THE USE OF SYNTACTIC CLUES IN DISCOURSE PROCESSING

Nan Decker
1834 Chase Avenue
Cincinnati, Ohio 45223, USA

ABSTRACT

The desirability of a syntactic parsing component in natural language understanding systems has been the subject of debate for the past several years. This paper describes an approach to automatic text processing which is entirely based on syntactic form. A program is described which processes one genre of discourse, that of newspaper reports. The program creates summaries of reports by relying on an expanded concept of text grounding: certain syntactic structures and tense/aspect pairs indicate the most important events in a news story. Supportive, background material is also highly coded syntactically. Certain types of information are routinely expressed with distinct syntactic forms. Where more than one episode occurs in a single report, a change of episode will also be marked syntactically in a reliable way.

INTRODUCTION

The role that syntactic structure should play in natural language processing has been a matter of debate in computational linguistics. While some researchers eschew syntactic processing as giving a poor return on the heavy investment of a parser (Schank and Riesbeck, 1981), others make syntactic representations the basis from which further work is done (Sager, 1981; Hirschman and Sager, 1982). Current syntax-based processors tend to work only within a narrow semantic domain, since they rely heavily on word co-occurrence patterns which hold only within texts from a particular sublanguage. Knowledge-based processors, on the other hand, can operate on a less restricted semantic field, but only if sufficient knowledge in the form of scripts, frames, and so forth, is built into the program.

This paper describes a syntactic approach to natural language processing which is not bound to a narrow semantic field, and which requires little or no world knowledge. This approach has been demonstrated in a computer program called DUMP (Discourse Understanding Model Program), which relies solely on syntactic structure to create summaries of one particular genre of discourse--that of newspaper reports--and to label the kinds of information given in them (Decker, 1985). The process for creating these summaries differs substantially from the word-list and statistical methods used by other automatic abstractor programs (Borko and Bernier, 1975). The DUMP program therefore depends on a predictable discourse genre or style, rather than a predictable sublanguage lexicon or body of world knowledge.

DUMP was developed from a corpus of over 5800 words representing twenty-three news reports from three daily newspapers: the New York Times, the Boston Globe, and the Providence Journal/Evening Bulletin. With one exception, each story appeared in the upper right-hand column of the front page. The stories in the corpus were chosen randomly and the only criterion for rejection was too large a percentage of quoted material. Only the first two hundred words or so of each story were included in the corpus in order to allow a greater sampling of reports. The discourse principles at work are fairly represented in an excerpt of this length.

The input to the DUMP program consists of a list of hand-parsed sentences making up each story. Ideally, these parse trees should be the output of a parsing program. In fact, about one-third of the sentences were passed through the RUS parser (Woods, 1973). RUS experienced difficulty with some of these sentences for a number of reasons: the parser was operating without a semantic component, and arcs from nodes were ordered with the expectation of feedback from semantics; RUS lacked some rules for structures which appear with regularity in the news; it attempted to give all the parses of a sentence, where DUMP only required one, and that not necessarily the correct or complete one (about which more later); and DUMP's rules call for certain syntactic labels which are not ordinarily assigned by parsing programs (negative and adversative clauses, for example). However, it should be stressed that none of these difficulties represents parsing problems of theoretical import. All could be resolved by extensions to existing components of the ATN and its dictionary.

THE DISCOURSE STRUCTURE OF NEWS REPORTS

The syntactic rules used by DUMP work because of the predictable, almost formulaic discourse structure of hard news reports*. Two journalistic devices above all else characterize hard news: the inverted pyramid, and the block paragraph (Green, 1979). The inverted pyramid refers to the convention of relating the most important facts of

* Features, sports reports, and so forth have their own discourse structure.

a news story in the first paragraph, followed by less important information given in descending order (or, it may be argued, random order) of importance. Thus, the news differs markedly from canonical story form in which material is given in chronological order. The block paragraph, the second device, is one which stands independent of paragraphs adjacent to it. This unit contains no logical connectives (however, in addition, moreover) which link it to preceding or following paragraphs. The avoidance of such connectives allows the newspaper editor to quickly delete paragraphs from a story in the morning edition to fit into the evening edition without rewriting. The block paragraph is short: over sixty percent of the paragraphs in the corpus are only one sentence long; about one-half have two sentences, and less than one percent have three sentences. The effect is that most sentences of the report are presented at the same level of importance: there is no orthographic unit larger than the sentence which reliably indicates that a group of sentences is related topically or episodically. In place of the normal paragraph, we shall see, is a highly reliable level of syntactic coding which links sentences into episodes.

