department of Computer Science.
[1] Without loss of generality, we will refer to the activity or state as the *act* and the participants and circumstances as *arguments of the act*.
[2] There are many more systems that use semantic representations other than cases, but any discussion of these systems and their underlying principles it beyond the scope of this paper.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

6 involves the extraction of information about a specified class of events and the filling of a template for each instance of such an event.' The templates consist of a set of predefined 'template elements' for people, organizations and artifacts involved in the event, i.e. the participants in the event. For example, a *purchasing* activity template consists of roles for the *seller*, *buyer* and *object purchased*. These template elements comprise an *application specific* case system, but a case system nonetheless. Even papers on systems that make explicit use of semantic cases (such as Oflazer and Yilmaz, 1996) rarely list a complete set of cases and even less often offer any justification for them.

Theoretical linguists have devoted more attention to the study of case systems (for example, Somers, 1987). However, these studies are usually driven by linguistic motivations such as language universals and psychological plausibility. The resulting case systems may not be useful to computer-based applications such as information extraction or computer aided knowledge acquisition from text.

In contrast with both these traditions, this paper describes the construction of a complete case system and offers practical justifications for the cases. Wilks *et al*. (1996) argue against validating a set of *semantic primitives* against some external set of entities (such as some unknown ideal set of psychological primitive concepts). They hold that there can be no direct, independent justification for individual primitives, only practical motivations for the inclusion of a given primitive in a set. In this spirit we eschew psychological arguments for inclusion of cases in our set in favour of practical considerations such as their adequacy in expressing and distinguishing different relationships, and their coverage of English texts.

The work reported in this paper is part of a larger project that deals with knowledge acquisition from technical text. The TANKA system (Text ANalysis for Knowledge Acquisition) aims to build a conceptual model of the domain described in a technical text in terms of relationships that can be recognized in surface linguistic phenomena. Restricting analysis to technical text requires some notion of what technical text is. Copeck *et al*. (1997) investigate the extent to which certain features of a text indicate its degree of technicality.

We propose to use as little *a priori* semantic knowledge as possible, and to offset the absence of such knowledge by including a human in the process, and by simple learning. The system first produces a detailed surface syntactic analysis of a text fragment, typically a clause. Next, it draws on its processing history and on publicly available lexical resources to produce a tentative semantic analysis which is then assessed by a participating user. Most often the user merely approves an analysis suggested by the system or requests another analysis. See Delisle *et al*. (1996) for a detailed description of TANKA.

Many decisions during the construction of the set of cases have been informed by the goals of the TANKA project. In particular, the decision to use cases as the semantic representation of choice is motivated by the close fit of cases to the surface syntactic form of clauses. Nevertheless, our case system should be useful in other NLP applications. We describe how TANKA's case system was built and defend it by referring to the practical criteria that guided its construction and to its coverage of real English texts.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

For the computational linguist this work extends and experimentally validates recent research on cases. For the natural language practitioner it provides a well tested set of semantic roles built from English lexical data.

1.2 Why case?

An advantage of using cases for engineering approaches to language processing is that there is a straightforward mapping from surface syntax to cases. Given a syntactic analysis of a text it is possible to arrive at a semantic analysis with little background knowledge and minimal user effort. However, the result is useful only if there are applications that can take advantage of the relative facility of such a mapping. There are several:

- *Information extraction* from declarative text involves the identification of essential participants and circumstances in acts. The participants and circumstances can be represented directly as cases.
- *Template filling* is one of the goals of information extraction. A template has a fixed set of slots to be filled with specific concepts from the text. For acts, these slots often have names quite like case names (such as Agent, Instrument, Location, etc.) and refer to the same concepts as cases.
- It is often claimed that case is a universal phenomenon across natural languages (Fillmore, 1968). Research has identified strikingly similar lists of cases for many human languages (see Campe (1994) for a multilingual bibliography of case). The case roles in a source language clause can be used as part of an interlingua in *machine translation*.
- *Question answering systems* are often faced with questions about the participants and circumstances of acts. Again, these often correspond to the cases in a proposition.
- The *construction of semantic lexicons* has become a popular area of research in natural language engineering. The verb entries in these lexicons often include information on the semantics of the expected arguments of the verb. For example, an entry for the verb *give* might include not just the information that *give* requires a subject and two objects, but that one of the syntactic objects is 'the thing given' (commonly called the Object or Patient) and the other is the Recipient. These features in verb entries also correspond to cases.

2 Background

2.1 Conventions and terminology

This section defines some terms already used and others that will be used in later sections. We also describe the typographical conventions of the paper. A single example at the end of this section illustrates both terms and conventions.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

2.1.1 Conventions

In any paper related to case theory there is a risk of ambiguity because the word *case* has so any different meanings. In this paper the word will always refer to its meaning in case theory, as defined below – we use no other sense. The first letter of names of individual cases (such as Agent) will always be capitalized. Abbreviations for cases appear in all capital letters (e.g. AGT). The individual case abbreviations are given with their definitions in section 4.4.

We will usually abbreviate parts of speech and common syntactic constituents. The abbreviations are standard or intuitive and include the following:

subject	→subj	preposition	→prep
direct object	→dobj	noun phrase	→np
indirect object	→iobj	prepositional phrase	→pp
complement	→compl	conjunction	→conj
adverbial	→adv		

When a part of speech or syntactic constituent and its corresponding case appear together, they are separated by a slash, e.g. subj/AGT.

It is often helpful to identify syntactic constituents or parts of speech in example sentences. We use the familiar bracket-subscript notation (see example sentence (1)) for both purposes, identifying only those syntactic phenomena relevant to the example. The bracket-subscript notation will also occasionally identify cases.

2.1.2 Terminology

Without claiming that the following definitions are universally accepted in case theory research, we will describe three common terms as they are used in this paper.

A *case* is a semantic relationship between the concept represented by a verb and the concept represented by some other constituent syntactically subordinate to the verb. Examples are the initiator of an act, the time at which an act takes place, and the entity experiencing a state.

A *case system* is a set of cases together with rules for their interpretation. For thorough discussions of the linguistic motivation for case theory and surveys of several case systems, see Fillmore (1968), Bruce (1975), Somers (1987) and Cook (1989).

Case Analysis (*CA*) is the task of identifying which syntactic constituents in a given clause represent cases, and what case each of these constituents represents.

The following four definitions are specific to case analysis in our own knowledge acquisition project. These terms will be used extensively when we describe the experimental validation of our case system.

A *case marker* (*CM*) is a lexical or syntactic element such as a preposition or direct object that indicates the presence of a case.

A *case filler* (*CF*) is an argument of an act that is involved in a case relationship with the act. The surface form of a case filler is often a noun phrase in the syntactic subject or object position of a verb, or the complement of a preposition.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

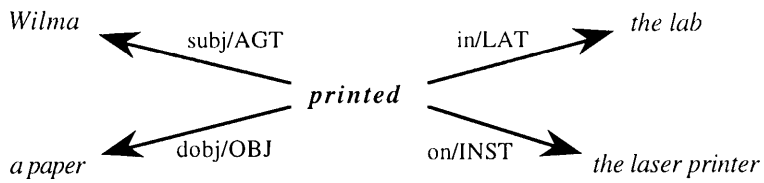


Fig. 1. Case analysis of example (1).

A *case marker pattern* (*CMP*) is the set of case markers found in a given clause.

A *case pattern* (*CP*) is the set of cases corresponding one-to-one with the case markers in a given clause.

2.1.3 Example

Four cases appear in example (1). There are two positional case markers (subject and direct object) and two prepositional markers (*in* and *on* within the two prepositional phrases (pp) postmodifying the verb *printed*). The CMP is therefore subj-dobj-in-on. The CP is Agent-Object-LocationAt-Instrument. The four CFs are *Wilma*, *a paper*, *the lab* and *the laser printer*. Fig. 1 is a graphic representation of the case analysis of (1).

(1) *[Wilma* $_{subj/AGT}$*] printed [a paper* $_{dobj/OBJ}$*] [in the lab* $_{pp/LAT}$*] [on the laser printer* $_{pp/INST}$*].*

2.2 A case theory primer

2.2.1 Case theory

Case theory (Fillmore, 1968; Somers, 1987) focuses on the simple clause and on the verb within it. Cases capture that part of the meaning of a constituent conveyed by its relationship with the verb. The occurrence of a case in a text is signaled by a case marker. These lexico-syntactic clues are either lexical (for example, a preposition that introduces a prepositional phrase) or positional (subject, direct object, indirect object). Traditional case theory requires that there be exactly one semantic case assigned to each syntactic argument in a proposition[3] (Fillmore, 1968).

Since case theory only deals with the representation of semantic relationships between an act and its arguments, the representation of other semantic relationships will require other mechanisms. In particular, the semantics of relationships between clauses and within complex noun phrases fall outside the realm of case theory. Any semantic analysis system based only on case theory will fail to capture knowledge

[3] The one-to-one assignment of cases to markers seems to be violated for verbs like *sell*, which has an overt Agent case and a covert Source case filled by the same syntactic argument. However, such subtle semantics are possible for just about any proposition. For example, in the absence of an overt Beneficiary, the Agent can almost always also be considered Beneficiary (as in *I shoveled the driveway*). Cook (1989) identifies three varieties of covert roles. Since covert roles are not marked syntactically case assignment in TANKA is limited to exactly one case per marker.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

expressed by other syntactic constructs. Section 5.1.2 discusses semantic analysis modules within the TANKA project that handle other structures.

It has been argued that case theory is nothing more than a notational device. However, the assignment of a semantic label (such as a case) from a finite set is a useful abstraction. Information Extraction and Question Answering are two examples of applications that benefit from the identification of the different roles played by the objects participating in an act. Since it is not the purpose of this paper to defend the use of case systems as a semantic representation, we refer the reader to Palmer (1990) and Wilks *et al.* (1996) for overviews of the debate.

2.2.2 Valency theory

Valency theory (Somers, 1987) deals with the types and number of verb arguments in a clause. The number of required verb arguments depends on the type of complementation of the verb. For example, a verb such as *give* is ditransitive and requires a subject and two objects. Cases assigned to required argument positions of a particular verb are considered *core* roles for that verb, while cases assigned to optional verb argument positions (such as adverbials) are considered *peripheral*.

Some cases (such as Agent and Object) are more often assigned to required verb argument positions and are therefore more frequently core. Others (such as temporal or locative cases) are more often assigned to optional verb argument positions and are therefore more frequently peripheral. Nonetheless, it is problematic to classify a given case as absolutely core or absolutely peripheral. Consider example (2).

