Analysis of Subordinating Conjunctions

The purpose of this paper is to establish a lexical knowledge base (LKB) for subordinating conjunctions and to describe the procedures by which it is created. The LKB is based on the subordinating conjunction definitions in Webster's Third New International Dictionary (hereafter W3). For the purposes of this LKB, 107 main entries and 374 distinct senses for subordinating conjunctions were identified in W3. The LKB is organized into an inheritance hierarchy with distinct primitives (as identified from W3), so that redundant information does not need to be stored with each entry. The procedures for its creation entail treating the entries and senses as nodes in a labeled directed graph (hereafter digraph), which enables the use of theorems from digraph theory in creating the LKB. The procedures are based on the goal of identifying the primitives, achieved by the inverse process of identifying those senses that cannot be primitives.

This paper is a work in progress. At this point, we describe (1) how entries and senses were selected; (2) procedures used to move toward identification of the primitives of subordinating conjunctions; (3) preliminary analysis of the syntactic and semantic structure of subordinating conjunctions, including (a) a WordNet analysis of key words used in subordinating conjunction definitions, (b) procedures for hierarchizing the observations about subordinating conjunctions in Quirk et al. and integrating the work of Ken Barker ("Interactive semantic analysis of clause-level relationships"), and (c) integrating the work of Alistair Knott ("A data-driven method for classifying connective phrases," *Journal of Language and Speech, 39* (1996)); and (4) a brief overview of the next stages of the analysis (which are under way). Any comments, criticisms, and suggestions are welcome, as are any offers of assistance for characterizing appropriate features to be assigned to the lexical entries.

1. Identification of Entries and Senses

A subordinating conjunction is a linguistic form that makes a clause "a constituent of another clause" {Quirk, et al. 1985: 44}. Subordinating conjunctions are not specifically identified in W3, but rather constitute a subset of entries marked as conjunctions. W3 was visually scanned for entries whose part of speech was given as a conjunction and the definitions of these entries were included for consideration. (Note 1) Definitions that clearly expressed a coordinating function were excluded from further analysis (primarily definitions of *and*, *but*, *either*, *neither*, *nor*, and *or*, although some, usually obsolete or archaic, definitions of these words were retained as subordinating conjunction definitions). Many senses among the conjunctions contain a usage note in addition to a definition. Some these usage notes indicated that the entry word could be used with another word in the particular sense. For example, one sense of *before* indicates that it is "sometimes used archaically with a postpositive *that*." This is taken to mean that the phrase *before that* may appear in a subordinating conjunction role and hence, should be treated as an entry.

Table 1 lists the 107 entries that were identified in W3 using the criteria listed above, 48 of which are phrasal entries. The table also shows the number of senses for each entry; this number does not correspond to the W3 identification of a sense, but rather uses each distinct defining phrase or synonymic cross-reference as a sense. For entries created on the basis of the

usage notes, a sense was created without any definition or usage note (an apparently empty sense); in the LKB, the meaning for such a sense is to be obtained by inheritance (described in more detail below), so to find the meaning, it is necessary only to link the sense back to the sense in the dictionary licensing its use. The result of these processes is that there are 374 senses in the initial LKB. In the table, 27 entries are identified as arising only from usage notes (having a total of 46 senses) and 3 entries have one or two empty senses (among others overtly in W3), for a total of 50 empty senses.

Table 1: Subordinating Conjunctions and Their Number of Senses

according as 5	after 3	†after that 1	against 3
albe 1	albeit 2	although 6	and 3
as 23	as if 3	as long as 2	as soon as 1
as though 1	because 8	before 2	†before that 2
being 2	†being as 2	†being that 2	but 11
†but that 2	but what 2	considering 1	†different than 1
†differently than 1	directly 2	†else than 1	except 3
†except that 2	for 4	for all that 2	for fear 1
‡for that 2	forasmuch as 3	†hardly than 1	how 7
howbeit 1	if 9	immediately 1	in case 4
in case that 2	‡in order that 2	†in that 1	inasmuch as 6
insofar as 2	insofar that 2	insomuch as 1	†just so 1
less 1	lest 3	like 7	†no sooner but 1
†not that 1	notwithstanding 1	†notwithstanding that 1	
now 3	†now that 4	on condition that 2	once 4
only 5	or 1	or else 1	†other than 1
provided 3	providing 3	save 3	†save that 3
†scarcely than 1	seeing 3	†seeing as 3	†seeing that 3
since 6	so 4	†so as 2	so long as 4
‡so that 4	†such as 2	†such that 1	supposing 2
than 3	that 15	though 6	till 17
†till that 4	†to the end that 1	unless 6	until 5
†until that 1	when 10	whenas 6	whencesoever 2
whenever 3	whensoever 3	where 13	whereas 10
wherefore 1	wheresoever 1	wherethrough 5	whereunder 1
whereup 1	wherever 2	whether 5	while 12
whilst 2	whither 2	without 2	yet 2

† Entries consisting of one or more senses with no definition or usage note legitimated only through usage notes, usually in definitions associated with the last word in the entry.
‡ One or more senses of these entries has no definition or usage note and is legitimated as in the previous note.

2. Identifying the Primitive Senses

Using DIMAP, the first set of links are established from the empty senses (that is, senses having no definitions and no usage notes) to the senses where they are licensed. These are superconcept or hypernymic relations and indicate that the meanings of these senses are inherited from the licensing senses. Thus, when the meanings of the licensing senses are fully specified, the meanings of the licensed senses can be taken as fully specified. By establishing these links, we have implicitly asserted that the empty senses are not primitives, since they derive their meaning from what they inherit. As a result, our digraph now contains only 80 entries and 324 senses to be analyzed; the primitives of the subordinating conjunctions must be found among

these entries and senses. It is this process of eliminating entries and senses from the possible primitives that is to be continued.

What we want to do at this point is to identify which of the definitions within this set can be primitives. We can establish which definitions cannot be primitive through the application of several rules. These were first enunciated in {Litkowski 1975; Litkowski 1976; Litkowski 1978; Litkowski 1980}, and are repeated here with particularization to subordinating conjunction definitions.