At a lower level of organization than the inverted pyramid and block paragraph are the two discourse units which DUMP relies on: the episode, and within the episode, the information field as found in the detached clause.

News reports may contain more than one episode. A new episode begins when the set of characters and/or setting (temporal or geographical) changes.

The detached clause is defined intonationally: it is bounded by pauses, has falling intonation at the end, or is preceded by a clause with falling intonation (Thompson, 1983). This clause is almost always set off in text with commas. So, for example, the following sentence from the ninth story in the corpus ("Arafat Forces Lose Key Position," Boston Globe, November 7, 1983) consists of four detached clauses, or information fields:

> (9:3)* Arafat's soldiers, who resisted the assault, fell back six miles to Beddawi, the remaining PLO stronghold in the area, and Nahr el Bared is now surrounded by Syrian soldiers....

The information fields here are: a nonrestrictive relative clause ("who resisted the assault"), an appositive ("the remaining PLO stronghold in the area"), and two main clauses ("Arafat's soldiers fell back..." and "Nahr el Bared is now surrounded...").

There are a small number of syntactic forms which reliably indicate the beginning of new episodes. Likewise, there is a strong correlation between the category of information the journalist conveys in each detached clause and the syntactic structures used for its expression. For example, the nonrestrictive relative clause in 9:3 expresses background events, the appositive expresses an identification of place, and the two main clauses express a main event and a current state, respectively. The next two sections will look at the syntactic correlates of the information field and the episode boundary in detail.

Syntactic Correlates of the Information Field

The syntactic rules used by DUMP reflect grounding principles found universally in discourse (Grimes, 1975). Certain assertional structures in text deliver foreground information, which tells the events of the narrative and moves the story forward. These events comprise a summary of the story. Less assertional structures are used to express background, supportive information which fleshes out the skeleton provided in the foreground but does not move the action forward. There is a strong correlation between the syntactic form and information type of this supportive material which allows DUMP to subcategorize it into the following classes: past events and processes leading up to the most recent development in the story; plans for the future; current state of the world; information of secondary importance; identifications; import of the story; effects of actions; comments made by participants in the story; and collateral (things which did not happen).

This division of material into foreground vs. background gives text its texture. A narrative in which everything is presented at the same level of prominence tends to be monotonous. One of the chief means of distinguishing foreground from background is tense and aspect, which has been called a sort of flow-of-control mechanism, allowing the reader to pick out the most important parts of a discourse (Hopper, 1979). Sentences with simple past verbs in the active voice are the chief conveyors of foreground material in news. This fact recalls the broader concept of transitivity put forth by Hopper and Thompson (1980), whereby certain properties of the verb and its arguments transfer the action from agent to patient more effectively than others. Foregrounded clauses have high transitivity, backgrounded clauses low transitivity.

High transitivity verbs are kinetic, telic, punctual, volitional, affirmative, and realis. Kinetic verbs allow easy transfer of action from subject to object. Throw is therefore kinetic, while the copular to be is not. Telic verbs are those which express an action with a natural endpoint. The verb make in "John is making a chair" is telic, while the verb sing in "John is singing" is not. Telic and atelic verbs can be distinguisned by their entailments: if John is interrupted while making a chair, it is not true that he has made a chair, but if he is interrupted while singing, it is still true that he has sung (Comrie, 1976). Punctual verbs (sneeze, kick) refer to actions with no obvious internal structure. Study and carry are examples of non-punctual verbs.

---

* The first number indicates the story in the corpus, the second the number of the sentence within that story.

