3 Case Relationships

Case relationships are assigned to a verb and its positional, prepositional and adverbial arguments. Evaluating lists of cases involves comparing them to other lists and checking their coverage of lexical and syntactic markers. The occurrence of cases in texts can be used as an indicator of their generality and their practical coverage.

This chapter is not about case analysis, the second part of HAIKU's three-part analysis. It is about the case relationships themselves.[1] Of HAIKU's three sets of semantic relationships, the cases are the oldest and have been used in the analysis of more texts than either the clause level relationships or the noun modifier relationships. The processes of constructing the case set and evaluating it described in detail in this chapter are similar for those other lists.

[1] Elements of case analysis inevitably creep into my presentation for coherence (sections 3.2.4, 3.6). Case analysis in HAIKU was a group undertaking in TANKA driven and implemented by Sylvain Delisle (Delisle 1994; Delisle *et al.* 1996). Development of the case list and case marker dictionary and evaluation of the case system were driven by me under the same TANKA wing.

3.1 Introduction

This introduction gives background on case theory and valency theory, and on other sets of case relationships. The next section (3.2) describes the lexical and syntactic elements that mark cases in clauses. The markers played a leading role in the construction and evaluation of the list of cases (as well as being used in the interactive case analyzer). In section 3.3 I describe the process of designing HAIKU's case system and define criteria for its evaluation. Case definitions appear in 3.4 along with examples. An evaluation of the case system is described in section 3.6.

3.1.1 Cases

Cases represent the semantic relationships expressed syntactically by a verb and its arguments. The verb expresses an event (see section 1.1.1); its arguments express the participants in and circumstances around the event. Researchers building natural language processing systems that use cases (or case-like relationships) 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-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.

The linguistics community has devoted more attention to the study of case systems (see Somers 1987). These studies, however, are usually driven by linguistic motivations such as, for example, language universals and psychological plausibility. The resulting case systems provide a starting point for case systems intended for computer-based applications such as information extraction or computer aided knowledge acquisition from text.

This chapter describes the construction of a complete case system that brings together elements of existing case systems and evaluates the coverage of the cases on syntactic phenomena and in real texts. The evaluation offers *practical* justifications for the cases, following the lead of Wilks *et al.* (1996) who 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.

3.1.2 Case Theory

Case theory (Fillmore 1968; Somers 1987) focuses on the simple finite clause and on the main verb within it. A case captures that part of the meaning of a syntactic element conveyed by its relationship with the verb. The occurrence of a case in a text is signaled by a case marker. Markers 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 (Fillmore 1968). The restriction can be traced back to case theory's roots in the study of grammatical roles and morphological case (Bruce 1975).

3.1.3 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, one sense of the verb *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. It is common for case system designers to distinguish cases as absolutely core or peripheral. But consider example (84).

(84) The meeting lasted [six hours $_{pobj}$*]*.

The case assigned to the direct object (written `pobj`—see section 3.2.1) 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, (85) shows a peripheral use of the same case, since it is assigned to an optional prepositional phrase argument.

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

Systems that contain core cases only sometimes have a locative case and rarely have a temporal case. To account for sentences such as (85), the temporal nature of the direct object is considered *incorporated* in the verb and a simple Object case is assigned. Alternatively, a verb such as *last* may be considered intransitive with a temporal preposition (such as *for*) lexicalized in the verb (Cook 1989).

For the purposes of automated text analysis the identification of cases that are core for a given verb sense would be a large knowledge engineering task and no existing dictionary offers help. This information would not really be useful in TANKA since we are concerned with collecting actual patterns of cases as they appear in a text: HAIKU's case analyzer captures the cases that are associated with specific arguments for the verbs in a text.

Given subcategorization rules for verbs, HAIKU could actually be used to build a dictionary of cases that appear as core and peripheral (see section 5.2.3).

3.1.4 Other Case Systems

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 6,600 references related to case. In this section I will discuss some of the more theoretical work in case theory then some of the computational work in case systems and text analysis. The distinction is not always perfect.

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 the six relationships listed in Table 14.

Agentive	Instrumental	Dative
Factitive	Locative	Objective

Table 14: List of cases from Fillmore (1968)

Bruce (1975) explores the relationship between surface case (syntactic level) and deep case (semantic level). He attempts to define a single subjective rule for deciding which cases are "central to an event description"—essentially, which cases are core for certain events. One of his observations is particularly relevant to the TANKA project. He recognizes that it is important to measure the "goodness" of a case system, although he doesn't offer a solution. He does mention that "one can only speak of the goodness of a case system relative to a problem situation" (Bruce 1975: 357) and that different problems require different levels of granularity in case systems (see section 3.3.2). Bruce also surveys several theoretical case systems and several practical case systems in use in early applications. Many of these systems are specific to the domain for which they were created.

Larson (1984) presents a list of cases to be used as an intermediate representation in a machine translation system. The computational setting demands more emphasis on identifying a *complete* set of roles (both core and peripheral) for all verb arguments. Larson's list of twelve cases appears in Table 15.

Agent	Instrument	Beneficiary	Affected
Resultant	Goal	Location	Time
Accompaniment	Causer	Manner	Measure

Table 15: List of cases from Larson (1984)

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: 20). Sowa provides a complete list of conceptual relations along with maximally general concepts they may link (according to a concept hierarchy). As is common, relations are defined (at least in part) by example sentences.

	source	***path***	***goal***	***local***
Active	instigator of action	instrument or means	intended result or active recipient	non-passive patient
Objective	original state or material	counter-instrument passive means	result state or factitive	undergoing change-of-state
Dative				
psychological	stimulus	medium	experiencer	content
possession	original owner	medium/price	recipient	thing transferred
Locative	place from where	space traversed	final destination	static position
Temporal	time since	duration	time until	time at which
Ambient	reason	manner	aim or consequence	condition

Table 16: The case grid from Somers (1987)

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 grid of four columns by six rows (Table 16). The four columns are inspired by the localist division of relations into Source, Path, Goal and Local dimensions (Anderson 1971). 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. The requirement that each category of cases map exactly onto the four dimensions seems at times artificial. Although the localist categories are intuitive, whether there is cognitive support for fitting all cases into such categories is unclear.