Subordinating conjunction definitions are similar to the better studied noun and verb definitions in having a genus term and differentiae. That is, each subordinating conjunction definition is defined in terms of another subordinating conjunction and, optionally, differentiae which act as restrictions on the meaning of the genus subordinating conjunction. (There are exceptions that will be noted and discussed below.) Like noun and verb definitions, disambiguating the sense of the genus term in a subordinating conjunction definition is an issue as well.

Let **E** be the set of definitions that have been eliminated from possible membership in the set of primitives. Initially, this set is empty; definitions are added to it through the application of reduction rules based on graph theoretic principles, to which are added lexicographic and lexicologic principles. The basic lexicographic principle for a definition to be primitive is that it cannot be defined by a **true hypernym**, provisionally defined to be a definition with a hypernym plus differentiae.

Initially, we view each subordinating conjunction as a node in the digraph, containing all its definitions. Later, we will establish nodes containing only a subset of the set of definitions for an individual subordinating conjunction, including possibly a single definition. Also initially, we view a link as the simple relation **R** "is used as the genus term in a definition of." Thus, we would say that **though R albeit**, since one of the definitions of *albeit* is "even though."

The first stage of the analysis of the subordinating conjunction digraph is to establish hypernymic links in the dictionary. This was done semi-automatically by examining each definition of a subordinating conjunction for the presence of a word or phrase that is identified in the dictionary as a subordinating conjunction. Hypernymic links were also established and entered manually for the empty senses licensed in the usage notes; the hypernym for such an empty sense is the licensing sense. Table 2 shows the number of uses of each word as a hypernym (or genus term). Fifty entries are used as hypernyms in 313 of the 374 subordinating conjunction senses. There are 20 senses that have no definitions and hence no hypernymic links, but they do have usage notes; they are found among the senses of *than* (2), *that* (15), and *whether* (3). These can be viewed provisionally as being among the primitives. All definitions of *that* fall into this category, so provisionally, it is counted as a primitive subordinating conjunction. The remaining 41 senses have relativizers in the definitions, indicating that the use of the subordinating conjunction establishes a relative clause attached to some element in the matrix clause rather than a subordinating clause; these sense may thus be excluded from the subordinating conjunction analysis. (Where a subordinating conjunction is part of a phrasal

subordinating conjunction, the search is conducted on the phrasal subordinating conjunctions before the single word, so that, for example, *as soon as* would have preference over *as*.)

Table 2: Subordinating Conjunctions Used as Genera

after (1)	albeit (1)	although (9)	as (15)
as if (4)	as long as (2)	as soon as (4)	as though (1)
because (5)	before (6)	being (4)	†but that (2)
but (5)	considering (1)	except (3)	how (1)
if (17)	in case (3)	in order that (4)	inasmuch as (4)
insofar as (2)	lest (1)	†notwithstanding that (1)	
now (3)	on condition that (4)	only (2)	or else (1)
provided (1)	save (2)	seeing (6)	†seeing that (2)
since (8)	so (3)	so long as (1)	so that (3)
supposing (1)	than (10)	that (111)	though (7)
till (3)	to the end that (1)	unless (6)	until (3)
when (20)	whenever (1)	where (2)	whereas (3)
wherever (2)	whether (2)	while (9)	
† (empty sense word)			

Reduction Rule 1 (RR1): If a subordinating conjunction or any of its definitions is not used as the genus term in a subordinating conjunction definition, it is not primitive and can be placed in E. In graph-theoretic terms, this is a node in the digraph that has 0 outdegree and positive indegree (that is, the definitions represented by the node are defined by one of more subordinating conjunctions); in other words, these nodes are leaves in the digraph. The 57 words in Table 3 satisfy this rule and can be placed in **E**; this can also be accomplished simply by not finding these entries in performing string searches for them in the definitions of subordinating conjunctions. We know that such words are not primitive because they are defined by more primitive words, that is, their genus terms.

Table 3: Words Not Used as Genera

according as	after that	against	albe
and	before that	being as	being that
but what	different than	differently than	directly
else than	except that	for	for all that
for fear	for that	forasmuch as	hardly than
howbeit	immediately	in case that	in that
insofar that	insomuch as	just so	less
like	no sooner but	not that	notwithstanding
now that	once	or	other than
providing	save that	scarcely than	seeing as
so as	such as	such that	till that
until that	whenas	whencesoever	whensoever
wherefore	wheresoever	wherethrough	whereunder
whereup	whilst	whither	without
yet			

There are 61 senses for which links were not established. Of these, 20 senses had no definitions. The remaining 41 require special analysis (they seem to have relativizers in the definition, indicating that the clause they introduce is intended to refer to some specific element of the matrix clause, thus making these senses not subordinating conjunctions, but rather

relativizers). Of these 41, 31 definitions had the word *which*. This leaves 10 using some other sort of definition; most of these contain the word *what* or *whatever*.

Lexicographic Rule 1 (LR1): If a definition has a usage label (such as *archaic, chiefly Brit, astronomy*), it cannot be used as the genus term in a definition without the same label and hence is not primitive and can be placed in E. The presence of a usage label marks a sense as existing inside a sublanguage, whether of time or subject area. Lexicographers do not define the core of the language on the basis of a sublanguage, so we can eliminate such definitions as non-primitive. Table 4 identifies 20 senses of nine entries still not eliminated after application of RR1 and having usage labels. (This rule could have been applied before RR1.)

Table 4: Senses Eliminated by Lexicographic Rule 1

as (3)	because (2)	but (4)	since (2)
so that (1)	though (2)	whereas (1)	whether (2)
while (1)			

By placing entries and definitions in E with RR1 and LR1, we have taken off a couple of layers (as of an onion) in moving toward a core. We can continue to identify layers on the basis of the proximity of the remaining entries and senses to the members of E. We do this successively by adding words to E that are used as genera only in the senses already in E. We can continue this process in successive passes, investigating the possibility only as long as a pass produces at least one new member of E. We formalize this process in another reduction rule.