Volitional verbs ("I wrote his name") have greater transitivity than non-volitional verbs ("I forgot his name")(Hopper and Thompson, 1980, p. 252). Affirmation distinguishes collateral information from all other types. And finally, the realis mode distinguishes events which have existed from those which only might have or would have. Main event clauses therefore never contain modals. The differential behavior of verbs from these semantic classes has been described by a number of taxonomers (Comrie, 1976; Mourelatos, 1981; Ota, 1963; Vendler, 1967).

Arguments high in transitivity are those which are strong agents, totally affected and highly individuated. Strong agents are human rather than non-human: "George startled me" has more transitivity than "The picture startled me" (Hopper and Thompson, 1980, p.252). Objects which are wholly affected lend greater transitivity than those which are only partially affected ("I drank the milk" vs. "I drank some milk"). Likewise, more highly individuated objects, defined as proper, human or animate, concrete, singular, count and definite, add more transitivity than less individuated ones.

These transitivity parameters assume a good deal of semantic knowledge about verbs and their arguments. In fact, the affirmative and realis features are the only ones reflected in DUMP's rules. But in another respect, Hopper and Thompson's notion of transitivity must be extended. An examination of tense and aspect alone is not sufficient to distinguish foreground from background in the DUMP corpus. The type of clause in which the verb appears is also crucial. So, for example, the simple past may be used to convey both foreground and background material, depending on the type of clause in which it occurs: in main clauses, it will always convey the most recent events in a story, while in relative clauses, it will always convey past events. The first two sentences of story 6 ("Stone Meets with Salvador Rebel Official," Boston Globe, August 1, 1983) illustrate the distinct uses of the two clause types.

> (6:1) After weeks of maneuvering and frustration, presidential envoy Richard B. Stone met face-to-face yesterday for the first time with a key leader of the Salvadoran guerrilla movement.

Here, the simple past is used in a main clause to foreground information.

> (6:2) "The ice has been broken," proclaimed President Belisario Betancur of Colombia, who engineered the meeting.

The simple past engineered in a relative clause indicates background material.

The information-bearing capacities of these two clause types, when they occur with the simple, active past, are in complementary distribution in newswriting. The main clause is more assertional than the relative clause; it is used to give information which the writer assumes the reader is seeing for the first time. The relative clause, on the other hand, is more presuppositional. The writer uses it to convey old information which is of lesser importance or which the reader may already have knowledge of.

Sentences 6:1 and 6:2 illustrate the way in which syntactic forms provide information which might otherwise need to be culled from world knowledge. We know that the planning of a meeting precedes its occurrence, but no such knowledge is necessary here, since the past verb form in a relative clause signals an event which occurred before the main event.

The so-called "hot news" present perfect in a main clause ("The president has resigned") signals a main event if it occurs in the first sentence of a story. Its appearance further down or in a non-main clause signals information about past events or states. Two sentences from story 16 ("Peronists Suffer Stunning Defeat in Argentine Vote," New York Times, November 1, 1983) illustrate this.

> (16:1) The leader of a middle-class party has swept to victory in Argentina's presidential elections....
>
> (16:4) The election, called by the ruling military, was a stunning defeat for the Peronists, who have dominated Argentina's political life since their party was founded in 1945 by Juan Domingo Peron.

In 16:1, the present perfect has swept is used in the hot news sense. In 16:4, the present perfect have dominated is used in a relative clause with an adverbial phrase ("since their party was founded in 1945...") to describe a state that has existed for decades. Note also that the verb dominate is atelic and non-punctual, and therefore low in transitivity. However, knowledge of the verb's semantic class is not necessary to identify the relative clause as supportive. The mere fact that the verb is in a relative clause or the fact that the present perfect appears after the first sentence suffices.

Syntactic clues may be used to avoid the need for time programs which determine the relative timing of events by interpreting adverbials. The following main clauses use the present perfect, but since they are non-initial, the states and events referred to in them must have occurred before the main event in the story ("O'Neill Now Calls Grenada Invasion 'Justified' Action," New York Times, November 9, 1983).

> (19:5) Pressures to pass a strict 60-day legal limit [to the stay of U.S. troops in Grenada] have eased in the past week.
> (19:6) Both houses have passed such measures, but the Senate version has been bottled up because it was attached to a debt-ceiling bill.
> (19:7) Other versions of the 60-day War Powers Resolution have been introduced but not acted upon.