Cook (1989) surveys several different models of case theory including his own Matrix Model. His model distinguishes *propositional* cases from *modal* cases. Propositional cases are those essential to a verb's meaning (roughly, core); modal cases are optional adjuncts of any predication (peripheral). Cook's work deals primarily with the propositional cases, though he recognizes the utility of modal cases to automated text analysis. The matrix model contains five propositional cases; Cook also provides a list of modal cases that should appear in any case system that includes modal cases (Table 17). The list of cases was derived from an analysis of the verb-argument relationships in 300 English sentences (Cook 1979) taken from Hemingway. A maximum of three

PROPOSITIONAL				
Agent	Experiencer	Benefactive	Object	Locative

MODAL			
Time	Manner	Instrument	Cause
Result	Purpose	outer Locative	outer Benefactive

Table 17: Propositional and modal cases from Cook (1989)

propositional 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. A novel aspect of Cook's work is a discussion of 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.

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 (86).

(86) *[Kettleman* $_{subj/Agent}$*] cut the bagel.*
[Fresh bagels $_{subj/Object}$*] cut easily.*
[Sharp knives $_{subj/Instrument}$*] cut well.*

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* (Table 18). Van Valin's Role and Reference Grammar theory holds that it is this ranking along with verb specific preference rules that allows humans to assign thematic relations to verb arguments.

macroroles	Actor					Undergoer
thematic relations	Agent	Effector	Experiencer	Locative	Theme	Patient

Table 18: Macroroles and thematic relations from van Valin (1993)

Sparck Jones & Boguraev (1987) avoid the arbitrariness of case lists based on intuition or assumptions about mental models. Instead, they define a set of cases based on language data. After deciding that their target set should consist of twenty to thirty 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 are defined only in terms of the example sentences from which they were derived. Without definitions, the distinction between similar cases is not always obvious. Sparck Jones and Boguraev also do not take into consideration the different syntactic functions of prepositions. Since prepositions do not always mark relationships between verbs 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 predicating adjectives. Finally, there is no evaluation of 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.

Slator *et al.* (1990) derive semantic classes based on a study of the twenty-five 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 (conventional) dictionary definitions. The classes are then given names 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, intuition-based 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 fifteen classes that appear to cover locative relationships. Nonetheless, the similarity between these purely empirically derived classes and case roles common to other systems lends empirical support to the kinds of semantic relationships intuited 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 that postmodify nouns. The output is 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 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 the Conceptual Graph notation (Sowa 1984) and appears to use common Conceptual Graph relations.

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: 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 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 & 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 Spark Jones & Boguraev's (1987) temporal cases, and also map roughly to the localist Path, Source, Goal and Local categories. Although Pustejovsky and Busa do not use the term *cases*, they show a mapping from prepositional (and other lexical) markers to the four time objects. For example, *in, throughout, for, late, early, beginning* of and *end of* mark the During object; the Before object is marked by *before, ending, until* and *by*.

3.2 Case Markers

A case marker is a syntactic or lexical indicator of a case in a sentence. HAIKU recognizes three types of case marker: positional, prepositional and adverbial.

3.2.1 Positional Markers

Positional markers mark a case by their syntactic position in the clause. They are the syntactic subject, direct object and indirect object of the verb. In this dissertation I use the symbols `psubj`, `pobj` and `piobj`, which refer to *positional subject, positional direct object* and *positional indirect object*. 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.

3.2.2 Prepositional Markers

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: 9.1). A preposition expressing a relationship between

its complement and a verb marks a case. 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 a noun phrase (as opposed to a verb) does not mark a case. That preposition is attached to a noun and marks a noun modifier relationship (see section 4.5).

In example (87), the prepositional phrase *on the printer* is attached to the verb *printed*. The noun phrase *the printer* fills that verb's Instrument case. In (88), 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. The prepositional phrase *on his desk*, however, postmodifies the noun *printer* and fills a noun modifier relationship of the noun *printer*, not a case of the verb *printed*.

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

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

Because certain prepositions and conjunctions are homographic, conjunctions may seem to mark cases. In (89), the conjunction *after* appears to mark TimeAt.

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

A conjunction joining clauses does not mark a case however because it does not signal a relationship between a verb and one of its arguments. Instead, it marks a clause level relationship.

3.2.3 Adverbial Markers

Adverbials are the third type of case marker. The same semantic relationships that are expressed by prepositional markers can also often be marked by adverbials. In (90), *at* marks and *the same time* fills the TimeAt case; in the paraphrase in (91), *simultaneously* expresses the same case.

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

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

Based on the similar distribution of prepositional phrases and adverbials in denoting the circumstances of events, adverbials are valid case markers in HAIKU. An adverbial case marker both marks a case and fills it.

3.2.4 Marker Order

For case analysis of a given clause, HAIKU builds an ordered pattern of markers for the main verb. In the pattern, positional arguments appear as the symbols `psubj`, `pobj` and `piobj`; prepositional phrase arguments are represented by their prepositional marker; adverbial arguments are represented by the symbol `adv`.

`psubj` > `pobj` > `piobj` > *preps* > `adv`

This ordering reflects the unmarked SVO (subject-verb-object) word order in English as well as the notion of clause structure element centrality (Quirk et al. 1985: 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. Clause element centrality has an obvious relationship to the distinction between core and peripheral cases. Those cases assigned to markers in more central positions would more likely be considered core roles for a verb.