Reduction Rule 2: If a subordinating conjunction or any of its definitions (with at least one or more of which having a hypernymic link) is used as the genus term only in subordinating conjunction definitions already in E, it is not primitive and can be placed in E. In graph-theoretic terms, these are nodes that have positive outdegree and positive indegree. Since they used to define only words that are not primitive, and they are defined in turn by more primitive words, they cannot be primitive. Table 5 shows that 25 entries are added to E in four passes through the digraph.

Table 5: Words Eliminated by Reduction Rule 2

<u>Pass 1</u>			
after	albeit	as though	being
how	lest	now	or else
provided	save	till	to the end that
whenever	where		
<u>Pass 2</u>			
as soon as	before	seeing that	so that
until	wherever		
<u>Pass 3</u>			
seeing	so	than	
<u>Pass 4</u>			
considering	in order that		

The results of applying RR1, LR1, and RR2 have induced the beginnings of a partial order on the set of subordinating conjunctions. Applying these rules has simplified the task of identifying the subordinating conjunction primitives. From the original 107 subordinating conjunctions, we have eliminated 82 entries and their definitions from the analysis. Presumably, when we clearly articulate and order the senses in the remaining 25 entries, we will be able to propagate these meanings back to the 82 entries. Moreover, we have reduced the complexity of examining individual uses as genera. For example, we originally had 17 uses of *if* as a genus term; in trying to assess its significance, we have eliminated 7 uses.

We can now proceed to a closer examination of the defining relationships in the digraph that remains at this point. To recall, the digraph is articulated through a list of its "adjacent-from" links. Thus, for example, the node *when* has links to {*as, since, whereas, while*}. We want to consider the definitional nature of these links. Some of the links between nodes are synonymic links, in which the defining word has no differentiae (such as *when* in one definition of *as*, in small capitals in W3), and others are non-synonymic links (such as *when* in one definition of *since*, "after the time in the past when"). The synonymic links may be viewed as carriers of all the senses of the defining word. Thus, for example, without further specification, we might view all the definitions of *when* as being carried to *as*, so that *as* has all its own senses and all those of *when* as well. (In all likelihood, the lexicographers did not intend this to be the case, but we have no basis for disambiguating just which senses of *when* to transmit to *as*, whereas the differentiae in *since* may provide some context for disambiguation.) Now, if we follow synonymic links, we observe that the carrier function introduces some redundancy in the digraph. For example, *while* is used to define {*as, so long as, though, when, whereas*}. By following the link to *though* and following its links, we see the definitional path (*while, though, although, when*) consisting only of synonymic links. Thus, there are two paths between *while* and *when*, either of which serves to carry the definitions of *while* to the meaning of *when*. We don't need both paths and so we eliminate the direct link (*while, when*) as redundant. This lead to the lexicographic rule.

Lexicographic Rule 2 (LR2): A direct link xRy can be eliminated from the digraph as redundant where the link is synonymic and there is another path in the digraph between x and y. These links are redundant not only lexicographically, but also from the standpoint of the digraph, in that their elimination does not change the connectivity of the underlying digraph because a path still exists between two nodes. However, their elimination reduces the complexity of the digraph, with fewer links that have to be considered.

LR2 is implemented through a depth-first search of the digraph, which is simplified through the application of various stopping rules. The digraph at this point in the analysis usually will have cycles, so these are taken into account. There are three events that stop a particular path: (1) reaching a node already in the path; (2) reaching a non-synonymic link; and (3) identifying a redundant link. The links eliminated as redundant are sensitive to the ordering of the nodes in the digraph, so that if the order in which two nodes are examined is reversed, a different link may be eliminated. This is of no significance, since by their nature, these links are redundant and do not change the carriage of definitions. Table 6 identifies 11 redundant links eliminated from the digraph of subordinating conjunctions.

Table 6: Words Eliminated by Lexicographic Rule 2

(if, when)	(though, as)	(although, while)
(while, when)	(while, as)	(while, whereas)
(because, as)	(since, as)	(since, as long as)
(but, only)	(as long as, so long as)	

A further class of synonymic links is the direct cyclical (or symmetric) link. In this case, we have **xRy** and **yRx**, where all the definitions of **x** are carried to **y** and all the definitions of **y** are carried to **x**. But these definitions include the words **x** and **y**, so that by performing the carriage and, for example, considering the now expanded set of definitions for **x**, we would find that one of its definitions is **x** itself. The presence of cycles in the digraph gives rise to many of these situations, frequently through the intermediation of lengthy paths. We will examine the full nuances of these cycles below, but for the moment, we will bite off a small chunk of this problem by considering only two cases in which the cycle is immediate between two nodes and where one of the words has only a single link remaining in the digraph.

The first case is where the link is synonymic for both words. In this case, word **y** is used only as the genus of word **x** and nowhere else in the digraph. When we carry the senses of **y** to **x**, we don't want to carry **x**, since this link provides no additional defining information for **x**. To prevent this, we eliminate the link **xRy** from the digraph.

Lexicographic Rule 3 (LR3): A direct link xRy can be eliminated from the digraph as a cyclical synonymic link if the link is synonymic, the direct link yRx also exists, and y has only the single link yRx in the digraph or additional links yRz_i where there is a synonymic path from x to z_i. In the latter case, the links yRz_i may also be removed.
Removing the direct cyclical link is equivalent to removing the sense **x** in the set of definitions of **y**. However, making this elimination does not remove any information from the digraph that remains at this point. If **y** has other uses $\mathbf{z_i}$ in the digraph, the synonymic path from **x** to any of these uses is preserved even when the links between **y** and these $\mathbf{z_i}$ are eliminated. Eliminating the link **xRy** also has an effect on the digraph, making the senses of **y** (now minus **x**) more primitive than **x**. Only one such link, *(if, in case)*, was removed from the subordinating conjunction digraph through application of the first part of the rule. The link *(if, on condition that)* was eliminated through application of the second part of the rule, along with the link *(on condition that, when)*.