(2) *The meeting lasted [six hours* $_{dobj}$*].*

The case assigned to the direct object clearly must be a temporal case (such as Duration). Since the direct object is a required argument for the monotransitive use of the verb *last* the Duration case is core in this example. In contrast, example (3) shows a peripheral use of the same case, since it is assigned to an optional prepositional phrase argument.

(3) *I read the minutes [for six hours* $_{pp}$*].*

Despite the difficulties, it is quite common in case theory to find systems containing only core roles. These systems sometimes contain a locative case and rarely contain a temporal case. To account for sentences such as (2), the temporal nature of the direct object is considered 'incorporated' in the verb. Alternatively, a verb such as *last* may be considered intransitive with a temporal preposition (such as *for*) lexicalized in the verb (see Cook, 1989).

For the purposes of text analysis, the identification of which cases are core for a given verb sense would be a large knowledge engineering task and no existing dictionary would offer help. This information is of no use in our broader project since we are concerned with collecting actual patterns of cases as they appear in a text. However, even if it were of use to us we would not undertake this classification since it would contradict the main tenet of our work: avoid precoding semantic knowledge if at all possible.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

2.2.3 Case markers

A case marker is a syntactic indicator of a case instance in a sentence. We distinguish three types of case marker: positional, prepositional and adverbial. The first type mark a case by their position in the sentence. They are the syntactic subject, direct object and indirect object of the verb. These three syntactic positions often mark cases such as Agent, Beneficiary, Experiencer, Object and Recipient. The constituents in these positions fill the cases marked by the position itself.

A prepositional phrase consists of a preposition and a noun phrase complement. Prepositions express a relationship between their complement and some other constituent in a sentence (see Quirk *et al.*, 1985, section 9.1). A preposition expressing a relationship between its complement and a verb marks a case. We say the prepositional phrase is attached to the verb and the prepositional complement is the case filler. A preposition expressing a relationship between its complement and some other noun phrase (as opposed to a verb) does not mark a case. The complement is instead considered a modifier of the other noun phrase and is not analyzed as a case even though it may express a case-like semantic relationship (see section 6.3).

In example (4), the prepositional phrase *on the printer* is attached to the verb *printed*. The noun phrase *the printer* fills that verb's Instrument case. In example (5), the prepositional phrase *on the printer on his desk* is also attached to the verb *printed*. The noun phrase *the printer on his desk* fills the Instrument case. However, the prepositional phrase *on his desk* postmodifies the noun *printer* and does not fill a case of the verb *printed*.

(4) *Bonnie printed the paper [on [the printer* $_{np}$*] *$_{pp}$*].*

(5) *Clyde printed the paper [on [the printer [on his desk* $_{pp}$*] *$_{np}$*] *$_{pp}$*].*

Adverbials are the third type of case marker. The exact same semantic relationships expressed by prepositional markers can also often be marked by adverbials. In example (6), *at* marks and *the same time* fills the TimeAt case; in the paraphrase in example (7), *simultaneously* expresses the same case.

(6) *The two events occurred [at the same time* $_{pp}$*].*

(7) *The two events occurred [simultaneously* $_{adv}$*].*

Based on the similar behaviour of prepositional phrases and adverbials in denoting the circumstances of acts, we consider adverbials case markers. An adverbial case marker both marks a case and fills it.

Because certain prepositions and conjunctions are homographic, conjunctions may seem to mark cases. In example (8), the conjunction *after* appears to mark TimeAt.

(8) *Fred came [after* $_{conj}$*] Barney left.*

However, conjunctions joining clauses do not mark cases because they do not signal a relationship between a single act and one of its arguments. Instead, they mark a relationship between two acts. This relationship would have to be handled by a mechanism to analyze relationships between clauses (see Barker and Szpakowicz, 1995).

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

3 Related work

There is a long history of work in case theory and research related to cases from many different areas, including linguistics, computer science and cognitive science. Campe (1994) offers a bibliography of over 6600 references related to case. This section describes selected references relevant to the project described in this paper.

Fillmore (1968) presents a small set of cases that has become the kernel of many subsequent case systems. His cases can be thought of as a set of semantic relationships underlying surface level case, which is expressed by morphology, word order, etc. Fillmore's original list consisted of Agentive, Instrumental, Dative, Factitive, Locative and Objective cases.

Bruce (1975) explores the relationship between surface case (syntactic level) and deep case (semantic level). He also attempts to define a single subjective rule for deciding which cases are 'central to an event description'. The remainder of the paper is a survey of several theoretical case systems and several practical case systems in use in various applications. Bruce notes the domain dependence of many of these systems and identifies the problem of how to measure the 'goodness' of a case system, leaving possible solutions to future research.

Larson (1984) presents a list of cases to be used for the intermediate representation in a machine translation system. The list consists of twelve cases: Agent, Instrument, Beneficiary, Affected, Resultant, Goal, Location, Time, Accompaniment, Causer, Manner and Measure.

The theory of conceptual graphs (Sowa, 1984) uses a set of 'conceptual relations' as semantic links between concepts in a graph. Conceptual relations linking acts with other concepts are borrowed directly from the case relations in case theory: 'Arcs of the graphs correspond to the *function words* and *case relations* of natural language' (Sowa, 1984, p. 20). Sowa provides a complete list of conceptual relations along with maximally general concepts they may link (according to a concept hierarchy). The list also includes example sentences for each relation.

Somers (1987) traces the history of case theory from its roots in valency theory through Fillmore and others to arrive at his own set of case roles. Somers' set of twenty-four cases takes the form of a four column by six row grid. The four columns correspond to Source, Path, Goal and Local dimensions. The six rows are the categories Active, Objective, Dative, Locative, Temporal and Ambient. Cases appear at the intersection of each row and column. This organization is an attempt to provide a principled basis for the construction of a case system, as opposed to the arbitrariness of systems based on intuition. However, the requirement that each category of cases map exactly onto the four dimensions seems at times artificial. Whether there is cognitive support for the organization of cases into such a grid is unclear.

Sparck Jones and Boguraev (1987) describe an attempt at a principled, data-driven approach to defining a set of cases. After deciding that their target set should consist of 20 to 30 cases, they studied a substantial list of example sentences illustrating the different senses of the English prepositions. They arrived at their final set of twenty-eight cases by mapping the preposition senses to candidate cases. The resulting cases

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

are defined only in terms of the example sentences from which they were derived. Without more formal definitions the distinction between similar cases is not always obvious. Sparck Jones and Boguraev also ignored the different syntactic functions of prepositions. Since prepositions do not always mark relationships between acts and arguments, their set of cases includes several non-case semantic relationships. For example, the Possessed-By relationship can only link nominals; the State relationship is included for predicate adjectives. Finally, there is no attempt at justifying the level of generality of each case: some of the cases are very specific (such as Force) while there is a single undiscriminated Location case to account for all locative relationships.

Cook (1989) gives an overview of several different models of case theory including his own Matrix Model. His model distinguishes 'propositional' cases (core) from 'modal' cases (peripheral). Propositional cases are those essential to a verb's meaning; modal cases are optional adjuncts of any predication. Cook's work deals primarily with the propositional cases, though he recognizes the utility of modal cases to text analysis. The matrix model contains five cases: Agent, Experiencer, Benefactive, Object and Locative. A maximum of three of these cases may appear in the case frame for a given verb. Object must appear in all case frames and may appear more than once (for stative verbs and object complements, for example). The other cases can appear at most once. Cook (1989) has several contributions relevant to this paper. He gives criteria for judging the completeness of a case system. These criteria include comparison to existing case systems and extensive use of the cases in text analysis. Cook also provides a list of modal cases that should appear in any case system that includes modal cases: Time, Manner, Instrument, Cause, Result, Purpose, outer Locative and outer Benefactive.

Slator *et al.* (1990) derive semantic classes based on a study of the 25 most common English prepositions. 184 different senses of the prepositions are clustered into 46 classes based on a semi-automatically computed measure of similarity of their dictionary definitions. The classes are named according to human reflection on the shared semantic properties of the senses in each cluster. What is striking in the results is the similarity between these empirically derived classes and the traditional cases found in more arbitrary, human built case systems. The classes that do not map straightforwardly onto common case roles often express non-case uses of the prepositions. These classes include relationships such as Attribute and Comparison. The relatively large number of classes can be attributed to these non-case classes and the small number of preposition senses per class (on average). The classes are correspondingly fine-grained, with as many as 15 possibly locative classes. Nonetheless, the similarity between these purely empirically derived classes and case roles common to other systems lends support to the kinds of semantic relationships arrived at intuitively by case researchers.

Velardi *et al.* (1991) attack the problem of acquiring shallow semantic knowledge from text semi-automatically. Their system uses a set of mapping rules between 'syntactic collocates' and semantic relations. The syntactic collocates consist of pairs of syntactically connected constituents, such as subject-verb pairs, verb-object pairs, adjective-noun pairs and prepositional phrases postmodifying nouns. The output is

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

a set of concept-relation-concept triples, where concepts are generalizations of individual words based on a small, domain dependent, hand coded concept hierarchy. The list of potential relations for a given syntactic collocate is narrowed by selectional restrictions between general concepts in the hierarchy. Finally, a human user can accept or reject analyses for inclusion in a permanent knowledge base. A complete list of semantic relations is not given, though the output is in Conceptual Graph notation (Sowa, 1984) and appears to use common Conceptual Graph relations.

In work not directly related to the development of case systems, Levin (1993) argues that verbs with similar syntactic behaviour share meaning components. She groups together English verbs that behave similarly with respect to diathesis alternations and attempts to identify elements of meaning common to the verbs in each group. For example, verbs like *cut* often have an Agent, an Object and an Instrument. Different alternations of the verb *cut* allow any one of these cases to occupy the syntactic subject position. Levin attempts to identify the semantic features of the verb that permit certain alternations. Her success in this endeavour provides support for the notion that part of a verb's meaning is reflected in the semantic roles allowed to appear in each syntactic argument position. Levin's work also shows the tremendous variation possible in the distribution of case roles in required verb argument positions. This variation shows that the same case can often be assigned to either a required or optional verb argument position, further blurring the distinction between core and peripheral cases.

Van Valin (1993) offers two tiers of semantic roles: 'thematic relations' and 'macroroles'. Thematic relations correspond to the case roles Agent, Effector, Experiencer, Locative, Theme and Patient. The macroroles Actor and Undergoer serve to rank these thematic relations in the order listed above, with Agent having the highest degree of 'Actor-ness' and Patient having the highest degree of 'Undergoer-ness'. In the author's Role and Reference Grammar theory, it is this ranking along with verb specific preference rules that allows humans to assign thematic relations to verb arguments.