3.3 Case System Design

In this section I describe the process of defining HAIKU's cases. The process, though guided by English marker data, example sentences and test texts, is not purely data-driven: the initial case set was inspired by cases from existing systems. Nonetheless, the role of language data in the evaluation of the cases grounds the set in linguistic fact, not just in intuition.

The first step in the process was to find a rough union of others' case systems (such as those described in section 3.1.4). Any case appearing in any other list was a candidate for inclusion in HAIKU's set. Only one case was selected among every strongly similar group of cases, even if they had different names in different systems. For example, HAIKU has a single Object case to account for cases variously named Affected, Object, Patient and ThingTransferred. Also ignored are cases not representing relationships between verbs and their arguments, such as the Attribute and Possessed-By cases in Sparck Jones and Boguraev (1987), since those non-case relationships are covered by the clause level relationships or the noun modifier relationships.

The second step in construction was to build a comprehensive list of case markers and map them to the cases. The set of cases was adjusted according to the extent to which cases were represented by the markers in that list. The distribution of data in the marker list is not the same as in complete texts. In this sense, using the case marker dictionary to test the coverage of the cases was more a *diagnostic* evaluation of the case set. The evaluation in section 3.6 represents a *performance* evaluation. Criteria for both evaluations are discussed in section 3.3.2.

3.3.1 The Case Marker Dictionary

The case marker dictionary was built by searching electronic and conventional dictionaries[2] for preposition and adverb senses. The positional case markers (`psubj`, `pobj` and `piobj`) were also included in the case marker dictionary.

For the prepositions each distinct sense was examined to decide which cases were appropriate to that sense. If a sense corresponded to a case, an entry mapping the preposition to that case was added to the 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 necessary.

The source dictionaries were also consulted to find adverbial cases markers, such as those in (92) and (93).

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

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

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 the adverb 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 sample of the entries for prepositions). In addition to providing the user with an illustration of a given marker→case pair (during case analysis), the example sentences helped 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, they might be merged into a single more general one.

---

[2] The dictionaries I used most were the *COBUILD* (Sinclair 1991), the unabridged *Random House* (Stein 1983) and Grady Ward's *Moby* lists. I also consulted *LDOCE* (Summers 1987) and *Fowler's Modern English Usage* (Fowler 1984) to confirm and clarify marker senses.

Implementation Note

The case marker dictionary is compiled with HAIKU and there is no facility to load new mappings at run time. Nonetheless, adding new entries and new mappings to existing entries is a simple matter of adding Prolog facts to the dictionary in the form shown in Appendix II. Since HAIKU saves all marker→case assignments in a separate output file, new mappings resulting from a HAIKU session can simply be copied over to the case marker dictionary.

3.3.2 Evaluation Criteria

I now identify three desirable characteristics of a case system. These are certainly not the only possible criteria for evaluating cases, but they have proven to be useful for finding weaknesses in the way the case set covers empirical data (both exhaustively enumerated data, such as the case marker dictionary, and data from real texts).

Generality

Cases must generalize to some degree the role of arguments of a verb. Individual verb-argument pairs themselves could not be used as cases, since they would be nothing more than a unique label for the relationship and would add nothing to its interpretation. For example, a BreakWindow case assigned to the direct object in (94) would not add any information to the interpretation of that sentence; and it would not account for the similarity of the roles played by the direct objects in (94), (95) and (96).

(94) Alphonse broke [the window $_{pobj}$]

(95) Alphonse broke [the pane $_{pobj}$]

(96) Alphonse shattered [the window $_{pobj}$]

At the other extreme, verb-argument pairs could be represented by just two roles corresponding to *participants* (P) and *circumstances* (C). This trivial case system would probably be too general for any natural language processing task since the designation of verb arguments as participants or circumstances is no more informative than the syntactic analysis of verb arguments as objects or adverbials.

(97) [Gaston $_{P}$] broke [the window $_{P}$] [yesterday $_{C}$] with [a rock $_{P}$].

The two extremes inspire the first criterion 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. A Participant case is too general to account for the different roles played by the subject, direct object and prepositional phrase in (97). On the other hand, cases should generalize the semantic relationships in a clause. That is, each case must account for relationships between distinct verb-argument pairs from different clauses. The BreakWindow case would account for only *one* verb-argument pair.

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. 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 I claimed was too general for (97) may be specific enough if the application is a template filling task to determine the entities involved in an act. If the target 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 fine distinctions between a number of locative cases.

Commonly, the overall level of generality of cases is not explicitly addressed by case system designers. 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 valuable in determining the practical generality of a set of cases, it would require that a large amount of text be analyzed to draw any significant 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 in texts than some cases with very restricted, specific definitions. Furthermore, the measure would change as more texts were analyzed, though more analysis would also increase confidence in its value.

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 texts, it is related to the first measure of a case's generality: the generality of a case using the first measure is bound from above by the frequency with which that case's 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 is easily calculated since the markers are usually fixed independent of any texts to be analyzed.

In Barker (1996) I investigate the degree to which each of HAIKU's cases is represented by markers in the case marker dictionary as well as in the *clouds* text. The report also discusses possible adjustments to the case set for cases that have marker representation above or below average. The main results from the investigation appear in section 3.6.2, where they are extended with results from the *small engines* experiment.

Completeness

A set of cases should cover any possible verb-argument relationship. 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. Generality, however, ensures that overly specific cases are not added to cover a single verb-argument relationship. A solution is to expand the definition of an existing case to account for a new relationship. The new definition should cover all of the relationships covered by the old definition along with the new relationship, and no others. Documentation of each case's coverage is therefore particularly important.