The second case of direct cyclical links is where one of the links is non-synonymic. We might describe this situation as $\mathbf{xR_{NS}y}$ and $\mathbf{yR_Sx}$. The non-synonymic link indicates that **x** appears with some differentiae. In this situation, the carriage of senses from **y** to **x** would result in **x** being defined by itself plus some differentiae. In general, this is an unacceptable lexicographic practice, so we want to eliminate it. (There may be situations where the differentiae are inherent in the meaning of **x**, so that carriage of the differentiae would actually add nothing but redundant information and would be acceptable and indicative of a true cycle in a dictionary. We will consider this situation in more detail below.) To eliminate this situation, we can eliminate the link $\mathbf{xR_{NS}y}$ so that $\mathbf{yR_Sx}$ will not carry the unacceptable sense back to **x**. (This is the first instance where we have modified the nature of the link **R** between two nodes,

from "is used to define" to "is used non-synonymically to define" and "is used synonymically to define". We will consider other modifications more systematically below.)

Lexicographic Rule 4 (LR4): A direct link $xR_{NS}y$ can be eliminated from the digraph as a cyclical non-synonymic link if the link is non-synonymic, the direct link yR_Sx also exists and is synonymic, and y has only the single link yRx in the digraph. Another way of viewing this situation is that what we have done is to partition the definitions of **y** into two sets, one containing the single sense containing **x** and the other containing the remaining definitions of **y**. We treat this as splitting the node **y** in the digraph into the nodes $\mathbf{y_1}$ (with the single sense containing **x**) and $\mathbf{y_2}$ (with the remaining senses). Since **y** has only one outgoing link in the digraph, we have to consider what will happen when we split the node. Since we have precluded the non-synonymic link from carrying the definition at $\mathbf{y_1}$ back to **x**, we have made $\mathbf{y_1}$ a leaf in the digraph with no uses. We now have to consider whether making this elimination removes any information from the digraph that remains at this point. The use involving **x** has clearly involved the use of differentiae, and so, intuitively, is probably not primitive. Eliminating the link $\mathbf{xR_{NS}y}$ also has an effect on the digraph, making the senses of **y** (now minus the definition involving **x**) more primitive than **x**. Only one such link, *(if, supposing)*, was removed from the subordinating conjunction digraph.

At this point, we have made several reductions in the size of the digraph, in the first several steps eliminating nodes and in the last few steps focusing eliminating links. We would like next to examine the overall effect and character of the digraph that remains. To do this, we make use of the most important characteristic of digraphs for our purpose, namely, that every digraph has a **basis set**, that is, a set of points from which all nodes in the digraph are reachable. This basis set is the set of primitives from which all other definitions are derived. A crucial notion in the determination of the basis set is that of a **strong component**, a set of nodes that are mutually reachable by at least one path of the digraph. A strong component is an equivalence class (based here on the relation "is used to define"). We can make use of an algorithm from digraph theory for partitioning the nodes of the digraph into its strong components (that is, equivalence classes), in order to view the superstructure of the digraph, which is itself a digraph.

In general, a digraph created in the manner described here, after the application of the several lexicographic and reduction rules, will not be a single equivalence class. As a result, examination of the superstructure will identify sets of nodes that are relatively more primitive than other sets. Identification of strong components will, in general, provide sets of nodes that are leaves in the superstructure to which we can apply the two reduction rules. Moreover, the superstructure has no cycles and provides a consistent topological sorting that will focus further analysis on the primitives, enabling us to put aside extraneous information. The next reduction rule identifies leaves in the superstructure that can be put aside.

Reduction Rule 3: Let S be a strong component of a digraph and let T be the set of all nodes that are defined with members of S and that are not members of S. If T is a subset of E and there is at least one node not in S and not in T used to define a node in S, then S and all its definitions can be placed in E. What this says is that (1) the members of S are mutually reachable (that is, there are cycles in the definitional paths between members of S),

(2) the members of S are used as superconcepts only for members of S and for subordinating conjunctions that have already been eliminated as non-primitive, and (3) there is at least one subordinating conjunction that is not in S or T that is more primitive than the subordinating conjunctions in S.

Analyzing the subordinating conjunction digraph at this point identifies nine strong components. This superstructure is shown in Figure 1. As can be seen, the superstructure digraph eliminates the internal structure for components with more than one word. Any component with more than one word has an internal cyclical structure so that there exists a path from any word to any other word. Following RR3, all nodes of the graph except the one containing the word *that* are eliminated as non-primitive. The other nodes containing one word (those for *whether*, *in case*, *supposing*, *notwithstanding that*, and *on condition that*) are essentially carriers of relatively small components of meaning into the large node containing what appear to be the 14 most dominant subordinating conjunctions. Figure 2 shows the internal structure of this strong component, where the remaining analysis will focus.

3. Analysis of Syntactic and Semantic Structure

To this point, we have used only the defining structure of the subordinating conjunctions to provide a general ordering. Now, we must delve deeper into the meaning components associated with these lexical items. As with all lexical items, the meaning must be captured in syntactic characteristics (components that describe the usage and context) and semantic characteristics (components that describe the meaning brought to the usage and context). We will use these meaning components to break the cycles shown in Figure 2. The essential method employed to accomplish this is to eliminate cycles that introduce an inconsistency. With respect to syntactic characteristics, we cannot allow a cycle to require that a lexical item have usage or context that differs. With respect to semantic characteristics, we cannot allow a lexical item to bring conflicting information to the usage and context and we cannot allow a more complex item to define a less complex item. In both cases, a simple way of viewing this is to say that we will break cycles by finding unification attempts that fail.

This section describes efforts to characterize the overarching meaning structure associated with subordinating conjunctions, particularly identifying features and meaning components that are associated with the subordinating conjunctions. To accomplish this, we attempt to integrate information and insights from WordNet, Quirk et al., Barker, and Knott. In summary, the results of this effort indicate an elegant hierarchical structure that provides an intermediate level of representation between sentence structure and discourse structure. The results suggest that subordinating conjunctions enable us to characterize clauses as descriptions of such things as times, causes, reasons, places, conditions, and points of reference.

3.1 WordNet Analysis

To begin, we note provisionally that only one of the 15 definitions of *that* seems to be the primitive from which the others are derived. This definition consists entirely of the usage note "used as a function word to introduce a subordinate clause that is joined as complement or modifier to a noun or adjective or is in apposition with a noun." According to Quirk et al. (p.