Wu (1993) restricts thematic role frames (case patterns) to a maximum of four thematic roles per frame, and relies on the nesting of role frames to account for verbs with more than four arguments. Nesting frames allows the system to assign only core roles in the top level frame while relegating peripheral roles to nested frames. According to Wu, a side effect of using nested structures is that 'surface cases do not map straightforwardly onto thematic roles in a one-to-one fashion.' (Wu, 1993, p. 327). In a system attempting to represent the semantics of a text in a form that closely matches its surface syntax, a one-to-one mapping between surface cases and semantic roles is highly desirable. Systems whose semantic representation does not map easily to surface syntax become much more dependent on the amount and form of precoded semantic background knowledge.

Pustejovsky and Busa (1995) present the details of the temporal template elements for the MUC competition. The four 'time objects' During, Before, After and On are equivalent to the Spark Jones and Boguraev (1987) temporal cases. Although Pustejovsky and Busa do not mention cases, they show a mapping from prepositional (and other lexical) markers to the four time objects. For example, *in*, *throughout*, *for*,

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

late, *early*, *beginning of* and *end of* mark the During object; the Before object is marked by *before*, *ending*, *until* and *by*.

4 Case system design

4.1 Requirements

In this section we identify three requirements that guided the construction of our case system. Section 5.3 will evaluate the extent to which our case system satisfies them.

4.1.1 Generality

In section 1, we said that cases were generalizations of the semantic relationships between an act and its participants and circumstances. A very simple case system could be composed of just two cases, Participant (P) and Circumstance (C).

(9) *[Clyde $_P$] broke [the window $_P$] [yesterday $_C$] with [a rock $_P$].*
(10) *[Bonnie $_P$] printed [a paper $_P$] in [the lab $_C$] on [the laser printer $_P$].*

This trivial case system would probably be too general for any natural language processing task since the designation of act arguments as participants or circumstances is no more informative than the syntactic analysis of verb arguments as objects or adverbials.

At the other extreme, individual verb-argument relationships themselves could be used as cases. This open ended case system would not be able to capture the fact that *a rock* and *the laser printer* were each instrumental to the execution of their respective acts in (9) and (10).

These two extremes suggest requirements for the construction of practical case systems. Cases should be *specific* enough to distinguish between the roles of each verb argument in a given clause. The Participant case was too general to account for the different roles played by the subjects, direct objects and adverbials in (9) and (10). On the other hand, cases should *generalize* the semantic relationships in a clause. That is, each case must be able to account for relationships between more than one distinct verb-argument pair. For example, a BreakWindow case would not account for the similarity of the roles played by the direct objects in (11), (12) and (13).

(11) *Alphonse broke [the window $_{dobj}$]*
(12) *Alphonse broke [the pane $_{dobj}$]*
(13) *Alphonse shattered [the window $_{dobj}$]*

Together the specificity and generality required of a case system will determine how many cases it contains and how closely definitions of the cases match the semantics of the specific verb-argument relationships they generalize. We call this the *generality requirement*. Choice of a particular point between the two extremes on the scale will depend on such things as the application using the case system, the target representation of case analysis and the type of text. For example, the Participant role that we discarded as too general may be specific enough if the application is a template filling task simply to determine the entities involved in an act. Or if the target

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

semantic representation to be produced from case analysis does not make subtle distinctions between semantic roles then a fine grained case set would be unnecessary. Certain types of texts may also dictate the required generality of a case system. The analysis of geographical texts, for example, may require very fine distinctions between a number of locative cases.

Commonly, the overall level of generality of cases is either not explicitly addressed by case system designers or set arbitrarily by choosing *a priori* a certain number of cases. There are ways, however, to measure the generality of an existing case set.

One measure is the number of times each case is assigned in the analysis of texts. Although this measure would be of great value in determining the practical generality of a set of cases, it would require that a large amount of text be analyzed to draw any reasonable conclusions. It also assumes that frequency of occurrence is directly related to generality. Although frequency may hint at which cases are more general than others, it is possible that certain quite generally defined cases are less common than some cases with very restricted, specific definitions. Furthermore, the measure would continually change as more texts were analyzed, though more analysis would also increase the confidence in the measure.

A second measure is the number of *syntactic* phenomena (i.e. markers) accounted for by each case. Although this measure tells us less about the applicability of each case to the analysis of actual texts, it does reflect an aspect of a case's generality: the generality of a case is bound from above by the frequency with which its markers occur in text. So a case with more markers or commonly occurring markers (such as the positional markers) will potentially account for more verb-argument relationships in a text than a case with fewer or less common markers. This second measure could also be calculated easily since the syntactic phenomena involved are usually fixed independent of any texts to be analyzed. Barker (1996) investigates the degree to which each of the cases presented in this paper is represented by markers. The report also discusses possible adjustments to the case set for cases that have above or below average marker representation. The main results from this investigation appear in section 5.3.

4.1.2 Completeness

The *completeness requirement* states that a set of cases should cover any possible relationship between an act and an argument. In practice a case system can never be shown to be perfectly complete; conversely, a single counterexample can show that it is incomplete. One measure of the degree to which a given case system is complete is the number of sentences in a text containing relationships covered by the case system. As more clauses are case analyzed within the confines of a given set of cases, confidence in the completeness of that set increases.

A potential pitfall of trying to satisfy the completeness requirement is the proliferation of new cases. If new cases are continually introduced to remedy incompleteness, their number may grow unchecked. However, generality ensures that overly specific cases are not added to cover a single verb-argument relationship. A solution is to elaborate the definition of an existing case to account for the new relationship. The new definition should cover all of the relationships covered by the

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

old definition along with the new relationship and no others. Documentation of each case's coverage is therefore particularly important (see section 4.4). Occasionally, a new case may still be needed.

4.1.3 Uniqueness

The *uniqueness requirement* states that there should be no superfluous or redundant cases. A case is superfluous if the relationship it describes never appears in any text between an act and argument. A case is redundant if every instance of the case is an instance of some other case in the set.

Empirically, a case can never be shown to be superfluous or redundant: the fact that a case has covered none of the encountered verb-argument pairs does not imply that it will never do so; the fact that the set of encountered relationships covered by a case is a subset of the set of relationships covered by some other case does not imply that the two sets will never diverge.

However, redundant and superfluous cases can be avoided by requiring that no case be added to the set arbitrarily. That is, inclusion in the case set must be restricted to cases that cover at least one verb-argument relationship not covered by any other case in the set. Redundancy can also be avoided by defining explicitly the types of semantic relationships covered by a case and by contrasting these with the definitions of other cases.

4.2 The design process

Case theory is not new. Numerous sets of cases have been proposed and criticized by linguists and computational linguists. Our first step in designing a case system was to look at sets of cases used by others, many of which were described in section 3. Using a rough union (see section 4.3.1) of others' case sets as our starting point gave us at least as much coverage as these other systems and also gave us a reasonable initial level of generality.

The second step was to compile a comprehensive list of case markers. This step served to identify any holes in the initial set or cases weakly represented by the markers, allowing us to move closer to satisfying the completeness and generality requirements. The third step was to define each case carefully. Finding examples and contrasting each case with similar cases helped improve uniqueness. The complete definitions also serve as a guide for anyone who wants to use the cases in case analysis. The use of the definitions to guide users of our own case analyzer is described in section 5.1.3.

4.3 Building the case system

4.3.1 Construction

The first version of our case set was built from the case systems described in section 3. Any case appearing in any other list was a candidate for inclusion in our own list. We included only one case for similar cases with different names from different systems. For example, we included a single Object case to account for cases variously named Affected, Object, Patient and ThingTransferred. We also ignored cases that

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

did not represent relationships between acts and their arguments, such as the Attribute and Possessed-By cases in Sparck Jones and Boguraev (1987).

We then compiled a comprehensive list of case markers from an exhaustive search of electronic and conventional dictionaries.[4] This dictionary search identified the preposition and adverb case markers. The positional case markers (subject, direct object and indirect object) were also included in the case marker list.

The case marker list included all prepositions, both single-word and multi-word (such as *instead of*). Each distinct sense of a preposition described in its conventional dictionary entry was examined to decide which if any of our cases were appropriate to that preposition sense. If a sense corresponded to a case, an entry mapping the preposition to that case was added to a case marker dictionary. If no case adequately captured a particular marker usage, the set of cases was reexamined to determine if a new case was needed or if the scope of an existing case should be redefined. At the end of the process any case that was weakly represented in the marker list was reexamined to see if it was redundant or overly specific, in an attempt to improve the generality and the uniqueness of the set.

The source dictionaries were also consulted to find adverbial cases markers, such as those in (14) and (15), marking Manner and Frequency, respectively.

(14) *This printer prints [quickly* $_{adv/MANR}$*].*

(15) *We backup the hard disk [daily* $_{adv/FREQ}$*].*

Adding mappings between all adverb case markers and cases would not be desirable since the class of adverbs is not closed. An adverb was only added to the case marker list if it had a sense corresponding to a case other than Manner, the most common case marked by adverbs. If one of these adverbs also marked Manner, a mapping between the adverb and Manner was included in the list. Any adverb not in the list is assumed to mark Manner only. Conversely, adverbs that do appear in the list are taken to mark Manner only if there is an explicit entry mapping them to Manner.

Every marker → case mapping in the marker list includes an example sentence (see Appendix II for a small sample of the entries for prepositions). In addition to providing the user with an illustration of a given marker → case pair, the example sentences helped us develop and tune the case set. An example sentence was sometimes a compelling argument to include a case. Conversely, cases for which an example sentence could not easily be invented were considered dubious. Another technique to tweak the case set was to try to construct an example sentence containing *two* similar cases. When it was not possible to compose a sentence containing both cases, we examined the cases to see if they could be merged into a single more general one.

[4] The dictionaries we used most in the construction of the case marker list were the *COBUILD English Language Dictionary* (Sinclair 1991), the unabridged *Random House Dictionary of the English Language* (Stein, 1983) and Grady Ward's *Moby* word lists. We frequently consulted *Longman's Dictionary of Contemporary English* (Summers 1987) and *A Dictionary of Modern English Usage* (Fowler, 1984) to confirm and clarify marker usages.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

4.3.2 Example

As noted above, the grain of our system was initially established by that of the existing systems from which cases were taken. We then used the English marker words to guide adjustments to the specificity of our system. If a single case was marked by several different markers with differing senses, the case was split into more specific cases. Similar cases marked by only one or two markers were merged into a single more general case. The temporal roles provide examples of both kinds of adjustment.