The appearance of the present perfect this far

into the story means that the time phrase in the past week does not have to be interpreted by a time program.

Likewise, the use of the passive simple past in a main clause indicates that the event is supportive material: main events, it turns out, are never expressed with passive voice in the corpus. In story 14 ("U.S. Says Moscow Threatens to Quit Talks on Missiles," New York Times, October 12, 1983), there is no need to interpret the adverbial in 1980 and in 1979 with a time program, unless relative ordering of background events is desired. The mere presence of the passive marks these events as occurring before the time of the main events in the story.

(14:8) Talks on a comprehensive test ban of nuclear devices were suspended in Geneva in 1980, and the Geneva negotiations were suspended in 1979.

Main events then are expressed in main clauses with simple past verbs. Events and states which existed before these main events are expressed with a greater variety of syntactic forms, from main clauses, to relative and subordinate clauses, down to noun phrases (which are not analyzed by DUMP). Nominalizations are perhaps the most frequent conveyors of background information in the news. The nominalization rule transforms a sentence into a noun phrase which can then be inserted into another sentence. It is a highly presuppositional structure, since the subject and object of the original verb are often deleted during the transformation and the reader must then supply these arguments from world knowledge. An example from the second story in the corpus ("Lebanon Needs Israeli Troops, Shultz Told," Boston Globe, March 14, 1983) shows the heavy use of nominalizations to create a very long prepositional phrase which contains not a single verb:

(2:2) In the first high-level contacts between the two governments since the start early this year of US-Israeli-Lebanese negotiations on the withdrawal of Israel's forces from Lebanon,....

We will see other uses of nominalization to express other information categories and to refer to episodes with a single word.

The following incomplete list gives a cursory look at the strong correlation between the remaining information categories in news reports and the syntactic forms used to express them. Most of the examples are from story 6, about envoy Stone's meeting with a Salvadoran guerrilla leader, and story 16, about the defeat of the Peronists in Argentina's elections. The next two categories, Current States and Plans, also locate events or states in time, and therefore must occur in finite clauses.

Current States: This category describes the state of the world at the time the report is written. Current states are expressed with simple present or present progressive verbs used in main clauses and in subordinate and relative clauses.

(6:10) Stone has repeatedly sought to meet with political leaders of the Salvadoran left, all of whom live in exile,....

(16:11) The country Mr. Alfonsin is due to govern is racked by a deep economic crisis.

Plans: These may be expressed with appropriate modals (will, might, would) in the same structures used for Current States.

(6:10) His mission is to encourage participation by the left in Salvadoran elections, which will probably be held in March 1984.

(16:10) Military officials said the ruling junta would consider it in a meeting Tuesday.

Certain verbs which express present planning (come, go, leave, start) can be used to indicate future time with the present tense: "Fiscal year 1983, which begins Oct. 1....".

It seems to be a discourse principle of journalese that while non-main events may be "promoted" to expression by the most assertive clause type, they may also be expressed with less assertional forms: subordinate and relative clauses, nominalizations, etc. The converse, however, is not true. Main events may never by "demoted" to expression by any other than the most assertive form.

The remaining information types do not locate actions in time, and therefore are free to appear in constructions without finite verbs.

Import: This category is occasionally expressed with equative sentences of the form: NP V-be NP. The subject and predicate NPs tend to be nominalizations, with the former referring to the main episode.

(16:4) The election...was a stunning defeat for the Peronists....

Election refers to the main event introduced in 16:1. 16:4 tells why that event is newsworthy.

Nonrestrictive PPs with nominalizations as heads may also express Import:

(4:1) The...Budget Committee, in a major blow to President Ronald Reagan, voted yesterday to hold the real growth in defense spending to 5 percent next year.... ("Senate Panel Trims Reagan Arms Budget," Boston Globe, April 8, 1983)

Identifications: With only one exception, all identifications in the corpus are made with prenominal modifiers ("Prime Minister Smith") or with appositives, which may be embedded recursively:

(6:3) ...Stone...talked with Ruben Zamora, the No. 2 leader of the Revolutionary Demo-

cratic Front, the political arm of the five Marxist-led guerrilla bands fighting government forces here.