1047 and pp. 1260-2), these are instances of postmodifier subordinate clauses in a noun phrase, so that the following clause would be treated as an appositive to the head noun (and linkable with the copula *be*), which must be a general abstract noun. A few examples of such definitions are *as soon as* "immediately at or just after the time that," *provided* "with the understanding that," and *inasmuch as* "in view of the fact that" and "for the reason that". A large number of the subordinating conjunction definitions are of the form **PP that**, with the PP of the form **P det N**, with det **the**. The determiner here is cataphoric, referring to the S following the subordinating conjunction and characterizing that S as the N. In W3, the 28 words in Table 7 fill the N position.

Table 7: Words Used to Characterize Sentences Following Subordinating Conjunctions

assumption	belief	cause	circumstance
concomitant	condition	consideration	degree
event	extent	fact	hope
manner	measure	moment	place
point	possibility	provision	purpose
qualification	reason	restriction	result
sort	time	understanding	way

Subordinating conjunctions with this pattern thus appear to be characterizing the subordinate clause. Moreover, the preposition at the beginning of the syntactic pattern indicates that the clause is serving as an adverbial and thus likely to fill one the seven semantic roles posited by {Quirk et al.: 8.2}, namely, **space**, **time**, **process**, **respect**, **contingency**, **modality**, and **degree**. Since subordinating conjunctions essentially serve a rhetorical function, we would view the subordinate clause as an entity functioning in a larger rhetorical structure. The clause may describe anything, but we are here concerned with the rhetorical role played by the description.

The words in Table 7 were examined in WordNet in order to characterize them and understand better the nature of the subordinating clause. Not surprizingly, the set of words exhibit hierarchical structure and patterns within the WordNet hierarchy. Considering all senses in WordNet, the words in this set fall into only 10 of the WordNet noun tops, only five of which seem legitimate characterizations for the subordinating clause (**abstraction**, **psychological feature**, **event**, **state**, and **location**), with the other five being senses with a different orientation (**act**, **entity**, **phenomenon**, **shape**, and **possession**). A clause may describe an act, phenomenon, or entity (as in senses of words such as *consideration* and *qualification*), but we are not concerned with this, but rather the rhetorical status of this act, phenomenon, or entity, perhaps as an abstraction or state.

It is useful to examine where these words fit within the WordNet hierarchy. For each sense of each word, we extracted the hypernymic path to a top and then merged all the paths. (This list, about six pages in length, is available, showing the subset of the WordNet hierarchy induced by the list of words in Table 7.)

For **abstraction** (glossed as "a concept formed by extracting common features from examples"), some of the words fell under five of the six hyponyms of this concept (that is, the synsets **time**, **space**, **attribute**, **relation**, and **{measure, quantity, amount, quantum}**). Most

of the hyponyms under **relation** were down a few hyponymic levels to the synset **{statement}**. Most of the hyponyms under **attribute** were under synsets for **quality**, **property**, and **trait**. There were relatively few words inducing the other branches of the tree under **abstraction**.

For **psychological feature** ("a feature of the mental life of a living organism"), the induced tree included all immediate hyponyms: **{cognition, knowledge}** ("the psychological result of perception and learning and reasoning", most of these falling under the synsets **{content, cognitive content, mental object}** and **{information}**), **{motivation, motive, need}** ("the psychological feature that arouses an organism to action", only *reason* and *purpose*), and **{feeling}** ("the psychological feature of experiencing affective and emotional states", only *hope*).

For **event**, the induced tree was shallow and contained only the hyponymic synset **{happening, occurrence, natural event}** ("an event that happens"), with only its synsets **{experience}**, **{case, instance}**, **{beginning}**, **{accompaniment, concomitant, co-occurrence}**, and **{ending, conclusion}**.

For **state** ("the way something is with respect to its main attributes"), the induced tree was also quite shallow, with only a small number of its hyponyms (**{condition, status}**, **{condition}**, **{situation, state of affairs}**, **{degree, level, stage, point}**, **{status, position}**, and **{being, beingness, existence}**.

For **location** ("a point or extent in space"), the induced tree was small and shallow, induced by the words *way, point,* and *place*.

To summarize the results from WordNet, a subordinating clause can be viewed as expressing an abstraction, a piece of knowledge, an event, a state, or a location, with more specific characterizations depending on the particular subordinating conjunction. The subtrees induced from WordNet correspond well to the semantic roles of adverbial clauses described by Quirk et al. (pp. 1077-1118), that is, clauses of time, contingency, place, condition, concession, contrast, exception, reason, purpose, result, similarity, comparison, proportion, and preference, but probably not and comment. It is an interesting observation that a word in Table 7 may appear several places in the induced WordNet subtree. For example, the word *place* in each of the five major categories, indicating that its use may convey several possibilities. (I hesitate to say that it is ambiguous; rather, I would say that it opens up opportunities for its use, enabling the merging of meanings. Thus, as Quirk et al. suggest (p. 1087), the meaning of *place* may merge with meanings of contingency, contrast, and time.)

3.2 Hierarchizing Quirk et al. and Integrating Barker

{Quirk et al.: 15.24-52} describes semantic roles for adverbial clauses. These are conveniently discussed in several categories, such as clauses of **time**. However, as noted (p. 1077), "many subordinators introduce clauses with different meanings," frequently combining meanings such as **time** and **purpose**. We can thus easily posit that, in general, subordinating conjunctions are composites of underlying semantic components and features. The issue, then, is one of identifying these components and features and arranging them in such a way as to ensure

consistency and appropriate composition (or unification). We thus began by identifying and grouping all the distinct semantic roles mentioned in this discussion, shown in Table 8.