Uniqueness

A case system should contain no superfluous or redundant cases. A case is superfluous if the relationship it describes never appears in any text between a verb and argument. A case is redundant if every instance of it is also an instance of some other case in the set.

Empirically, a case can never be shown to be absolutely superfluous or redundant: the fact that a case has covered none of the encountered verb-argument pairs in texts 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.

Nonetheless, 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 semantic relationships covered by a case and by contrasting these with the definitions of other cases.

3.3.3 Using the Criteria to Guide Case Selection

The initial grain of HAIKU's case system was suggested by that of the existing systems from which cases were taken. The case markers further guided adjustment to the specificity of the 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 corresponding to the four localist categories. HAIKU's cases include Frequency in addition to the common TimeAt, Duration, Before and After, inspired by the prevalence of adverbial case markers for such a role: *hourly*, *daily*, *monthly*, etc. Frequency 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 & 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.

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, leaving their examples to be covered by the more general TimeAt case.

3.4 The Cases

The resulting set of cases (Table 19) was then compared against those in the systems described in section 3.1.4 to ensure that it had at least as much coverage as those systems.

The cases appear in five groups. The PARTICIPANT group consists of cases in which the argument is usually directly involved in the event expressed by the verb. The CAUSALITY group includes relationships enabling or opposing the event. SPACE and TIME group relationships that place the event 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.

The groups are more useful when talking about the cases than when using them. For example, the division into groups 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; TimeThrough is more likely redundant because its instances are covered by TimeAt than because they are covered by Agent.

3.4.1 Case Glossary

This section describes the meaning and coverage of each case. Each description is followed by one or more example sentences chosen to reflect typical usage or to illustrate the scope of each case.

Implementation Note

In practice, for any particular assignment the choice of case depends on the reader's interpretation of the sentence, especially for more abstract examples. HAIKU allows the user to invent new cases if finer grain is required for a text. These user-defined cases are stored in a separate file that can be loaded at the beginning of a new session. When assigning a case in a sentence the user can choose one from the original set or from the user-defined cases.

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)	

Table 19: The cases (with abbreviations)

Participant

Accompaniment (ACMP)

The Accompaniment case represents one or more entities accompanying another entity (usually the Agent, Experiencer or Object) involved in an event. Examples (98), (99) and (100) show accompaniment of Agent, Experiencer and Object.

(98) [I $_{AGT}$] eat supper with [my family $_{ACMP}$].

(99) [I $_{EXPR}$] live with [my family $_{ACMP}$].

(100) I eat [my peas $_{OBJ}$] with [honey $_{ACMP}$].

Agent (AGT)

The Agent case represents the initiator or performer of an event. 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 an act or actively participates in an event. The Agent is expressed by the syntactic subject of the verb or as the object of the preposition *by* in passive clauses.

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

Example (101) shows an Agent, *The database manager*. Although database manager modules 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 (102) the Agent would be *The analyst* and the *database manager* fills the role of Instrument.

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

Beneficiary (BENF)

The Beneficiary case represents the entity that benefits from the state resulting from the event. The state 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 event.

(103) I wrote [Smilla $_{piobj/BENF}$] a reference letter [to prospective employers $_{pp/RECP}$].

(104) I wrote [Smilla $_{piobj/BENF}$] [a reference letter [to prospective employers $_{pp}$] $_{pobj/OBJ}$].

(105) This year's rains produced a bumper crop [for the farmer $_{pp/BENF}$].

Example (103) shows the indirect object (*Smilla*) filling the Beneficiary case with *prospective employers* filling the Recipient case. In (104) there is no verb argument to fill the Recipient case (*to prospective employers* is attached to *letter*, not *wrote*). Nonetheless, since the writing was done to the advantage (or disadvantage) of *Smilla*, Beneficiary is again an appropriate case. Example (105) illustrates the fact that the filler of the Beneficiary case need not be the *intended* beneficiary of the event. Clearly, the production of a bumper crop for the farmer's benefit was not intentional.

Exclusion (EXCL)

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

(106) Everybody slept except [Jeff $_{EXCL}$].

(107) Jeff played instead of [Howard $_{EXCL}$].

Experiencer (EXPR)

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

(108) [Ajax $_{psubj/EXPR}$] is sleeping.

Instrument (INST)

The Instrument case represents an entity that is applied or employed in an event. The Instrument for events of transfer is often the medium of the transfer.

(109) He broke the window with [a brick $_{INST}$].

(110) The system administrator notified the users via [email $_{INST}$].

In (109) *a brick* is the entity applied to accomplish the event of breaking the window. In (110) *email* is the medium of transfer of notifying the users.

Object (OBJ)

The entity directly acted upon by the event 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.

(111) Ethel printed [the file $_{pobj/OBJ}$].

(112) [The window $_{psubj/OBJ}$] broke.

(113) They stripped him [of his pride $_{pp/OBJ}$].

Example (111) has the syntactic direct object (*the file*) filling the Object case; (112) shows the less common situation of the syntactic subject (*the window*) filling this case; (113) is an example of the rare occasion where the Object case is marked by the preposition *of*.[3] *His pride* is the entity that has been stripped while *him* is the location from which it was stripped.[4]

---

[3] Object marked by *of* is rare but not merely idiomatic. Similar examples using different verbs support the decision to treat *of* as a valid marker for Object: *The deprived him [of his rights $_{pp/OBJ}$]*; *They partake [of the bread $_{pp/OBJ}$]*.

[4] Example (113) has a person as the abstract location of a feeling. An alternative to allowing both abstract and concrete locations would be to refine each locative case into two separate cases distinguished by the feature `±concrete`.