Many older case systems have only one or two temporal cases. More recent work distinguishes between multiple temporal cases, often four.[5] We added Frequency to the common TimeAt, Duration, Before and After, inspired by the prevalence of adverbial case markers for such a role: *hourly*, *daily*, *monthly*, etc. This is clearly a temporal role not captured by TimeAt, Duration, Before or After.

The Before and After cases were richly represented by several different markers, as evidenced also in Pustejovsky and Busa (1995). For example, Before is marked by the prepositions *before*, *by*, *ending*, *till*, *to*, *towards*, *until* and *up to*. With the exception of *before* and possibly *by*, these markers suggest a slightly more specific role describing the ending point of an act of duration. To capture this distinction we split the Before case into TimeBefore and TimeTo. Similarly, we split the After case into TimeAfter and TimeFrom, bringing the number of temporal cases up to seven.

However, with the introduction of TimeTo and TimeFrom cases, TimeBefore and TimeAfter became very weakly represented by the markers. Only *before* exclusively marked TimeBefore and only *after* exclusively marked TimeAfter. We deleted these two cases, relegating their roles to the more general TimeAt case.[6]

4.3.3 The cases

The resulting set of cases (Table 1) was then compared against those in the systems described in section 3 of this paper to ensure that it had at least as much coverage as those systems.

We present our cases in five groups. The PARTICIPANT group consists of cases whose entities are directly involved in the act. The CAUSALITY group includes relationships with events or entities enabling or opposing the act. Under SPACE and TIME we group relationships that place the act at an absolute or relative position in space or time. Finally, QUALITY groups the remaining cases that represent various other relationships between a verb and its arguments. These groups are more useful when talking about the cases than when using them. For example, the division into groups

[5] The choice of four temporal cases is almost certainly inspired by the localist approach, where case systems are often divided into source, path, goal and local.

[6] We have noted that the Before and After relationships are much more strongly represented at the inter-clausal level where the time of one event represented by a clause is often expressed relative to the time of another event represented by a second clause. This relationship has its own role in our set of clause-level relationships, the inter-clausal analog of the verb-argument cases. See Barker and Szpakowicz (1995).

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

Table 1. *List of cases* (*with abbreviations*)

PARTICIPANT			
Accompaniment	ACMP	Experiencer	EXPR
Agent	AGT	Instrument	INST
Beneficiary	BENF	Object	OBJ
Exclusion	EXCL	Recipient	RECP
CAUSALITY			
Cause	CAUS	Opposition	OPP
Effect	EFF	Purpose	PURP
SPACE			
Direction	DIR	LocationThrough	LTRU
LocationAt	LAT	LocationTo	LTO
LocationFrom	LFRM	Orientation	ORNT
TIME			
Frequency	FREQ	TimeThrough	TTRU
TimeAt	TAT	TimeTo	TTO
TimeFrom	TFRM		
QUALITY			
Content	CONT	Measure	MEAS
Manner	MANR	Order	ORD
Material	MATR		

shows clearly the parallel between each of the temporal cases and their spatial analogs. Moreover, since each group contains cases that share some property, clashes or overlaps between cases are much more likely to occur within groups. For example, it is more likely that Effect and Purpose will need to be distinguished explicitly than Purpose and LocationTo; TimeAt and TimeThrough are more likely redundant than Agent and Cause.

4.4 Case glossary

This section describes the meaning and coverage of each case. It is meant as a guide for anyone wanting to use the case system. It is also used as a guide by our own users, as described in section 5.1.3. Each description is followed by one or more example sentences chosen to reflect typical usage or to illustrate the scope of each case. The entries are subject to change – in particular, to elaboration – as new semantic phenomena are encountered in text and dealt with as described in section 4.1.2. When consulted in the context of a given input sentence these definitions and examples should help the user to assign cases consistently.

In practice, for any particular assignment the choice of case depends on the reader's interpretation of the sentence, especially for more abstract examples. Our own

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

semantic analysis system allows the user to invent new cases for more refined domain specific modeling.

4.4.1 Participant

Accompaniment (ACMP)

The Accompaniment case represents one or more entities accompanying another participant (usually the Agent, Experiencer or Object) involved in the act. (16), (17) and (18) show accompaniment of Agent, Experiencer and Object respectively.

(16) [I $_{AGT}$] eat supper with [my family $_{ACMP}$].
(17) [I $_{EXPR}$] live with [my family $_{ACMP}$].
(18) I eat [my peas $_{OBJ}$] with [honey $_{ACMP}$].

Agent (AGT)

The Agent case represents the initiator or performer of an act. An Agent is typically a sentient being or some entity treated as sentient to some degree within the domain. This case differs from the Experiencer case in that the Agent intentionally performs or actively participates in the action. The Agent is expressed by the syntactic subject of the verb or as the object of the preposition *by* in passive clauses.

(19) [The database manager $_{subj/AGT}$] retrieved the records.

Example (19) shows an Agent, *The database manager*. Although database managers are not usually thought of as sentient, in the domain of Computer Science they often perform complex tasks. In the absence of a truly sentient initiator, other entities (such as database managers) can be Agents. In (20) the Agent would be *The analyst* and *the database manager* fills the role of Instrument.

(20) [The analyst $_{AGT}$] retrieved the records with [the database manager $_{INST}$].

Beneficiary (BENF)

The Beneficiary case represents the entity that benefits from the situation resulting from the act. The situation may be to the Beneficiary's advantage or disadvantage, and it may be intentional or accidental. Typically the Beneficiary is an animate being or an organization. It may correspond to the syntactic indirect object of the verb if the indirect object is not the Recipient of the object of the act.

(21) I wrote [the girl $_{iobj/BENF}$] a reference letter [to prospective employers $_{pp/RECP}$].
(22) I wrote [the girl $_{iobj/BENF}$] [a reference letter [to prospective employers $_{pp}$] $_{dobj/OBJ}$].
(23) This year's rains produced a bumper crop for [the farmer $_{BENF}$].

Example (21) shows the indirect object (*the girl*) filling the Beneficiary case with *prospective employers* filling the Recipient case. In (22) there is no verb argument to fill the Recipient case. Nonetheless, since the writing was done to the advantage (or disadvantage) of the girl, Beneficiary is again an appropriate case.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

Example (23) illustrates the fact that the filler of the Beneficiary case need not be the *intended* beneficiary of the act. Clearly, rains did not intend to produce a bumper crop for the farmer's benefit.

Exclusion (EXCL)
The Exclusion case represents an entity not included in a group or not accompanying another entity or entities. It can also represent the entity that substitutes for another whose involvement in the act is expected, or whose lack of involvement is significant.

(24) *Everybody slept except [him* $_{EXCL}$*].*
(25) *John went instead of [Mary* $_{EXCL}$*].*

Experiencer (EXPR)
The Experiencer case represents the entity experiencing a state or a sensation. Unlike an Agent, an Experiencer does not intentionally perform or actively participate in an act. The Experiencer is typically a sentient being or some entity treated as sentient. It corresponds to the syntactic subject of the verb.

(26) *[Fred* $_{subj/EXPR}$*] is sleeping.*

Instrument (INST)
The Instrument case represents an entity that is applied or employed to accomplish an act. The Instrument for acts of transfer is often the medium of the transfer.

(27) *He broke the window with [a brick* $_{INST}$*].*
(28) *The system administrator notified the users via [email* $_{INST}$*].*

In (27) *a brick* is the entity applied to accomplish the act of breaking the window. In (28) *email* is the medium of transfer of notifying the users.

Object (OBJ)
The entity directly acted upon by the verb's action fills the Object case. The Object often corresponds to the syntactic direct object of the verb but may also be marked by syntactic subject or by a preposition.

(29) *Jim printed [the file* $_{dobj/OBJ}$*].*
(30) *[The window* $_{subj/OBJ}$*] broke.*
(31) *They stripped him [of his pride* $_{pp/OBJ}$*].*

Example (29) has the syntactic direct object (*the file*) filling the Object case; (30) shows the less common situation of the syntactic subject (*the window*) filling this case; (31) is an example of the rare occasion where the Object case is marked by the preposition *of*.[7] *His pride* is the entity that has been stripped while *him* is the location from which it was stripped.[8]

[7] We consider this instance of the Object case rare but not merely idiomatic. The existence of similar examples using different verbs supports the decision to treat the preposition *of* as a valid case marker for the Object case: *They deprived him* [*of his rights* $_{pp}$]; *They partake* [*of the bread* $_{pp}$].

[8] Note that this interpretation makes a person an abstract location of a feeling. An alternative to allowing both abstract and concrete locations fill the locative cases would be to refine each locative case into two separate cases distinguished by the feature {+|-}concrete.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

Recipient (RECP)
The Recipient case represents the entity that directly receives the object of an act. The Recipient must be distinguished from the closely related Beneficiary case. Whereas the Beneficiary benefits from the realization of an act, the Recipient takes possession of the act's object. Recipient frequently appears with acts describing dative relationships and often corresponds to the syntactic indirect object of the verb.

(32) *I sent [the prospective employers* $_{iobj/RECP}$*] a reference letter for her.*
(33) *I wrote [the girl* $_{iobj/BENF}$*] a reference letter [to prospective employers* $_{pp/RECP}$*].*

In (32) the indirect object of the verb fills the Recipient case. In (33) the indirect object fills the Beneficiary case while the Recipient case is marked by the preposition *to* (see the definition of Beneficiary above).

In sentences that contain both cases, Beneficiary is more often marked by the preposition *for* while Recipient is more often marked by the preposition *to*.

4.4.2 Causality

Cause (CAUS)
The Cause of an act is the situation or event that makes it take place. The Cause case may also represent an environment that allows an act to be performed or a state to exist.

(34) *He died of [thirst* $_{CAUS}$*].*
(35) *We acted on [his advice* $_{CAUS}$*].*

Effect (EFF)
The Effect case represents a state that is the outcome of an act or the result of a situation.

(36) *The battle will end in [death* $_{EFF}$*].*

Opposition (OPP)
The Opposition case represents an entity that contrasts with or opposes the act but is insufficient to prevent it from happening.

(37) *Despite [my warning* $_{OPP}$*] they persisted.*

In (37), *my warning* opposes their persistence but does not prevent it.

Purpose (PURP)
The Purpose case represents the situation intended to result from the act's execution. This case implies initiation of the act by a sentient being. Purpose differs from Effect in that Purpose is the intended though not necessarily occurring result whereas Effect is the occurring though not necessarily intended result.