Table 8: Semantic Roles of Adverbial Clauses

temporal	**duration**, **repetition**, **temporal**, **time-before**, **time-after**, **time overlap**, **time beginning**, **time proximity**
contingency	**cause**, **reason** (**motivation**), **circumstance**, **condition** (**presupposition**), **purpose**, **prevention**, **result**, **concession** (**unexpectedness?**, **fulfillment?**)
exception	(although distinguished, this feature always seems to be blended with **condition**)
comparison	**contrast** (**antithesis**), **preference**, **proportion**, **similarity** (points of reference)
place	(although distinguished, this feature always seems to be blended with either **contingency** or **comparison**)
modality	**fulfillment**, **negation**, **plausibility**, **unexpectedness**

The key factor underlying the groups shown in Table 8 arose from trying to deal with **contingency**. Each of the semantic roles identified in this group was discussed in several places throughout the referenced sections in Quirk et al. As such, they seemed to stand independent of the others and not fit into an overall structure. However, in their analyses, both Barker and Knott make considerable reference to an implicational, propositional structure of the form P_1 ... P_n Q. When the various semantic roles identified under **contingency** were considered in light of this propositional structure, it appeared that the semantic roles were ways of characterizing the clauses on either the left- or right-hand side of the rule. It was this observation that made it possible to consider all the semantic roles as part of an overarching hierarchy that could elaborate the one posited by Barker.

Barker suggests three types of clause-level relationships (each with subtypes): conjunctive, temporal, and causal. Moreover, he suggests that there is a ranking of the types: (1) conjunctive relationships merely state a number of propositions (P_i) without any additional information; (2) temporal relationships add temporal ordering to the propositions (while keeping the conjunctive relations); and (3) causal relationships add causal ordering to the propositions (while keeping temporal and conjunctive relations). (This ranking seems to correspond well to the nature of rational thought, first positing a number of propositions, then noticing a temporal ordering, and finally articulating a causal ordering.) This process of adding information to the relationships is one that is well suited to making sense of the semantic roles in Table 8. Thus, for example, we can see that **reason** adds the information that one of the P_i is to be treated as not just a **cause**, but also that a human has attached significance to the proposition as the articulation of a driving force. Similarly, **contrast** adds a component that says P_i is to be viewed as a proposition that stands in contrast with Q (whether or not there is a real logical derivation showing that this is the case). (We remind the reader here that this analysis is focusing on the meanings of subordinating conjunctions, and not coordinating conjunctions, so that, for the most part, we are always dealing with concepts that are essentially temporal or causal.)

An important aspect of understanding the import of this hierarchy is that we can begin to see something about the relationship of particular senses of subordinating conjunctions. For

example, we can see how the causal senses of *since* can be derived from its temporal senses. Similarly, we can see how a causal sense can be added to the meaning of *before*, something not currently recognized in any dictionary. With this general view of the semantic roles of adverbial clauses, the components listed in Table 8 can be further arranged. We can pair the **comparison** group with **conjunctive** relationships, the **temporal** group with **temporal** relationships, and the **contingency** group with **causal** relationships.

From the general grouping in Table 8, it is now necessary to provide a further structure and hierarchy. We begin with the **temporal** group, as perhaps the simplest and as an excellent starting point for articulating the nature of the structure and hierarchy. {Quirk et al.: 8.2} suggest the presence of the underlying features of *position*, *duration*, *frequency*, and *relationship* in describing TIME. We can formalize these features by hypothesizing the presence of an underlying structure, an abstract class, containing variables (attributes or feature names) that have specific values in characterizing anything that has an associated TIME. This abstract class must be assigned some value for one or more of its features, but many features may remain without values (that is, undefined).

Following {Quirk et al.: 8.4, 8.51-77, 8.97-98}, we can tentatively hypothesize the features POSITION (denoting a point or period of time, and possibly also providing attention to another period of time), DURATION (evoking a starting and an ending position, which may be indefinite), SPAN (requiring a position, a duration, and a direction), FREQUENCY (usually implying or specifying a span and which is definite (period or occasion) or indefinite (usual, continuous/continual/universal, high, or low)), and RELATIONSHIP (involving a temporal sequence or position, and possibly growing into causal or comparitive relationships). It appears that a TIME structure may be associated with a clause, with a RELATIONSHIP expressing how the TIMEs of two clauses are related. Subordinating conjunctions essentially characterize the relation between the subordinate and main clauses, and generally characterize some of the temporal features associated with one or both of the clauses. The essential relationship, as {Quirk et al.: 15.26} suggest, is characterizing "the time of the situation denoted in the [subordinate] clause to the time of the situation denoted in the matrix clause." This encompasses **time_before**, **time_after**, and **time_overlap**, and sometimes **time_beginning** or **time_proximity**, that is, characterizing the position. Sometimes, the subordinating conjunction may additionally provide some information about the **duration** or the **frequency**. Thus, one temporal sense of *since* indicates that we associate a time beginning, a time duration, and a backward span to the subordinate clause and that we characterize the main clause as occurring after the subordinate clause. Below, we will identify specific definitional phraseology associated with each of these temporal components.

The structure of **causal** or **contingency** components is considerably more complex than the **temporal**. We first agree with Barker that the basic causal structure involves (or inherits) a temporal ordering (that is, "a cause temporally precedes its effect"). {Quirk et al.: 8.7} identify **cause**, **reason**, **purpose**, **result**, **condition**, and **concession** as the semantic relations involved here. We will use a **cause** -> **result** abstraction as the underlying structure involved in all these cases, with various features further elaborating either side of the relation or the nature of the relation itself. Thus, we will allow the left-hand side to be further characterized further as **reason** (**motivation**), **condition**, **circumstance**, **presupposition**, or **concession**; we will allow

the right-hand side to be further characterized as **purpose**; and, we will allow the relation itself to be further characterized as **condition** or **prevention**.

As a first approximation into the elaboration of the basic structure, we find five distinct major nodes in the type hierarchy:

> **Basic Causation**: **cause** has some **result** (Barker's *causation*), with prevention as a strong cause in which the result has a negation (Barker's *prevention*);
> **Point of View**: a personal and subjective assessment is made on one of the two sides, **reason** (cause position) or **purpose** (result position);
> **Purpose Fulfillment**: **result** (result position) may be the fulfillment of **purpose** (cause position);
> **Condition**: introducing uncertainty into the nature of the relation between the left-hand side and the right-hand side, **condition** (cause position) provides circumstances that **result** is achieved (Barker's *entailment*); **presupposition** (cause position) is equivalent to condition (Barker's *entailment*); bare **circumstance** (cause position) is the absence of any further characterization (Barker's *enablement*); and
> **Concession**: **concession** (cause position) provides circumstances that result is achieved despite, that is, the result of a cause is blocked or inoperative (Barker's *detraction*).