Recipient (RECP)

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

(114) I sent [the prospective employers $_{piobj/RECP}$*] a reference letter for Smilla.*

(115) I wrote [Smilla $_{piobj/BENF}$*] a reference letter [to prospective employers* $_{pp/RECP}$*].*

In (114) the indirect object of the verb fills the Recipient case. In (115) the indirect object fills the Beneficiary case while the Recipient case is marked by the preposition *to* (see the definition of Beneficiary).

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*.

Causality

Cause (CAUS)

The Cause is the state or event that makes another event take place. The Cause case may also represent an environment that allows an event to be performed or a state to exist.

(116) He died of [thirst $_{CAUS}$*].*

(117) We acted on [his advice $_{CAUS}$*].*

Effect (EFF)

The Effect case represents a state that is the outcome of an event or the result of another state.

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

Opposition (OPP)

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

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

Purpose (PURP)

The Purpose case represents the state intended to result from the event. This case implies initiation of the event 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.

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

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 event of motion to be an arrow. The tail of the arrow represents the starting point of the event (LFRM). The head represents the destination of the event (LTO). The direction of the arrow corresponds to the event's bearing in space (DIR). The shaft joining the tail to the head corresponds to the space through which the event passes or extends (LTRU).[5]

(121) I drove [east $_{DIR}$] on [autoroute 640 $_{LTRU}$] from [Laval $_{LFRM}$] to [Repentigny $_{LTO}$].

(122) Look inside [yourself $_{DIR}$] for the answer.

(123) They stripped [him $_{LFRM}$] of his pride.

(124) Strange images flitted through [my mind $_{LTRU}$].

(125) His life is going [nowhere $_{LTO}$].

Example (121) shows all four cases together in concrete (physical) usage. Examples (122) to (125) 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 event occurs or a state exists. This case applies to non-motion events that occur statically in space, or to specific points along the path of a motion event. 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.

(126) We stopped at [a restaurant $_{LAT}$] on our way home.

(127) Memory of the accident was hidden in [his subconscious $_{LAT}$].

In (126), a restaurant is a specific point along the path of motion; (127) illustrates an abstract usage of the LocationAt case.

---

[5] In the localist framework, LocationFrom, LocationThrough, LocationTo and LocationAt correspond to Source, Path, Goal and Local. There is no direct analogue for Direction.

Orientation (ORNT)

The Orientation case represents the bearing an entity involved in an event in terms of the 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.

(128) The tower lay on [its side $_{ORNT}$*].*

(129) The tree stood [erect $_{ORNT}$*] despite the heavy ice.*

Time

Frequency (FREQ)

The Frequency case represents the rate at which an event or state recurs.

(130) Buses arrive [in five minute intervals $_{pp/FREQ}$*].*

(131) Jennifer washes her truck [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 event takes place or a state exists. It is the temporal analog of the spatial case LocationAt. 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 another event indicating when the event took place.

(132) He traveled extensively [last year $_{TAT}$*].*

(133) He made his fortune during [his visit to Europe $_{TAT}$*].*

The case filler in (132) is a time unit with measurable extent. In (133) the nominalized event *his visit to Europe* indicates the time when the event (making his fortune) occurred.

TimeFrom (TFRM)

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

(134) He has been blind since [the war $_{TFRM}$*].*

(135) I will stay at your house from [Tuesday $_{TFRM}$*] to Saturday.*

In (134) *the war* marks the time at which the state began. *He* has been blind from the war until now. In (135), 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 event or a state. In contrast to TimeAt, the filler of this case must have extent.

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

TimeTo (TTO)

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

(137) We worked from nine to [five $_{TTO}$].

(138) I will stay at your house until [Tuesday $_{TTO}$].

Quality

Content (CONT)

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

(139) John wrote about [volcanoes $_{CONT}$].

(140) Albert is concerned about [the economy $_{CONT}$].

(141) He filled the container with [milk $_{CONT}$].

Manner (MANR)

The Manner case represents a way in which an event is performed or takes place. This case accounts for many common relationships between a verb and its arguments that describe qualitative characteristics of the event or state. Manner 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.

(142) Emily writes [with style $_{pp/MANR}$].

(143) Michael played [beautifully $_{adv/MANR}$].

Material (MATR)

The Material case represents the physical substance composing an entity involved in an event.

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

Measure (MEAS)

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

(145) Secretariat won by [three lengths $_{MEAS}$].

(146) I bought the car for [five hundred dollars $_{MEAS}$].

Order (ORD)

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

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

3.5 Assigning Cases

The case analyzer 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. The technique is reminiscent of case-based reasoning, with previous CMPs and verbs comprising the base of cases.[6]

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 most similar previous marker pattern CMP_s among previous clauses. The measure of CMP distance for the similarity calculation is defined in Delisle *et al.* (1993). Case patterns previously assigned to CMP_s contain candidate cases for the current CP.

If the system cannot 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

[6] I will not make any other references to case-based reasoning; the term *case* is already overloaded.

definitions of the cases (section 3.4.1) to guide case selection. Once the user has accepted a suggested CP or supplied one, 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.

The implicit assumption in using previous case analyses to guide assignment is that the cases are generalizations that can apply to different verb-marker pairs. The number of times previous case patterns suggested to the user are accepted testifies indirectly to the generalizing power of the cases themselves.

3.6 Evaluation

I have so far described the construction of a set of cases driven by English case marker data from dictionaries and word lists. Evaluation of the set was accomplished by using HAIKU in the case analysis of sample texts.

3.6.1 Case Analyzer Evaluation

I claimed in section 3.3.2 that the evaluation of a case system includes its coverage of clauses from real texts. This section briefly presents the results of case analysis of the *clouds* text and the *small engines* text.