Effects: Detached participial phrases are used to tell the effects of the actions described in main clauses.

(16:1) The leader of a middle-class party has swept to victory in Argentina's presidential elections, handing the union-based Peronists their first election defeat in nearly four decades.

Comments: Comments are simply quotations from people involved in an event. While in other narratives, dialogue is often the chief means of telling a story and moving the action forward, this is not the case in newswriting. Here, quotes from participants add flavor and give supplementary information, but they are never the sole vehicle for informing readers of an event. This is a lucky fact, since the syntactic forms used in quoted speech are usually much less constrained than those in non-quoted portions.

(16:5) "We are entering a new stage," the 56-year old Mr. Alfonsin, whose politics are left of center, said in a television interview early today.

Collateral: News reports tell what did not happen in a story, what events and processes never were, with surprising frequency. This information category is expressed by negations of clauses, including negative existentials, negative subordinate clauses, and various negative prefixes and prenominal modifiers.

(6:7) Salvadoran officials had no immediate comment on what they heard from Stone....

(6:9) Stone had been unable to arrange a meeting with the Salvadoran rebel leaders... earlier this month.

If it were the case that the correspondence between a syntactic form and the information types it expresses was one-to-many, this relation would not be of much help in automatic processing. In fact, the correspondence is closer to one-to-one, so that, for example, equatives only express import and not identifications, as would be natural in conversational English ("Smith is mayor of the city").

DUMP was successful in creating good summaries and labeling the information types for all but two of the twenty-three stories in the corpus. These two exceptions were highly eventful, chronological accounts and DUMP had difficulty distinguishing minor events from major ones. In addition, after the completion of the program, it performed well with a final story not from the corpus.

Syntactic Correlates of Episode Boundaries

About one-third of the stories in the DUMP corpus consist of more than one episode. Story 17, given here with its DUMP-derived analysis of information, contains three minor episodes in addition to the major one introduced in the first sentence of the report. The discussion below of syntactic forms used to indicate episode boundaries will call upon this story for examples.

* * * * *

Story 17
The New York Times, November 4, 1983
"Senate Approves Secret U.S. Action Against Managua"

By Martin Tolchin
Special to the New York Times

Washington, Nov. 3 - 1. The Senate today approved by voice vote continued aid for covert operations in Nicaragua. 2. The approval was made contingent upon notification to the intelligence committee of the goals and risks of specific covert projects.

3. The action would provide only $19 million of the $50 million that the Administration sought for covert operations in Central America, mostly in Nicaragua. 4. Those funds are expected to run out in less than six months, when the Central Intelligence Agency would have to give an account of its activities as it sought the rest of the funds.

5. The vote followed an hourlong debate that focused on covert United States activity in Nicaragua, which was banned in a House-passed bill. 6. The House bill would provide $50 million in open assistance to any friendly Central American government. 7. House and Senate conferees will now seek to resolve differences in the two measures, and the Nicaraguan dispute is expected to be a stumbling block in the negotiations.

Judge Orders Investigation

8. In San Francisco, a Federal district judge ordered Attorney General William French Smith to conduct a preliminary investigation of charges that President Reagan and other Government officials violated the Neutrality Act by supporting the activities of paramilitary groups seeking to overthrow the Nicaraguan government. 9. The ruling came in a lawsuit filed by Representative Ronald V. Dellums, Democrat of California [Page A9].

10. Senator Daniel Patrick Moynihan, the New York Democrat who is vice chairman of the Intelligence Committee, told the Senate that the Administration had modified its covert policy last summer, and was not supporting the insurgents seeking to overthrow the Sandinista government.

Summary of Main Events: The Senate today approved by voice vote continued aid for covert operations in Nicaragua. Senator Daniel Patrick Moynihan told the Senate that the Administration had

* Dump does not analyze either subtitles, which not all newspapers use, or titles.

modified its covert policy last summer and was not supporting the insurgents seeking to overthrow the Sandinista government.

Past Events: ...which [covert US activity in Nicaragua] was banned in a House-passed bill.

Current State: Those funds are expected to run out in less than six months.
...the Nicaragua dispute is expected to be a stumbling block in the negotiations.