(38) *The drug was invented for [pain relief* $_{PURP}$*].*

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

4.4.3 Space

Direction (DIR), LocationFrom (LFRM), LocationThrough (LTRU), LocationTo (LTO)

The spatial cases Direction, LocationFrom, LocationTo and LocationThrough represent positions in some (possibly non-physical) space. To distinguish between them consider an act of motion to be an arrow in three-dimensional space. The tail of the arrow represents the starting point of the action (LFRM). The head represents the destination of the action (LTO). The direction of the arrow corresponds to the act's bearing in space (DIR). The shaft joining the tail to the head corresponds to the space through which the act passes or extends (LTRU).

(39) *I drove [east $_{DIR}$] on [autoroute 640 $_{LTRU}$] from [Laval $_{LFRM}$] to [Repentigny $_{LTO}$].*
(40) *Look inside [yourself $_{DIR}$] for the answer.*
(41) *They stripped [him $_{LFRM}$] of his pride.*
(42) *Strange images flitted through [my mind $_{LTRU}$].*
(43) *His life is going [nowhere $_{LTO}$].*

Example (39) shows all four cases together in concrete (physical) usage.

Examples (40) to (43) illustrate abstract (non-physical) uses of Direction (*yourself*), LocationFrom (*him*), LocationThrough (*my mind*) and the LocationTo case (*nowhere*).

LocationAt (LAT)

The LocationAt case represents the space in or at which an act or event occurs. This case typically only applies to non-motion acts that occur statically in space, or to specific points along the path of a motion act. There is no restriction on the physical extent of a point that fills the LocationAt case. As with other locative cases, LocationAt may be interpreted abstractly (non-physically), and either absolutely or relative to another location.

(44) *We stopped at [a restaurant $_{LAT}$] on our way home.*
(45) *Memory of the accident was hidden in [his subconscious $_{LAT}$].*

In (44), *a restaurant* is a specific point along the path of motion; (45) illustrates an abstract usage of the LocationAt case.

Orientation (ORNT)

The Orientation case represents an object's bearing in terms of, or its rotation about, the three-dimensional axes on whose origin the object is centred. Less strictly, Orientation may represent the position of parts of an object with respect to each other.

(46) *The carton lay on [its side $_{ORNT}$].*
(47) *The statue stood [erect $_{ORNT}$] despite the heavy winds.*

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

4.4.4 Time

Frequency (FREQ)

The Frequency case represents the rate at which an act recurs.

(48) *Buses arrive [in five minute intervals* $_{pp/FREQ}$*].*
(49) *He washes his car [daily* $_{adv/FREQ}$*].*

The two example sentences show the Frequency case being marked first by a preposition and then by an adverb.

TimeAt (TAT)

The TimeAt case represents the time at which an act or event takes place or a state exists. It is the temporal analog of the spatial LocationAt case. There is no restriction on the extent of the time unit filling the TimeAt case – the filler need not refer to a measurable instant in time. It may also be an event indicating when the act took place.

(50) *He traveled extensively [last year* $_{TAT}$*].*
(51) *He made his fortune during [his visit to Europe* $_{TAT}$*].*

The case filler in (50) is a time unit with measurable extent.

In (51) the nominalized event *his visit to Europe* indicates the time when the action (making his fortune) occurred.

TimeFrom (TFRM)

TimeFrom is the case that represents the time of the beginning of an act or event. It may also represent the moment at which a state began to exist or will begin to exist.

(52) *He has been blind since [the war* $_{TFRM}$*].*
(53) *I will stay at your house from [Tuesday* $_{TFRM}$*] to Saturday.*

In (52) *the war* marks the time at which the state began. He has been blind from the war until now.

In (53), the moment at which the state of staying at your house will begin is in the future (*Tuesday*).

TimeThrough (TTRU)

The TimeThrough case represents the duration of an act, an event or a state of existence. In contrast to TimeAt, the filler of this case must have extent.

(54) *The meeting lasted for [six hours* $_{TTRU}$*].*

TimeTo (TTO)

The TimeTo case represents the time at which an act or event of duration ended or will end. It may also represent the moment at which a state ceased to exist or will cease to exist.

(55) *We worked from nine to [five* $_{TTO}$*].*
(56) *I will stay at your house until [Tuesday* $_{TTO}$*].*

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

4.4.5 Quality

Content (CONT)

The Content case represents the subject of any type of communication or consideration. It may also represent the physical filling of a container.

(57) *He wrote about [birds* $_{CONT}$*].*
(58) *I am concerned about [the economy* $_{CONT}$*].*
(59) *He filled the container with [milk* $_{CONT}$*].*

Manner (MANR)

The Manner case represents a way in which an act is performed. This case accounts for many common relationships between a verb and its arguments that describe qualitative characteristics of the act or state. MANR is often used in the absence of a more suitable case and can thus be considered a default. This is particularly true for adverbial markers.

(60) *She works [with style* $_{pp/MANR}$*].*
(61) *He sings [beautifully* $_{adv/MANR}$*].*

Material (MATR)

The Material case represents the physical substance composing an object involved in an act.

(62) *We build houses with [brick* $_{MATR}$*].*

Measure (MEAS)

The Measure case represents some quantitative property describing the extent of an act or the amount, number or value (including economic worth) of an entity involved in the act. An instance of Measure implies a scale of measurement.

(63) *The horse won by [three lengths* $_{MEAS}$*].*
(64) *I bought the car for [five hundred dollars* $_{MEAS}$*].*

Order (ORD)

The Order case represents the relative position of an entity within a sequence, or within another structured arrangement of entities.

(65) *He filed the Baker file before [the Abel file* $_{ORD}$*].*

5 Evaluation

We have so far described an approach to building a set of cases driven by English case marker data from dictionaries and word lists. Evaluation of the set was accomplished by performing case analysis of text. Although the system has been used in many experiments using different texts, this section will concentrate on a single experiment designed explicitly to validate the coverage of the cases.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

5.1 Semi-automatic text analysis

The case analysis of a large number of sentences can be an onerous task when performed by hand. In order to lessen the burden of the task, we have used the automatic syntactic analysis and semi-automatic semantic analysis tools of TANKA to identify the appropriate verb arguments in a clause and to suggest likely case assignments for them.

5.1.1 Syntactic analysis

Syntactic analysis is performed by DIPETT (Domain Independent Parser of English Technical Texts), a broad coverage Definite Clause Grammar parser (Delisle, 1994) whose rules are based primarily on Quirk *et al*. (1985).

DIPETT automatically produces a single initial parse.[9] This first good parse tree is a detailed representation of the constituent structure of the clause at hand; in particular it allows us to extract case markers and fillers. Pronoun references are resolved semi-automatically immediately following syntactic analysis.

5.1.2 Semantic analysis

The HAIKU semantic analysis module (Delisle, 1994; Delisle *et al*., 1996) performs semi-automatic semantic processing on three levels: clause level relationships, noun modifier relationships and cases. The division of semantic relationships into these three levels is motivated by their syntactic manifestation at different levels. We recognize however that the same concept may be realized syntactically at any one of the levels. For example, the surface representation of an act is usually a clause, but can also be a noun phrase. This fact has been taken into consideration in the design of the different semantic analysis submodules of HAIKU.

Clause Level Relationships (*CLRs*) are semantic relationships between acts, events or states realized in text by finite clauses connected by conjunctions or conjunctive punctuation. Under the supervision of a user, CLR analysis assigns semantic labels to the connection of pairs of such clauses. The set of CLRs and the fully implemented CLR analyzer are described in Barker and Szpakowicz (1995).

Noun Modifier Relationships (*NMRs*) are semantic relationships between the head noun in a phrase and its modifiers. Noun modifier relationship analysis assigns a semantic label to each modifier of a head noun, including adjectives, premodifying nouns and postmodifying prepositional phrases. Noun phrases identified as nominalizations of acts often have the arguments of the act attached to the head noun as postmodifying prepositional phrases. NMR analysis allows the assignment of case roles to such arguments (see also section 6.3). Early research on NMRs and NMR analysis appears in Barker (1997).

Cases are the verb-argument roles under investigation in this paper. Within HAIKU cases are assigned to all verb-argument pairs in finite clauses, including relative

[9] In general, there may be more than one valid parse for a given English sentence. An alternative to producing multiple parse trees is to generate a single tree that can be edited if necessary (for example, to reattach a misattached prepositional phrase). This single parse alternative is explored by various researchers, including Jensen *et al*. (1993, chapter 3) and Delisle (1994).

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

clauses syntactically postmodifying noun phrases. The need for a usable set of clearly defined cases for HAIKU partially motivates this paper.

5.1.3 A closer look at case analysis in HAIKU

The case analysis module in HAIKU extracts all case markers in a given clause based on their syntactic labels as identified by DIPETT. It assembles these markers into a case marker pattern (CMP). HAIKU then attempts to determine the corresponding case pattern (CP) by comparing the current CMP and verb to CMPs and verbs stored from previous clauses.

If the current case marker pattern CMP_C has already appeared in some previous clause $Clause_P$, then the previous case pattern CP_P is suggested to the user for the current clause. If CMP_C has never been encountered, a partial matching algorithm finds the nearest neighbour marker pattern CMP_N among previous clauses. The measure of CMP distance for the nearest neighbour calculation is defined in Delisle *et al*. (1993). Case patterns corresponding to CMP_N contain candidate cases for the current CP.

If the system is unable to find good candidate cases to suggest to the user (for example when processing the first sentence of a text), the user must supply a case pattern using the formal definitions of the cases (section 4.4) to guide case selection. Once the user has accepted a suggested CP or supplied her own, the CP is stored along with CMP_C and the current verb V_C. These stored patterns are accumulated to help in the processing of subsequent clauses.

As an example, consider (66) and (67).

(66) *[I $_{subj/AGT}$] [usually $_{adv/FREQ}$] print [my documents $_{dobj/OBJ}$] [on the laser printer $_{pp/INST}$]*

(67) *[Alice $_{subj/AGT}$] prints [her papers $_{dobj/OBJ}$] [on the bubblejet $_{pp/INST}$]*

After (66) has been analyzed, HAIKU stores $CMP_{(66)}$ (subj-dobj-on-adv) and $CP_{(66)}$ (AGT-OBJ-INST-FREQ) along with the verb *print*. When analyzing (67), HAIKU compares the current $CMP_{(67)}$ (subj-dobj-on) and verb (*print*) with previous CMPs and verbs. It finds that the nearest neighbour CMP was $CMP_{(66)}$. The candidate cases for $CP_{(67)}$ are the corresponding cases from $CP_{(66)}$, namely, AGT-OBJ-INST.