The *clouds* text contains 512 sentences from which DIPETT parsed 439 finite, non-stative clauses usable for case analysis. The system assembled 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 above. Over all 439 clauses, the system made on average 4.5 CP suggestions per CMP. 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. Figure 4 shows the number of clauses for which the correct pattern was among the system's suggestions (*system*) contrasted with the number that had to be supplied by the user (*user*).

In the *small engines* experiment, 584 clauses received case analysis. The system made on average 4.4 suggestions for clauses for which it could make any suggestions at all. After fewer than a hundred clauses, the number of system assignments overtook the user assignments for good. By the end of the experiment, the system had suggested a correct case pattern for 66% of the assignments. Figure 5 plots the cumulative number of assignments made by the user against the cumulative number of correct assignments

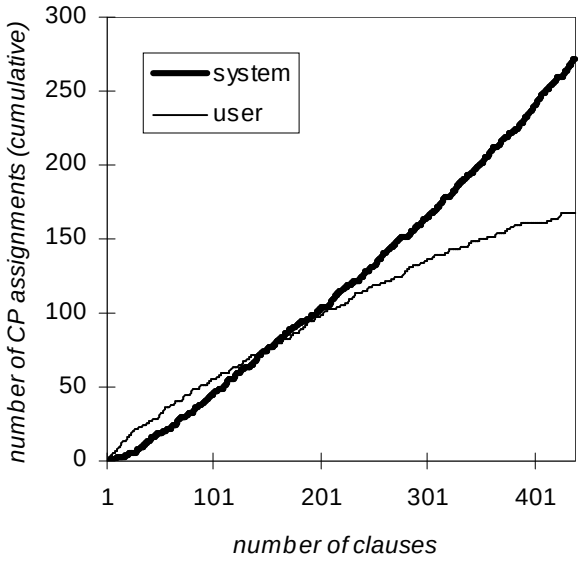

Figure 4: CP assignments in the clouds *experiment*

among the system's suggestions. The similarity of the trends in Figure 4 and Figure 5 supports the contention that the case analyzer can learn to make better suggestions to the user.

In order to evaluate the proportion of clauses in the text that received a case analysis, I sampled 100 verbs at random from the *small engines* text, determined their case marker

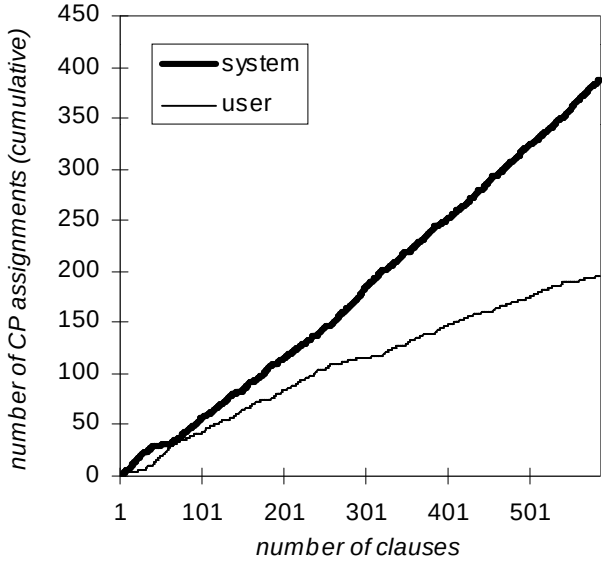

Figure 5: CP assignments in the small engines *experiment*

pattern manually and found that 97 of the sampled verbs and their correct CPs appeared in HAIKU's output. With 95% confidence HAIKU extracted between 91.5% and 98.9% of the case patterns in the text. This high coverage is due to the fact that the user is given the opportunity to correct case marker patterns that are incorrect due to misparses.

Finally, for both the *clouds* and *small engines* experiments, a user onus rating was recorded for each user interaction (see section 1.4.4). For the *clouds* experiment user onus was 0 (lowest) for 384 interactions (87%). For 50 interactions (11%) the onus was somewhat higher (1) and for only five interactions (1%) the onus was rated 2. No interactions were rated onus level 3. In the *small engines* experiment the corresponding onus ratings were: 480 onus 0 interactions (82%), 89 onus 1 interactions (15%) and 15 onus 2 interactions (3%). Again, no interactions were rated onus 3.

No new cases were needed to capture the semantic relationships in either experiment. On the other hand, four cases were never assigned in the *clouds* experiment and five were never assigned in the *small engines* experiment. Section 3.6.2 looks at these cases more closely to determine if they might be superfluous or redundant.

3.6.2 Case System Evaluation

This section revisits the case system evaluation criteria from section 3.3.2. The evaluation is based on the results of the *clouds* and *small engines* experiments and on a survey of the representation of cases among the case markers (Barker 1996). Figure 6 shows the distribution of each case with respect to the prepositional and adverbial markers as well as among the cases assigned in each of the two experiments.

Each case *C* has associated with it four bars in Figure 6. From top to bottom, the bars represent: 1) the percentage of case assignments in the *small engines* experiment accounted for by *C*; 2) the percentage of assignments in the *clouds* experiment accounted for by *C*; 3) the percentage of adverbial marker→case mappings in the case marker dictionary accounted for by *C*; 4) the percentage of prepositional and positional marker→case mappings in the case marker dictionary accounted for by *C*.

For example, Direction accounted for only 0.7% of the cases assigned in the *small engines* experiment and 2.6% of those in the *clouds* experiment, but it is well represented by the markers (20.0% of adverb→case mappings; 8.9% of prepositional or positional mappings). Object is weakly represented by the markers (1.2% of the prepositional or positional mappings and 0.0% of the adverb mappings), yet it accounts for 24.8% of all case assignments in *clouds* and 39.1% of all case assignments in *small engines*.