Plans: Sentence 3.
...when [in less than six months] the Central Intelligence Agency would have to give an accounting of its activities as it sought the rest of the funds.
Sentence 6.
House and Senate conferees will now seek to resolve differences in the two measures.

Secondary:* The approval was made contingent upon notification to the intelligence committee of the goals and risks of specific covert projects.

Identifications: ...Moynihan, the New York Democrat who is vice chairman of the Intelligence Committee.

The remaining uncategorized sentences are episode markers and will be discussed below.

* * * * *

As noted earlier, orthographic paragraphs are not used in newswriting to indicate episode boundaries. In their place are a small number of constructions which regularly introduce new episodes, relating them temporally to previous episodes. These structures include the double container sentence, the sentence introduced with a non-restrictive location PP, the LinkS, and the detached time adverbial with a nominalization in it.

The first four sentences of story 17 concern the main episode. A new, minor episode is introduced by the double container in sentence 5. This kind of structure has a verb from the small class (e.g. precede, follow, result in) which may take a nominalization in both subject and object position. The subject refers to an old episode and the object to a new one.

> (17:5) The vote followed an hourlong debate that focused on covert United States activity in Nicaragua....

The subject vote refers back to the story's main event, the Senate vote in the first sentence. The object, or new episode, is the nominalization debate. The object also tells of another episode concerning passage of a House bill. This bill episode is developed in 17:6 and 17:7.

The second minor episode is introduced with a simple detached PP of location in 17:8. This structure is used to shift the setting from the dateline location to a new place. In this case, the action moves from Washington to San Francisco:

> (17:8) In San Francisco, a Federal district judge ordered Attorney General William French Smith to conduct a preliminary investigation of charges that President Reagan and other Government officials violated the Neutrality Act....

This episode is not developed any further in this report, but is interrupted in the next sentence, a LinkS, by the third minor episode. The LinkS is of the form:

NP-nom came { PP / conjunct S }

The nominalized subject refers back to a previous episode and the object of came refers to a new episode. The conjunct or preposition shows the new episode's temporal relation to the old.

> (17:9) The ruling came in a lawsuit filed by Representative Ronald V. Dellums, Democrat of California. [Page A9.]

The lawsuit episode is developed elsewhere in the paper. The page reference closes this episode, and therefore, since 17:10 contains no reference to a new place or time, and has a simple past main verb (told), it must by default be part of the original, main episode. This decision is supported by the eleventh sentence in the story (not included in the corpus):

> After this policy change, Mr. Moynihan said, the committee approved additional funds.

There is no example of the final episode marker in story 17--the sentence introduced by a detached time adverbial with a nominalization in a time phrase ("Two hours before the vote"; "During the Pope's visit"). The nominalization refers to a previous episode and the main sentence to which the whole adverbial phrase is attached introduces the new episode. Story 10 ("French Jets Retaliate, Hit Shiite Positions," Boston Globe, November 18, 1983) begins with French planes bombing Iranian-backed militia in Lebanon. A related episode starts in sentence 5:

> (10:5) Six hours after the French air attacks, gunmen fired rocket-propelled grenades and automatic weapons at a French peacekeeping post in the Shiite Moslem neighborhood of Khandik Ghamik in West Beirut.

Each episode in a report has the potential to contain its own main events, background events, plans, current states, identifications, and so forth. An extension of DUMP's labeling ability would be the creation of a discourse tree for each news report, with a root node dominating episode nodes, which in turn dominate relevant information categories.

---

* This category is not a very reliable one. It includes clauses with passives and copulas.

THE DUMP PROGRAM

DUMP works very simply. It takes as input parsed sentences of a story and searches through them for the kinds of syntactic labels described above (declarative sentence, detached PP, etc.). These labels introduce information fields, each of which is stored on a stack. A set of rules is then applied to each entry on the stack, and assignment of each entry made to one of the information categories on the basis of the structural label and optional tense/aspect marker.