Since all distinct (V CMP CP) triples are stored, there has been concern that the number of CPs suggested to the user will become unmanageable as more clauses are analyzed. However, the experiment documented in Barker and Delisle (1996) shows several encouraging results. More details on this experiment appear in section 5.2, but in general,

- the maximum number of CPs suggested to the user increases slowly throughout the analysis of a text;
- interaction with the user does not become overly burdensome, even as numbers of suggested CPs increase;
- as more clauses are analyzed, the system learns to make better suggestions, meaning that the user is more likely to find the appropriate CP among the suggestions and less likely to be required to supply the CP herself;

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

- the interaction time spent by the user on each sentence throughout the analysis of a text decreases on average.

If none of the cases in the existing set are satisfactory, HAIKU allows the user to define new cases during analysis. These would subsequently be examined closely for inclusion in the set. However, in tests to date no user defined cases have been needed. Full details of HAIKU's case analyzer are presented in Delisle *et al*. (1996).

5.2 Experimental results

We have argued that the validation of our case system comes from its use in the case analysis of some large number of real English sentences. Previously, we have tested HAIKU's case analyzer on such English texts as the Canadian income tax guide and a fourth generation computer language manual. Since settling on the current set of twenty-eight cases, no tests have identified a need for new cases. However, none of these tests were conducted primarily to validate the coverage of the cases themselves. The test described in Barker and Delisle (1996), and referred to here as the *case test*, was designed with several explicit goals, among them the experimental validation of the case system.

The case test was conducted on the book *Junior Science Book of Rain*, *Hail*, *Sleet & Snow* (Larrick, 1961). This book uses less complicated language than many technical texts, allowing us to get a high percentage of sentences parsed well enough to exercise the HAIKU semantic analyzer. Furthermore, the majority of the information in the book is carried in finite clauses, which makes it more useful for evaluating the case system in particular. Texts that are heavy on other constructs (such as complex noun phrases) would exercise other parts of HAIKU (such as the noun modifier relationship analyzer).

Larrick (1961) contains 513 sentences from which DIPETT extracted 439 finite, non-stative clauses usable for case analysis. Stative clauses were parsed but not case analyzed, since stative clauses usually represent noun modifier relationships, not cases. The system was able to assemble automatically the correct CMP for 69% of the clauses. The user supplied the correct CMP for the remaining clauses.

Starting with an empty processing history, the system suggested zero or more case patterns for each CMP, depending on previous processing as described in section 5.1.3. Over all 439 clauses, the system made on average fewer than five CP suggestions per CMP. After processing all clauses, the maximum number of CP suggestions made for any single clause was fourteen. The increase in this maximum was quick for the first half of the experiment (jumping from zero to eleven over the first 200 clauses) but slowed for the second half, where the maximum increased by only three over the last 200 clauses.

The correct case pattern was among the system's suggestions for 62% of the 439 clauses – 50% for the first 208 clauses and 72% for the next 231 clauses. Fig. 2 shows the number of correct case patterns determined automatically by the system over the course of the experiment contrasted with the number that had to be supplied by the user.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

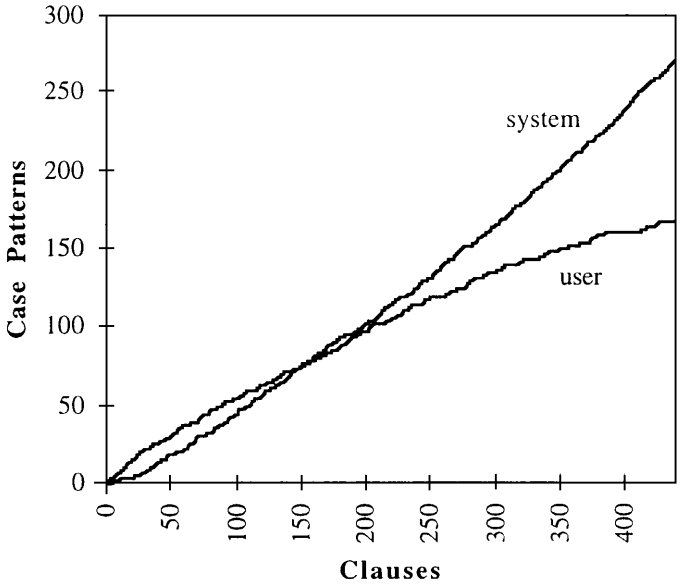


Fig. 2. Number of correct case patterns determined by the system vs. those supplied by the user.

The experiment also recorded a 'user burden' for each analysis, measuring the degree of difficulty of case assignment of a clause. The measurement was based on the amount of time the user spent on a given assignment along with a more subjective evaluation of how difficult it was for the user to find appropriate cases for the clause. User burden was low for 87% of the clauses, whether the case pattern was suggested by the system or supplied by the user. By low burden we mean that for each verb-argument relationship there was a single obvious case that best captured the semantics of the relationship and the case was identified quickly (within a few seconds). Case assignment for the remaining 13% of the clauses was more difficult, requiring some consultation of the definitions of the cases. The complement ing-clause of (68) was among these more difficult clauses.

(68) *So weather stations began*
[sending up [airplanes $_{dobj}$*] [to get reports on the clouds* $_{adv}$*]* $_{compl}$*].*

In the main clause, *weather stations* is the Agent of *began* and the ing-clause *sending up airplanes to get reports on the clouds* is the Object. Within the ing-clause, *airplanes* is the direct object of the verb *sending* and the to-infinitive clause *to get reports on the clouds* is parsed as an adverbial of the verb *sending*. The word *up* is an adverbial particle considered part of the verb sequence and is not case analyzed. HAIKU had no suggestions for the verb *send* with the CMP dobj-adv, so the user had to supply the case pattern Object-Purpose. That is, *airplanes* is the semantic Object of *sending up* and *to get reports on the clouds* is the Purpose of *sending up*. This example was considered slightly more difficult than most because of the clausal structure of the sentence. The verb *began* has only two case fillers: *weather stations* and the *sending*

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

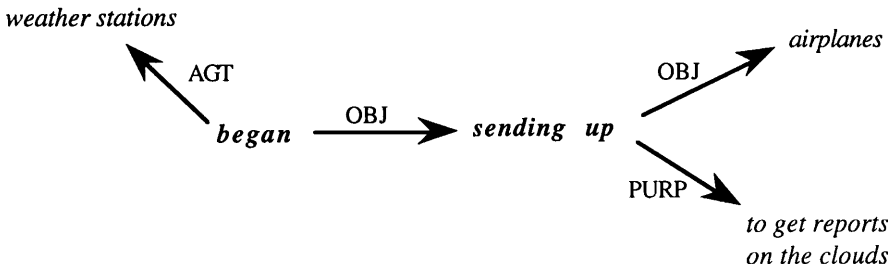


Fig. 3. Case analysis of (68).

up clause. The *sending up* clause also has only two cases: *airplanes* and the to-infinitive *get* clause. The (incorrect) intuition was to assign *weather stations* an Agent role in the *sending up* clause. Fig. 3 offers a graphic view of the correct analysis.

No new cases were needed to capture the semantics of the 55 difficult clauses. On the other hand, four cases were never assigned. Section 5.3.3 looks at these four more closely to determine if they are superfluous or redundant.

5.3 Revisiting the requirements

We now revisit the requirements we identified in section 4.1 and attempt to evaluate the extent to which our case system is *general*, *complete* and *unique*, based on experimental results and on a survey of the distribution of cases among the case markers (Barker, 1996).

5.3.1 Generality

In section 4.1.1 we proposed two different methods for measuring the generality of a set of cases: 1) the number of times each case is assigned in the analysis of a large body of text; 2) the number of syntactic phenomena accounted for by each case.

The experiment described in section 5.2 is too small to be used for the first measure. A single text with only 439 case analyzable clauses is insufficient for drawing general conclusions about the applicability of each of the cases.[10] Until the data from many more texts are available, we will restrict our evaluation of generality to the second measure.

Barker (1996) tabulates the number of markers for each of TANKA's 28 cases. These data are then used to identify the degree to which the markers support the inclusion of each case in the set. Cases whose marker representation is below average are potentially over specific and considered candidates for combination into a more general case. Cases whose marker representation is above average are potentially over general and considered candidates for splitting into more specific cases. The case-by-case analysis identified a possible combination of Effect and Purpose into a more general case for applications with no need to distinguish intended from unintended

[10] It is difficult to know how much text would be enough to draw conclusions about the generality of the cases. In the corpus statistics community it is not uncommon to deal with tens or even hundreds of thousands of sentences.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

outcome or realized from unrealized outcome. It also identified possible refinements of LocationAt and TimeAt into more specific cases for applications sensitive to the subtleties of spatial or temporal phenomena. Finally, the analysis identified three very common marker→case pairs. Although Agent, Experiencer and Object are weakly represented by the markers, the very high frequency of positional subjects and direct objects in text result in these three cases accounting for a large number of case assignments in practice.

We have also noted cases that seem to have general definitions but low marker representation and low occurrence rates in text. For example, the definition of Content allows for both concrete (contents of a container) and abstract (content of communication) interpretations. It could be argued that these are two separate cases. However, since the Content case is already weakly represented by the markers and experimental data, and since the two different interpretations seem not to appear together in the same sentence, we have left the Content case as is. More data might justify a split after all.

Overall, the markers support the inclusion of each case in the set, meaning that each case accounts for a reasonable number of syntactic phenomena and can be expected to generalize a reasonable number of verb-argument relationships in practice. As more texts are analyzed using our cases, the extent to which this expectation is realized will be monitored.

5.3.2 Completeness

As mentioned in section 4.1.2, a practical measure of the completeness of a case set is the number of act-argument relationships in a text covered by *only* the cases in the set.

After several years of use with hundreds of sentences in several different technical domains we have not yet found a verb argument role that is not covered by our set of 28 cases. In particular, the experiment described in section 5.2 assigned 965 cases among the 439 clauses. All 965 case instances were covered by the existing cases within the scope of their current definitions. As we continue to analyze more and more sentences, our confidence in the completeness of TANKA's case set increases.

5.3.3 Uniqueness

The definition of uniqueness that we offered in section 4.1.3 has two parts: no redundant cases and no superfluous cases.

The presentation of formal case definitions in this paper is an attempt at avoiding redundant cases. The use of consistent language in these definitions may make them uninteresting, but it helps highlight similarities between cases. Where definitions exposed such similarities, an explicit distinction based on semantic or syntactic grounds was offered with the definitions, as for Agent and Experiencer, Recipient and Beneficiary, Effect and Purpose.