Generality

In section 3.3.2 I proposed two measures of 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 (*i.e.*, markers) accounted for by each case.

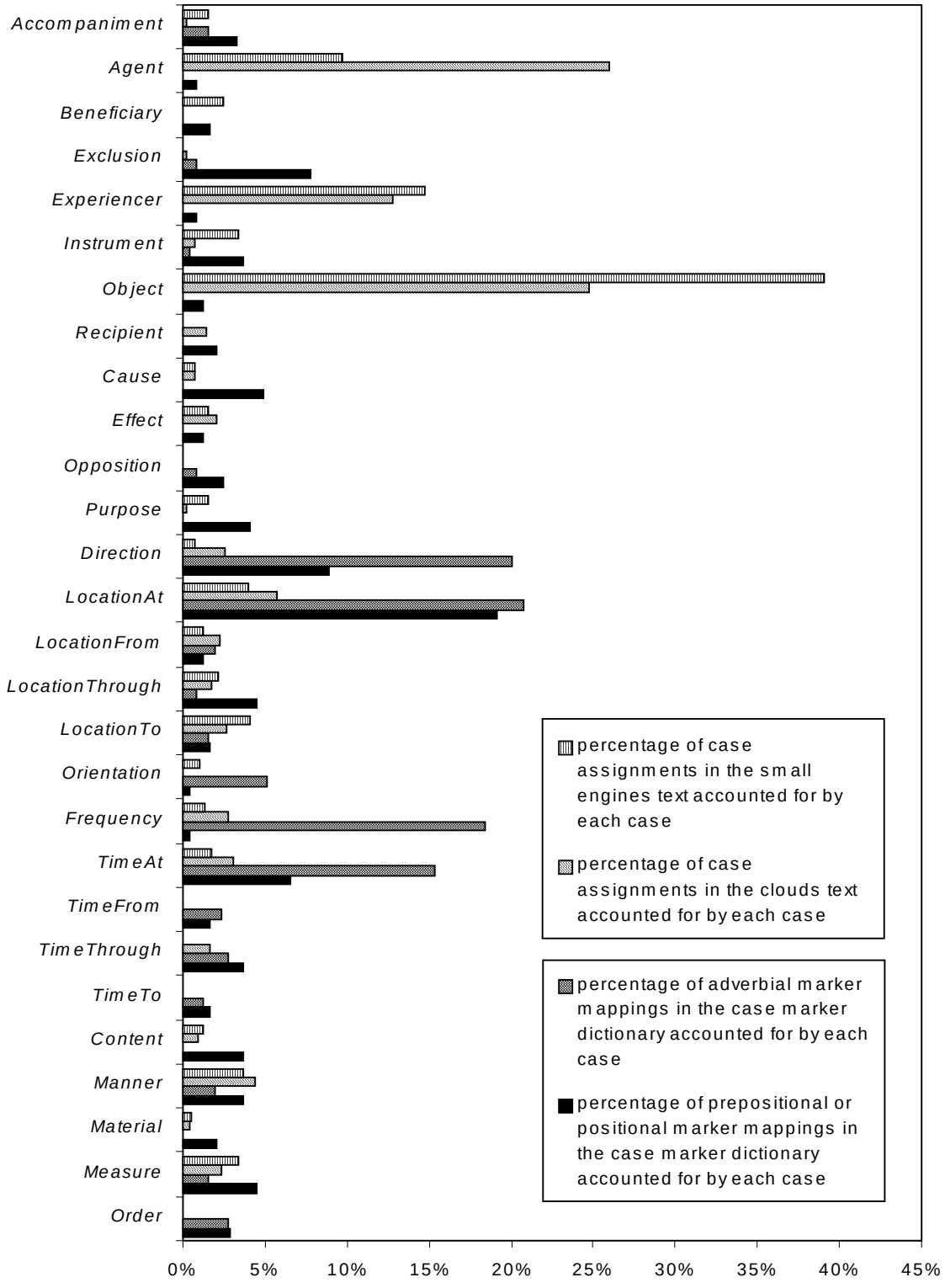

Figure 6: Distribution of the cases in the clouds *and* small engines *experiments, and among the markers*

The *clouds* and *small engines* experiments cannot be considered "a large body of text" for the first measure. Two texts are insufficient for drawing general conclusions about the applicability of each of the cases. For example, below average representation of a particular case in two texts does not necessarily mean that the case is significantly over-specific. Nonetheless, the experiments assigned case patterns to 1023 finite clauses for a total of 2043 individual case assignments in two quite different (though both technical) domains. The data from these assignments are a useful start when investigating the completeness and uniqueness of HAIKU's case set.

The marker→case mapping data charted in Figure 6 can be 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 may be candidates for combination into a more general case if they are semantically related. Cases whose marker representation is above average are potentially over-general and may be candidates for splitting into more specific cases.

Effect and Purpose are cases with similar definitions and are both weakly represented by the markers (0.6% and 2.0% of all marker→case mappings respectively). These two cases could be merged into a single more general case. They differ in that Effect describes a `+realized` outcome that is `±intended`. Purpose represents a `+intended` outcome that is `±realized`. For applications with no need for such distinctions, a single case might suffice. Effect and Purpose are also weakly represented by the experimental data.

LocationAt and TimeAt are candidates for splitting into more specific cases based on their large number of markers (20.0% and 11.0% of all marker→case mappings respectively). Their representation among the case assignments in the experiments is not nearly so high, with LocationAt accounting for 5.7% of the *clouds* assignments and 4.0% of the *small engines* assignments; TimeAt 3.1% for *clouds* and only 1.8% for *small engines*. One possibility would be to split LocationAt and TimeAt on the `±concrete` feature, though doing so would result in inconsistency with the other locative and temporal cases, whose abstract and concrete interpretations are rolled into individual cases.