DUMP does not need a full parse of a sentence to assign syntactic structures to a particular information category. For example, it does not need to know anything about the attachment of clause-internal PPs, a difficult problem for parsing programs. Furthermore, newswriting (with the exception of quoted portions, which DUMP does not need parsed) does not reflect the use of a full grammar of English. The corpus contains no question forms and a number of the "stylistic" transformations (pseudo-cleft, topicalization are examples) do not appear. The question of whether some kind of "fuzzy" parser with a limited number of rules could provide adequate output for DUMP is one for further research.

On the other hand, whatever parser is used to prepare input for DUMP will need certain labels not ordinarily found in parse trees: sentences are not usually distinguished as equative or double container in type. Furthermore, DUMP requires some non-standard features on words. For example, we have seen in a number of instances how crucial it is to mark nouns as nominalizations.

RELATION TO OTHER WORK

The DUMP program embodies principles useful both to the processing of sublanguages and to AI research. In the former case, these principles allow preliminary automatic processing of texts within the same genre, regardless of the breadth of the semantic field. As noted earlier, current work with sublanguages relies on word co-occurrence classes which result from their very constrained subject matter. Newswriting covers a wide range of topics and therefore word co-occurrence classes are not an efficient method of automatic processing. However, these reports do show predictable constraints in the use of syntactic constructions to express particular kinds of information and it is this regularity that DUMP depends upon.

In the case of AI research, DUMP can serve as a support program to knowledge-based processors. The FRUMP program (DeJong, 1979), for example, creates summaries from sketchy scripts by looking for key requests, or main events, in the text. So, the script for an earthquake story might contain key requests for information about the quake's rating on the Richter Scale, the amount of property damage it did, where the epicenter was located, and how far shock waves were felt. FRUMP would then look to the newspaper text for evidence of each of the key requests in the script. The scripts are written by the programmer, based on his or her assumption of the most important information likely to be found in all stories about a particular topic. DUMP is freed from reliance on such scripts because of the fact that the news reporter, however unconsciously, encodes key requests syntactically. DUMP can locate these key requests easily and also signal the beginning of new eipsodes, thus facilitating one of the tasks which FRUMP finds most difficult--that of script selection. (Imagine the confusion that could result in story 17 when the Congressional script is interrupted in the eighth sentence by an episode requiring a judicial script.) Once all of the detached clauses and episodes in a report have been correctly labelled by DUMP, a knowledge-based processor could then go about building conceptual representations for each unit.

It is expected that DUMP's approach could be extended to other genres of writing, since most texts achieve texture by distinguishing foreground from background. However, texts vary in the proportion of foregrounded to backgrounded material and in their preference for certain forms to convey grounding. The literary style of a discourse will therefore influence the design of automatic text processing programs. The style of news reports is relatively subordinated, non-redundant, and predicationally dense. The sentences in the DUMP corpus average 2.88 predications per sentence, as compared to a high of 2.78 in the informative sections of the Brown corpus and 2.64 across all genres (Francis and Kucera, 1982). The term predication refers to both the finite and non-finite types, and therefore the 2.88 figure indicates that the news corpus is characterized by a great deal of embedding of both types: finite clauses (relative clauses, adverbial clauses), and well as non-finites (infinitive complements, reduced relatives, participials).

It can be hypothesized that a highly predicated writing style such as journalese will show greater variety in its syntactic structures than a style with few predications per sentence. This syntactic diversity will reflect a text with less foregrounded material--in short, a text with greater texture. A further hypothesis is that in a predicationally dense style there will be a stronger correlation between syntactic forms and the particular information types expressed by these forms. It seems likely that a genre which uses few predications per sentence would consist chiefly of main clauses used as the workhorse to express all kinds of information: background, main events, plans, import, and so forth. Some of these information categories will be distinguishable by verb tense, aspect, mood and voice, as in the news. But others will have to rely on world knowledge for categorization. As an example, consider a revised version of the opening of story 6, rewritten so that embedded clauses in the original are expressed as main clauses:

> Richard B. Stone met face-to-face today with a key leader of the Salvadoran guerrilla movement. He spent several frustrating weeks

maneuvering the meeting.

"The ice has been broken," proclaimed President Belisario Betancur of Colombia. He engineered the meeting.

Knowledge about the way plans are made would be needed to distinguish foreground from background in these sentences.