Case marker distribution was also used to avoid redundant and superfluous cases. Cases with identical sets of markers might be redundant; a case whose markers are

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

a strict subset of another case's markers might be superfluous. Marker distribution was the main evidence used in the decision to reject TimeBefore and TimeAfter as superfluous cases in our set.

Example sentences were very useful for improving the uniqueness of the case set. In addition to the examples given for each definition in this paper, an example sentence is stored with every entry in our dictionary of 500 marker → case pairs. These example sentences were carefully constructed to cover all of the case marking senses of each marker. They provide a justification for the inclusion of a case as marked by a given marker, which in turn justifies the inclusion of the case in the case system.

Redundant Cases

In an experimental setting, redundancy in the case system would be uncovered if, when analyzing a clause, two or more cases were appropriate for a given interpretation of a particular verb-argument relationship. To date, we have not encountered this problem using the existing set of cases. Although a sentence may itself be ambiguous, once a single interpretation has been chosen the choice of cases is unambiguous. For example, (69) has two interpretations depending on whether the noun phrase *different size hammers* fills the Instrument case or the Material case.

(69) *I make my sculptures with [different size hammers* $_{???}$*].*

The ambiguity is not due to any redundancy between Instrument and Material. If the intended meaning of (69) is that the sculptures themselves consist of some arrangement of hammers, then *different size hammers* unambiguously fills the Material case. If the intended meaning is that the hammers are used to construct the sculpture (out of wood and nails, for example), then the Instrument case is more appropriate.

Note that if two cases can appear in the same clause (as Instrument and Material do in (70)), they are not redundant since a single case may not appear twice in the same clause.

(70) *I make my sculptures out of [wood and nails* $_{MATR}$*]*
with [different size hammers $_{INST}$*].*

Superfluous Cases

Any general purpose, domain independent case system may include cases that are superfluous to the analysis of a particular text. In the case test described in section 5.2 four cases were not used: Opposition, Order, TimeFrom and TimeTo. We now examine these cases to determine if they are indeed superfluous or merely not represented in the text analyzed in the case test.

The Opposition case represents an entity that contrasts with or opposes the act but is insufficient to prevent it from happening. Despite the fact that this case was not used at all in the analysis of the test text, the only other case that might be able to capture the relationship marked by appropriate senses of such markers as *against*, *considering*, *despite*, *in spite of*, *notwithstanding*, *versus*, *nevertheless* and *nonetheless* is

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

Exclusion. (71) shows both Exclusion and Opposition being used in the same sentence, which prevents us from folding Opposition into the definition for Exclusion.

(71) They went without [me $_{EXCL}$] despite [my pleas $_{OPP}$].

Order represents the relative position of an entity within a structured arrangement of entities, as marked by *after*, *ahead of*, *before*, *below*, *beneath*, *beyond*, *underneath*, *first*, *last*, *lastly*, *next*, *primarily*, *second*, *secondly*, etc. The only other case in the set that could capture these relationships is LocationAt. Example (72) contains instances of both LocationAt and Order.

(72) Put the Åkerblads file in [my filing cabinet $_{LAT}$] after [the A's $_{ORD}$].

The other two cases not used were TimeFrom and TimeTo. These are the natural temporal counterparts of the locative LocationFrom and LocationTo cases, both of which appeared in the text in the case test. The TimeFrom and TimeTo cases are also well represented by the case markers, as shown in Barker (1996).

6 Future work

The areas of future work described in this section concern the use of the case system in the TANKA project. It is through application that the case system will evolve and its validation continue.

6.1 Case analysis interaction

We argued in sections 5.1 and 5.2 that the number of case pattern suggestions made by HAIKU increases slowly and does not become unmanageable during the analysis of a text. The experiments that allowed us to make these claims revealed another interesting result.

Regardless of the surface order of constituents in a sentence, the case analyzer assembles case markers into a case marker pattern in a consistent order:

$$\text{subj} > \text{dobj} > \text{iobj} > \text{prep(s)} > \text{adv(s)}$$

This ordering was chosen based on the SVO (subject-verb-object) word order in English as well as the notion of clause structure element centrality (see Quirk *et al.* (1985, section 2.13)). The notion of centrality comes from the observation that the verb is the most central element of the clause; the adverbial is the least central, or most peripheral. Between these two extremes, the subject, then the objects and then any other non-adverbial complements (such as prepositional phrases) are decreasingly central.

The cases within a case pattern are stored in the order of the corresponding case markers. For example, in (73), the order of the case marker pattern would be subj-dobj-iobj-in-adv and the corresponding case pattern would be AGT-OBJ-RECP-LAT-TAT.

(73) [Yesterday $_{adv}$] [I $_{subj}$] gave [Sandra $_{iobj}$] [a diskette $_{dobj}$] [in the lab $_{pp}$].

When suggesting multiple case patterns to the user, the system sorts them

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

alphabetically so that all CPs with the same case in the subj position appear together. This allows the user to consider or reject *groups* of suggested CPs, making the number of suggestions more manageable. For example, consider the simple (74).

(74) *[The program $_{subj}$] printed [the document $_{dobj}$].*

The case marker pattern subj-dobj is probably the most common CMP and is likely to have many corresponding case patterns. At one point in one of our experiments, twelve CPs had been accumulated for the CMP subj-dobj. These CPs were stored with the case corresponding to the subj first followed by the case corresponding to the dobj and were presented to the user in the following order:

(1) AGT-DIR
(2) AGT-EFF
(3) AGT-OBJ
(4) CONT-EXPR
(5) EXPR-CAUS
(6) EXPR-MANR
(7) EXPR-MEAS
(8) EXPR-OBJ
(9) EXPR-TAT
(10) OBJ-MEAS
(11) RECP-MEAS
(12) RECP-OBJ

The user first looks at CP (1) considering AGT as a potential case for the subject *the program*. Once the user decides that AGT is indeed the correct case for *the program*, she must choose a case for the direct object *the document*. But now there are only three suggestions to choose from, since none of the others assign AGT to the subject of the sentence.

A possible modification to the system would be to present the user with CP suggestions one case at a time. Instead of suggesting twelve CPs for (74), HAIKU would first present only the cases corresponding to the subject position:

(1) AGT
(2) CONT
(3) EXPR
(4) OBJ
(5) RECP

If the user chose case (1) for the subject, she would then be shown only those object position cases appearing in patterns with AGT in the subject position:

(1) DIR
(2) EFF
(3) OBJ

This partial case assignment method would need to be tested in a large experiment to confirm that such an interaction is in fact easier for the user.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

6.2 Noun semantics

Currently HAIKU makes use of information from earlier analyses to make good case pattern suggestions for a given sentence, as described briefly in section 5.1.3 and in more detail in Delisle *et al.* (1996). If the current verb has occurred previously with the same or a similar case marker pattern, HAIKU will suggest the same case pattern. This process could be extended to consider the semantics of case fillers.

Suppose a current case marker is *for* and its noun filler belongs to the *human* semantic category. If *for* has previously occurred with a *human* case filler and the corresponding case was determined to be Beneficiary, then there is some probability that the current case is also Beneficiary. When stored case patterns become many, this extra knowledge would help to determine a preference among the CPs associated with a given CMP.

Implementation would require that every noun in the dictionary be given a semantic tag. However, one of the tenets of TANKA's philosophy is to minimize the use of hand coded *a priori* semantic knowledge. To avoid hand coding, we would need to acquire semantic tags automatically or semi-automatically from an existing online source. The WordNet system (Miller, 1990) can supply consistent semantic class information automatically for many nouns. This information could be used to improve HAIKU's ability to make good case assignment suggestions in the manner proposed above. Delisle (1996) describes LEXICOGRAPHER, a module that semi-automatically constructs a lexicon for a given text or corpus. For every entry in the lexicon, LEXICOGRAPHER can record part of speech information, morphology, concordances as well as information from WordNet. Other applications of WordNet in TANKA are discussed in Feng *et al.* (1994), Li *et al.* (1995) and Szpakowicz *et al.* (1996).

6.3 Noun modifier relationships

The case system described in this paper is limited to the analysis of verb-argument relationships (in keeping with case theory). However, often the choice of a surface representation is arbitrary, as illustrated by sentence (75) and noun phrase (76), both of which describe the same event.

(75) *A man was murdered yesterday with a handgun because a jealous husband returned home early.*

(76) *the murder of a man yesterday with a handgun because of the early return of a jealous husband*

It would be desirable to assign the same semantic labels to the corresponding arguments in each example. In both situations, *a handgun* is the Instrument of a *murdering* event, *a jealous husband* is the Agent of a *returning* event, etc. There is obviously some overlap in the types of semantic relationships possible at each of the three levels we discussed in section 5.1.2.

We need a way to identify which noun phrases would be more suited to analysis using case relationships than noun modifier relationships (such as noun phrase (76)).

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

There are at least two approaches open to us for identifying *deverbal* nouns: a morphological/lexical approach and a semantic approach.

The morphological/lexical approach would only require an online dictionary with part of speech information. Such dictionaries are readily available and were used in the construction of our case marker list. Head nouns having both noun and verb entries in the dictionary would be likely candidates for semantic analysis using the cases. Similarly, head nouns with the known deverbal suffixes (such as -ant, -ee, -er, -or, -age, -al, -ation, -tion, -sion, -ing, -ment, etc.) could be massaged into likely verbal forms. If the verbal form appears in the dictionary as a verb, the noun is again a likely candidate for analysis using the cases.

Noun semantics could be used for the same purpose. Large noun taxonomies are less common and often incomplete and arbitrary. However, a resource such as WordNet could be used to find semantic information such as whether a noun represents an activity type concept rather than an entity type concept. Such nouns would be likely candidates for analysis with cases.

6.4 Experimental validation

We have already tested the HAIKU semantic analyzer on many English texts, including a computer software manual, a Canadian income tax guide, the Ontario building code and others. Nonetheless, the case system can only be strengthened by more experimental validation. There is almost certainly a compelling example somewhere that is not covered by our set of 28 cases, though that example becomes less statistically significant as more sentences *are* covered. Moreover, the case analyzer itself and the other semantic analysis modules within HAIKU can always benefit from more testing, especially as bits of Future Work make their way into Past Work.

Acknowledgments

The authors wish to thank the members of the extended TANKA family for input during the development of TANKA's case system. We would also like to thank them for subsequently trying to break it. We thank André Vellino for casting an objective eye on an earlier version of this paper and on the undertaking in general. Finally, we would like to express appreciation for the useful comments and pointers provided by anonymous reviewers.