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 (*clouds*: 26.0%, 12.7% and 24.8%; *small engines*: 9.7%, 14.7% and 39.1%). The frequency of these cases in texts is not very surprising, since they are most often assigned to required verb arguments.[7]

There are also 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. Yet Content accounts

[7] Recall that in Cook's Matrix Model (Cook 1989), Object is a required case in all case frames.

for only 1.8% of the marker→case mappings and 0.9% and 1.2% of the *clouds* and *small engines* assignments. Since the two different interpretations seem unlikely to appear together in the same sentence, the Content case might be left as is until more data justify a split.

Completeness

As claimed in section 3.3.2, a practical measure of the completeness of a case set is the number of verb-argument relationships in a text covered by only the cases in the set.

The *clouds* and *small engines* experiments assigned 2043 individual cases in 1023 clauses and did not uncover a verb-argument role not covered by HAIKU's twenty-eight cases. These and other smaller experiments continue to increase the confidence in the completeness of the set.

Uniqueness

The definition of uniqueness from section 3.3.2 requires two things: that the case set have no redundant cases and that it have no superfluous cases.

The presentation of careful case definitions in 3.4.1 is an attempt to avoid redundant cases. The 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 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.

Example sentences were useful for improving the uniqueness of the case set. In addition to the examples given for each definition in the glossary, an example sentence is stored with every entry in the case marker dictionary (see Appendix II). These example sentences were meant 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, HAIKU's users have not encountered this problem using the existing set of cases. Whenever there has been disagreement among users about which case to assign, it has resulted from ambiguity in the sentence, not redundancy in the case set. Once a single interpretation of an ambiguous sentence has been chosen the choice of cases is unambiguous. For example, (148) has two

interpretations depending on whether the noun phrase *various tools* fills the Instrument case or the Material case.

(148) Marty makes sculptures with [various tools $_{????}$].

The ambiguity is not due to any redundancy between Instrument and Material. If the intended meaning of (148) is that the sculptures themselves consist of some arrangement of tools, then *various tools* unambiguously fills the Material case. If the intended meaning is that the tools are used to construct the sculpture (out of wood and stone, for example), then the Instrument case is appropriate.

Note that if two cases can appear in the same clause (as Instrument and Material do in (149)), they are not redundant since a single case may not appear twice in the same clause.

(149) Marty makes sculptures out of [wood and stone $_{MATR}$] with [various tools $_{INST}$].

Superfluous Cases

Any general purpose, domain independent case system may include cases that are superfluous to the analysis of a particular text. In the *clouds* experiment four cases were not used: Opposition, Order, TimeFrom and TimeTo. The absence of a case in a single text is not evidence that the case is superfluous. In *small engines*, however, *the same four cases were never assigned* (Recipient was also not used). In fact, the temporal cases were in general weakly represented by the data in those experiments.

From its definition, Opposition represents an entity that contrasts with or opposes an event but is insufficient to prevent it from happening. Despite the fact that this case was not assigned in *clouds* or *small engines*, the only other case that might capture the relationship marked by appropriate senses of such markers as *against*, *considering*, *despite*, *in spite of*, *notwithstanding*, *versus*, *nevertheless* and *nonetheless* is Exclusion. (150) shows both Exclusion and Opposition in the same sentence.

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

Traditional case theory requires that no case appear more than once in a given clause and therefore prevents the merging of Opposition and Exclusion. Exclusion, though, is also weakly represented, with only two assignments in *clouds* and one single assignment (out of more than a thousand) in *small engines*. If Opposition is not theoretically superfluous, it has certainly proven to be practically superfluous. Opposition and Exclusion may well be candidates for merging.

Order represents the relative position of an entity within a structured arrangement of entities involved in an event, 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 (151) contains instances of both LocationAt and Order.

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

The other two cases not used in either experiment were TimeFrom and TimeTo. These are the natural temporal counterparts of the locative LocationFrom and LocationTo cases, both of which appeared frequently in the texts. The TimeFrom and TimeTo cases are also fairly well represented by the case markers (2.0% and 1.4% of marker→case mappings).

Finally, the absence of Recipient in *small engines* suggests that this case might be over-specific, though it did occur fourteen times in *clouds* (1.5%). In case systems that do not contain a Recipient case, the relationship is often covered by some Goal case (Somers 1987). HAIKU's case set does not include a generic Goal case.

3.7 Chapter Summary

This chapter described the construction and evaluation of HAIKU's list of cases. The first step in choosing cases was to reconcile existing lists into a consistent amalgam. A dictionary mapping positional, prepositional and adverbial markers to the cases assists case analysis, but it has also helped expose weaknesses in the case list. The cases were further evaluated on the criteria of generality, completeness and uniqueness using two measures: the number of occurrences of each case in sample texts and the number of marker→case mappings in the case marker dictionary. Evaluation identified cases that might be over-general or over-specific. For the *clouds* and *small engines* experiments, HAIKU's case list was complete, that is, no new cases were required for those texts. The cases were also seen to be unique in that no case's markers are a proper subset of the markers of some other case and there were no case assignments in the experiments for which more than one case was deemed appropriate.

In this chapter I also reported on performance evaluations of the case analyzer. Experiments showed that as more sentences are analyzed HAIKU learns to make better case pattern suggestions to the user. The average and maximum number of suggestions presented to the user support the claim that case analysis is not overly burdensome. The claim is further supported by the low average onus rating for CA interactions (0.17 average over the *clouds* and *small engines* experiments).

This chapter described the semantic relationships within clauses between a verb and its syntactic arguments, many of which are noun phrases (or contain noun phrases). In the next chapter I go inside the noun phrases to investigate the relationships between a noun and its modifiers.