One further metric can be hypothesized for determining discourse genres suitable for syntactic analysis. In syntactic theory there is a well-known correlation between the flexibility of word order in a language and its use of morphosyntactic inflections. Languages like English which have lost most of their inflectional markers rely on rigid word order to establish syntactic relations. On the other hand, highly inflected languages like Latin can afford greater flexibility in word order since inflections on the ends of words indicate their function in the sentence.

An analogy might be drawn in which syntactic structures correspond to morphosyntactic inflections and information order in discourse corresponds to word order. The discourse structure of news reports violates canonical story form. The writer does not start at the beginning and relate events through to the end. The potential confusion introduced by this unpredictability is compounded by the density of new information in news reports. Perhaps the great regularity in the use of distinct syntactic forms to express the types of information conveyed in the news serves to compensate for the flexibility in discourse structure. It is as though the strong correlation between syntactic form and information type frees the reader to process the large amount of new information being delivered. Just as inflectional endings allow the listener to assign words to their functional slots regardless of the order in which they appear, so the syntactic correlates to information types allow the news reader to quickly assign phrases their function in the discourse. Stories which adhere to a standard story grammar do not need such syntactic regularity, since the position of the material in the text indicates its function.

The extension of a program like DUMP to other discourse genres would require, first, the identification of the information categories expressed by the kind of text. Cookbooks, for example, convey instructions and descriptions, not main events, effects and identifications. Secondly, correlations between syntactic form and information type and the syntactic means for indicating episode boundaries must be determined. The degree of correlation between syntactic form and information type in non-news genres is a matter for further investigation.

ACKNOWLEDGMENTS

This research was carried out under grant G008101781 from the U.S. Department of Education, Program for the Hearing Impaired.

REFERENCES

Borko, Harold and Bernier, Charles. 1975. Abstracting Concepts and Methods. New York: Academic Press.

Comrie, Bernard. 1976. Aspect. Cambridge: Cambridge University Press.

Decker, Nan. 1985. Syntactic clues to discourse structure: A case from journalism. Ph.D. dissertation, Brown University.

DeJong, Gerald. 1979. Skimming stories in real time: An experiment in integrated understanding. Research Report #158, Department of Computer Science, Yale University.

Francis, W. Nelson and Kucera, Henry. 1982. Frequency Analysis of English Usage. Boston: Houghton-Mifflin Company.

Green, Georgia. 1979. Organization, goals and comprehensibility in narratives: newswriting, a case study. Technical Report #132. The Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Grimes, Joseph. 1975. The Thread of Discourse. Janua Linguarum, Series Minor, no. 207. The Hague: Mouton.

Hirschman, Lynette and Sager, Naomi. 1982. Automatic information formatting of a medical sublanguage. In R. Kittredge and J. Lehrberger (Eds.), Sublanguage: Studies in Language in Restricted Semantic Domains. New York: Walter de Gruyter.

Hopper, Paul. 1979. Aspect and foregrounding in discourse. In T. Givon (Ed.), Syntax and and Semantics, vol. 12. New York: Academic Press.

_____ and Thompson, Sandra. 1980. Transitivity in grammar and discourse. Language 56: 251-299.

Mourelatos, Alexander. 1981. Events, processes and states. In P. Tedeschi and A. Zaenen (Eds.), Syntax and Semantics, vol. 14. New York: Academic Press.

Ota, Akira. 1963. Tense and Aspect of Present-Day American English. Tokyo: Kenkyusha.

Sager, Naomi. 1981. Natural Language Information Processing: A Computer Grammar of English and its Applications. Reading, MA: Addison-Wesley.

Schank, Richard and Riesbeck, Christopher. 1981. Inside Computer Understanding. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, Sandra. 1983. Grammar and discourse: The English detached participial phrase. In F. Klein-Andreu (Ed.), Discourse Perspectives on Syntax. New York: Academic Press.

Vendler, Zeno. 1967. Linguistics in Philosophy. Ithaca, NY: Cornell University Press.

Woods, William. 1973. An experimental parsing system for transition network grammars. In R. Rustin (Ed.), Natural Language Processing. Englewood Cliffs, NJ: Prentice-Hall.