Appendix I: Sample output

This appendix shows a typical HAIKU interaction for a single sentence. The text is direct output from the implemented HAIKU system with minor editing (formatted headings, some details of the output removed for clarity, etc.). User input is shown boldface. Details of the processes behind the CLR interaction are presented in Barker and Szpakowicz (1995). The case analyzer is described in full in Delisle (1994) and Delisle *et al*. (1996).

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

```
Current sentence:
= = = = = = = = = =
Since Rosa usually supports her daughter, she can claim
childcare expenses on her income taxes.
```

⟨*parse tree not shown*⟩
⟨*anaphora resolution interaction not shown*⟩

```
Clause-Level Relationship analysis:
= = = = = = = = = = = = = = = = = = = =
There is an Enablement Clause-Level Relationship marked by
'since':
     'rosa usually supports her daughter'
        ⟨enables⟩
     'rosa can claim childcare expenses on her income taxes'

Do you accept this assignment? Y

Case analysis:
= = = = = = =
For the verb 'claim':
     SUBJECT    : rosa
     COMPLEMENT : child care expenses
                  on her income tax

Do you accept the CMP subj-dobj-on? Y

The most likely candidate CPs are:
     (1) agt-obj-lat
Please enter a number between 1 and 1 (or 0 if none is
appropriate): 1

For the verb 'support':
     SUBJECT    : rosa
     COMPLEMENT : her daughter
     MODIFIERS  : usually

Do you accept the CMP subj-dobj-adv? Y
The most likely candidate CPs are:
     (1) agt-obj-freq
     (2) agt-obj-lat
     (3) agt-obj-manr
Please enter a number between 1 and 3 (or 0 if none is
appropriate): 1
```

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

Appendix II: Case markers

This appendix is an excerpt from HAIKU's domain independent case marker dictionary showing only the entries for preposition case markers beginning with the letter 'a'. In the example sentences, markers appear boldface and case fillers appear in capital letters. The complete list contains more than 500 marker→case mappings. The dictionary ranks every case marked by a given marker as more common or less common. Less common marker→case pairs are shown here in a smaller, italic font.

Marker	*Case*	*Example*
aboard	*lat*	*He lived **aboard** THE SHIP.*
	lto	*He climbed **aboard** THE SHIP.*
about	cont	He wrote about MATHEMATICS.
	ltru	*He wandered **about** THE CITY this afternoon.*
	lat	*The flowers were arranged **about** A SINGLE WHITE ROSE.*
above	lat	Put the horseshoe **above** THE MANTLE.
	dir	Look **above** THE CLOUDS.
across	lat	I live in the house **across** THE RIVER.
	dir	She looked **across** THE STREET to find a pay phone.
	ltru	We drove **across** THE BRIDGE.
after	ord	File it **after** THE ALPHABETIC ENTRIES.
	tat	File it **after** FIVE.
	lat	He drove **after** THE BUS.
against	opp	She spoke **against** THE PROPOSAL.
	dir	The rain beats **against** THE WINDOW.
	lat	Mount the picture **against** A BLUE BACKGROUND.
	purp	*He is saving **against** RETIREMENT.*
ahead of	lat	He stood **ahead of** ALL THE OTHER PLAYERS.
	ord	She finished **ahead of** THE REST OF THE CLASS.
along	lat	We stood **along** THE BANK as the boat went by.
	ltru	He ran **along** THE RIVER to the bridge.
alongside	*lat*	*The tug came **alongside** THE LINER.*
amid	*lat*	*He stood **amid** THE GUESTS at the party.*
amidst	*lat*	*He stood **amidst** THE GUESTS at the party.*
among	lat	Stand **among** US or you'll be volunteered.
	acmp	Divide it **among** THEM.
amongst	lat	Stand **amongst** US or you'll be chosen.
	acmp	Divide it **amongst** THEM.
apart from	excl	**Apart** FROM THAT, what is left to do?
around	ltru	We walked **around** THE COURTYARD sightseeing.
as	purp	She acted **as** MY AGENT in Europe last year.
as of	*tfrm*	*The system is up **as of** YESTERDAY.*
aside from	*excl*	***Aside from** THAT, what is left to do?*
at	dir	The deer ran right **at** THE HUNTERS.
	lat	I stood **at** THE DOOR.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

	tat	The convict will be executed **at** HIGH NOON.
	manr	The car moves **at** HIGH SPEED.
	cont	She is good **at** ARTS.
	meas	It stopped **at** FIFTY.
	caus	She was amazed **at** HIS GALL.
atop	*lat*	*Place the book **atop** THE BUREAU.*

References

Barker, K. (1996) The assessment of semantic case roles using English positional, prepositional and adverbial case markers. *TR-96-08*, Department of Computer Science, University of Ottawa.

Barker, K. (1997) Noun modifier relationship analysis in the TANKA system. *TR-97-02*, Department of Computer Science, University of Ottawa.

Barker, K. and Delisle, S. (1996) Experimental validation of a semi-automatic text analyzer. *TR-96-01*, Department of Computer Science, University of Ottawa.

Barker, K. and Szpakowicz, S. (1995) Interactive semantic analysis of clause-level relationships. *Proceedings of the Second Conference of the Pacific Association for Computational Linguistics*. Brisbane, Australia. Pp. 22–30.

Bruce, B. (1975) Case systems for natural language. *Artificial Intelligence* **6**: 327–360.

Cook, W. A. (1989). *Case Grammar Theory*. Washington: Georgetown University Press.

Copeck, T., Barker, K., Delisle, S., Szpakowicz, S. and Delannoy, J.-F. (1997) What is technical text? *Language Sciences* (to appear).

Delisle, S. (1994) Text processing without a-priori domain knowledge: semi-automatic linguistic analysis for incremental knowledge acquisition. *PhD thesis*, *TR-94-02*, Department of Computer Science, University of Ottawa.

Delisle, S. (1996) Le traitement automatique du langage naturel au service de l'ingénieur de la connaissance: le système reader. *Actes de la Conférence Internationale sur le Traitment Automatique des Langues et ses Applications Industrielles*, Moncton. Pp. 60–66.

Delisle, S., Copeck, T., Szpakowicz, S. and Barker, K. (1993) Pattern matching for case analysis: a computational definition of closeness. *Proceedings of ICCI-93*. Pp. 310–315.

Delisle, S., Barker, K., Copeck, T. and Szpakowicz, S. (1996) Interactive semantic analysis of technical texts. *Computational Intelligence* **12**(2): 273–306.

Feng, C., Copeck, T., Szpakowicz, S. and Matwin, S.. (1994) Semantic clustering: acquisition of partial ontologies from public domain lexical sources. *Proceedings of AAAI Knowledge Acquisition for Knowledge-Based systems Workshop*, Banff, Canada. Pp. 1–16.

Fillmore, C. J. (1968) The case for case. In: E. Bach and R. T. Harms (eds.), *Universals in Linguistic Theory*. Rinehart and Winston, New York.

Fowler, H. W. (1984) *A Dictionary of Modern English Usage*: *second edition* (*revised by Sir Ernest Gowers*). Oxford University Press.

Grishman, R. (1995) MUC-6. http://cs.nyu.edu/cs/faculty/grishman/muc6.html.

Jensen, K., Heidorn, G. E. and Richardson, S. D. (eds.). (1993) *Natural Language Processing*: *the PLNLP approach*. Kluwer Academic, Boston.

Larrick, N. (1961) *Junior Science Book of Rain*, *Hail*, *Sleet* & *Snow*. Garrard Publishing, Champaign, IL.

Larson, M. L. (1984) *Meaning-Based Translation*: *A guide to cross-language equivalence*. University Press of America, Lanham.

Levin, B. (1993) *English Verb Classes and Alternations*: *A preliminary investigation*. University of Chicago Press.

Li, X., Szpakowicz, S. and Matwin, S. (1995) A WordNet-based algorithm for word sense disambiguation. *Proceedings of IJCAI-95*, Montréal, Canada. Pp. 1368–1374.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press

Miller, G. A. (ed.). (1990) WordNet: an on-line lexical database. *International Journal of Lexicography* **3**(4).

Oflazer, K. and Yilmaz, O. (1996) A constraint-based case frame lexicon. *Proceedings of COLING-96*, Copenhagen, Denmark.

Palmer, M. S. (1990) *Semantic Processing for Finite Domains*. Cambridge University Press.

Pustejovsky, J. and Busa, F. (1995) A revised template description for time (v3). http://cs.nyu.edu/cs/faculty/grishman/time-guidelines.v3–1.html

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. Longman, London.

Sinclair, J. (ed.). (1991) *Collins COBUILD English Language Dictionary*. Collins, London.

Slator, B. M., Amirsoleymani, S., Andersen, S., Braaten, K., Davis, J., Ficek, R., Hakimzadeh, H., McCann, L., Rajkumar, J., Thangiah, S. and Thureen, D. (1990) Towards empirically derived semantic classes. *Proceedings of the Fifth Rocky Mountain Conference on Artificial Intelligence*, Las Cruces. Pp. 257–262.

Somers, H. L. (1987) *Valency and Case in Computational Linguistics*. Edinburgh University Press.

Sowa, J. F. (1984) *Conceptual Structures*: *Information processing in mind and machine*. Addison-Wesley, Reading, MA.

Sparck Jones, K. and Boguraev, B. (1987) A note on the study of cases. *Computational Linguistics* **13**(1–2): 65–68.

Stein, J. (ed.). (1983) *The Random House Dictionary of the English Language*. Random House, New York.

Summers, D. (ed.). (1987) *Longman Dictionary of Contemporary English*: *new edition*. Longman, Essex.

Szpakowicz, S., Matwin, S. and Barker, K. (1996) WordNet-based word sense disambiguation that works for short texts. *TR-96-03*, Department of Computer Science, University of Ottawa.

Van Valin, R. D. (ed.). (1993) *Advances in Role and Reference Grammar*. John Benjamins, Amsterdam.

Velardi, P., Pazienzi, M. T. and Fasolo, M. (1991) How to encode semantic knowledge: a method for meaning representation and computer-aided acquisition. *Computational Linguistics* **17**(2): 153–170.

Wilks, Y. A., Slator, B. M. and Guthrie, L. M. (1996) *Electric Words*: *Dictionaries*, *computers*, *and meanings*. MIT Press.

Wu, D. (1993) An image-schematic system of thematic roles. *Proceedings of the First Conference of the Pacific Association for Computational Linguistics*, Vancouver, Canada. Pp. 323–332.

https://doi.org/10.1017/S1351324997001460 Published online by Cambridge University Press