There are many nuances or variations on these themes, principally as further elaborations on **condition**. The several subordinating conjunctions that express these relations do so by assigning values to attributes of the elements of the basic structure. In addition, and quite importantly, the use of many subordinating conjunctions entails a sort of masking of these assignments. Knott unravels many of these masks by showing what transformations are involved in making the proper assignments of feature values.

3.3 Integrating Knott

Most of Knott's work deals with the underlying coordinating conjunctive structure, that is, the top of the hierarchy outlined above. Interestingly, many of these coordinative structures apparently provide alternative mechanisms for expressing temporal and causal structures that are frequently lexicalized into subordinating conjunctions. Therefore, understanding the coordinative structures and the transformations relative to the underlying propositional structure lends credence to Barker's hierarchy and provides mechanisms for its further elaboration as discussed above. We thus first need to recast Knott's results in terms of the hierarchies proposed above.

3.3.1 Source of Coherence

Knott describes this feature, SOURCE OF COHERENCE, through consideration of the lexical items *it follows that* (PRAGMATIC) and *as a result* (SEMANTIC), both of which are essentially coordinative in nature. The essential difference between lexical items which have one of these feature values has to do with the nature of the **cause - result** relation. Lexical items with the SEMANTIC value are articulating that it is only necessary for the reader to believe the relation

holds, whereas items with the PRAGMATIC feature are expressing relations that actually hold in the real world.

In Knott's analysis, subordinating conjunctions that have the feature SEMANTIC are *until*, *before*, *whereas*, and *when*. Thus, we can expect that these subordinating conjunctions will have senses where we would label the relation of our general structure as **reader-belief**. (Note that *before* has only temporal definitions in all dictionaries surveyed, while {Quirk et al.: 15.27} indicate these other sense extensions; I have inquired whether the addition of these senses to the dictionary is likely.) In Knott's analysis, subordinating conjunctions that have the feature PRAGMATIC are *unless*, *because*, and *in order that*. Thus, we can expect that these subordinating conjunctions will have senses where we would label the relation of our general structure as **real-world**. (These names are arbitrary; the essential point is that the nature of the relation itself, rather than anything pertaining to the left- or right-hand side, is provided with a feature value.)

3.3.2 Polarity

Knott describes this feature, POLARITY, through consideration of the lexical items *so* (POSITIVE) and *but* (NEGATIVE), both of which are essentially coordinative in nature. What Knott suggests is that the use of a word with this feature indicates that a defeasible statement of implication or rule of inference ($P \rightarrow Q$) is present and what is being communicated is the success (POSITIVE) or failure (NEGATIVE) of the rule. The crucial difficulty is recovering the rule, because the indication of failure makes the surface expression a transformation of the underlying rule. In particular, it is necessary to negate either the matrix or subordinate clause to recover the rule; the specific clause varies with the lexical item. In Knott's analysis, subordinating conjunctions that have the feature POSITIVE are *in order that* (introducing Q) and *because* (introducing P); in both cases, no transformation is necessary. In Knott's analysis, subordinating conjunctions that have the feature NEGATIVE are *even if*, *unless*, *until*, and *whereas*. In these cases, the subordinate clause is P and remains unaffected, while the matrix clause is Q and is negated to recover the rule of inference.

All these cases are instances of the **causal** relation. It seems that every causal subordinating conjunction needs to be marked for this feature, although it is mainly crucial for those that are NEGATIVE, where we may then more precisely characterize the argument structure of a text. These NEGATIVE subordinating conjunctions will be found in the hierarchy under Barker's *prevention* and *detraction* classes.

3.3.3 Pattern of Instantiation

Knott describes this feature, PATTERN OF INSTANTIATION, through consideration of the lexical items *admittedly .. but* (undefined for this feature) and *despite this* (BILATERAL), both of which are essentially coordinative in nature. In both cases, Knott suggests, there is a presumed rule of the form, $P_1 \wedge ... \wedge P_n \rightarrow Q$. The key issue is identifying the rule that is presumed. For the BILATERAL cases, there is no difficulty, with the matrix and subordinating clauses mapping directly into the rule. In Knott's analysis, subordinating conjunctions that have the feature BILATERAL are *unless*, *until*, *because*, *in order that*, *so*, and *provided that*.

The difficulty arises in the UNILATERAL case, because the rule is implicit and the consequent is not necessarily even recoverable from the text. To begin with, the subordinating clause must be negated. Then, the clause is added to the list of premises. The conclusion or consequent is not either the matrix or the subordinating clause. However, based on the examples Knott provides, it is possible that the conclusion comes before these premises and that the premises are thus stated as supportive of the conclusion. It is also possible that the conclusion is stated further on in the text, but signaled by a conclusionary cue phrase (such as *accordingly* or *hence*). In Knott's analysis, the only subordinating conjunction that has the feature UNILATERAL is *whereas*.

3.3.4 Rule Type

Knott describes this feature, RULE TYPE, through consideration of the lexical items *despite this* (CAUSAL) and *whereas* (INDUCTIVE), with the former essentially coordinative and the latter subordinative. In both cases, Knott suggests, there is a presumed defeasible rule, $P_1 \wedge ... \wedge P_n \rightarrow Q$, that is defeated. For lexical items labeled with CAUSAL, the presumed rule is recoverable from the text, with the matrix clause as the antecedent and the clause introduced by the conjunctive item as the consequent. For lexical items labeled with INDCUCTIVE, the presumed rule is not recoverable from the text. Rather, the lexical item signals (1) the presence of some underlying property over the objects in some class, (2) an inferential rule that inductively generalizes for this property, and (3) the failure of this rule. The subordinating clause presents contrary information precluding the generalization.

In Knott's analysis, subordinating conjunctions that have the CAUSAL feature are *if, even if, unless, until, because, in order that, before*, and *provided that*. Only *whereas* had the INDUCTIVE feature. The feature is uncertain for *when*, and undefined for *while*. (Since *when* seems to have distinct senses derived from *if* and *while*, it is possible that it has the CAUSAL feature in those senses derived from *if* and is undefined for those derived from *while*.) Thus, it would seem that Knott's analysis indicates that the nature of the inferential rule be labeled for this feature, with some cases where it is undefined.

3.3.5 Anchor

Knott describes this feature, ANCHOR, through consideration of the lexical items *despite this* (CAUSE-DRIVEN) and *unfortunately* (RESULT-DRIVEN), both of which are essentially coordinative in nature. Knott's analysis suggests that this feature pertains to the nature of the inferential rule, whether we are predicting the right-hand side of the rule from its premises (CAUSE-DRIVEN) or attempting to achieve the right-hand side of the rule through the premises (RESULT-DRIVEN). When this feature is used, there are several transformations that put the clauses into their appropriate positions in the inferential structure and assign feature values to the constituents. Because they are complex and pertain primarily to coordinating conjunctions, we will not go into detail for those uses.

In Knott's analysis, only three subordinating conjunctions are defined for this feature: *unless* and *until* are CAUSE-DRIVEN and *in order that* is RESULT-DRIVEN. In these cases, it seems useful to label the relation itself, as the basis for indicating that the argumentative structure of the

underlying text is purposive in nature, rather than simply expositional. These cases would correspond to the **Point of View** set in the hierarchy.

3.3.6 Focus of Polarity

Knott describes this feature, FOCUS OF POLARITY, through consideration of the lexical items *but* (COUNT) and *otherwise* (ANCH), both of which are essentially coordinative in nature. What does this mean? First, this allows us to make inferences about the nature of the matrix (S_A) and subordinate clauses (S_C), that is, the ones introduced by *but* and *otherwise*. Because both are BILATERAL CAUSE-DRIVEN, $S_A = A$, $S_C = C$ and $A \in P_1 \wedge ... \wedge P_n$ and $P_1 \wedge ... \wedge P_n$ is true and C' (the counterpart after the polarity transformation) = Q in $P_1 \wedge ... \wedge P_n \rightarrow Q$.

The fact that *but* is COUNT means that F = C and F' = C' (that is, the focus of polarity after the polarity transformation equals the counterpart after the polarity transformation) and that I = A and I' = A' (that is, the invariant is the anchor, and the invariant after the polarity transformation is the same as the anchor after the polarity transformation). The fact that *but* is NEG means that F is ¬F', which means that C' is ¬C. (See {Quirk et al.: 13.32}.) Hence, the negation of C (what occurs after the polarity transformation) is an expectation that has been violated. Now this bears a striking similarity to **concession**. We could rephrase the sentence, putting the matrix clause into the subordinate position with *although*, making the coordinate clause introduced by *but* the matrix clause. In Knott's analysis, other subordinating conjunctions that have this feature are *if, even if, because, provided that,* and *when*. Thus, we can expect that these subordinating conjunctions will have senses where we would label the left-hand side of the general structure as **concessive**.

The fact that *otherwise* is ANCH means that F = A and F' = A' (that is, the focus of polarity after the polarity transformation equals the anchor after the polarity transformation) and that I = C and I' = C' (that is, the invariant is the counterpart, and the invariant after the polarity transformation is the same as the counterpart after the polarity transformation). The fact that *otherwise* is NEG means that F is ¬F', which means that A' is ¬A. Hence, in the operative rule, the left-hand side of the rule contains ¬A; but in the instant case, the rule does not fire because A is true. In other words, we can transform a sentence involving *otherwise* into a **contingency**, specifically, a **condition**, with the negation of the matrix clause as the condition and the coordinate clause introduced by *otherwise* as the **result**. In Knott's analysis, subordinating conjunctions that have this feature are *unless* and *until*. Thus, we can expect that these subordinating conjunctions will have senses where the nature of the relation is identified as **conditional**, with the negation of the left-hand side as the **condition** and the matrix clause as the **result**.

3.3.7 Presuppositionality

Knott describes this feature, PRESUPPOSITIONALITY, through consideration of the lexical items *while* and *when* (PRESUPPOSED) and *meanwhile* (NON-PRESUPPOSED), where the issue has to do with the necessary temporal ordering of the clauses involved. All of these lexical items "convey information about temporal simultaneity." Thus, they correspond to the temporal feature **time_overlap**. The issue is whether the time span involved in the clauses is fluid or

allows arbitrary rearrangement. The solution is that lexical items having the PRESUPPOSED value do not allow rearrangement, but rather are expressing something of a necessary logical ordering: the subordinate clause must precede the matrix clause. In addition, Knott indicates that, for these lexical items, there must be preceding context that needs to be joined with the content of the clause introduced by the lexical item in reaching the **result** expressed in the matrix clause (with this preceding context identified as a *precondition*). For lexical items with the NON-PRESUPPOSED value, there is no logical ordering. In Knott's analysis, subordinating conjunctions that have the PRESUPPOSED feature are *if, even if, unless, until, because, in order that, before, whereas, provided that, when,* and *while*. No subordinating conjunctions have the NON-PRESUPPOSED value (supporting the thesis expressed above that all subordinating conjunctions involve more than the simple conjunction). It would appear, therefore, that this feature does not provide any additional information for our basic structure. However, Knott's suggestion of a precondition is interesting; it may be that no clause introduced by a subordinating conjunction is sufficient unto itself, but rather there are further presuppositions involved. Barker uses the sample sentence, "The file printed because the program issued a print command," as the prototypical example of a causation relationship. What Knott's analysis suggests is that there needs to be an additional presupposition, something like "The program contains code that will enable a file to be printed."

3.3.8 Modal Status

Knott describes this feature, MODAL STATUS, through consideration of the lexical items *if* (HYPOTHETICAL) and *when* (ACTUAL), both of which are essentially subordinative in nature. While *if* seems to be prototypically hypothetical, the analysis of this feature goes deeper and asserts that the more appropriate distinction is at a level characterizing whether the protagonist/writer knows the presupposition or precondition underlying the subordinating clause. In other words, as with the PRESUPPOSITIONALITY feature, lexical items marked with either value of this feature seem to require knowledge about the cause underlying the statement in the subordinate clause. In Knott's analysis, subordinating conjunctions that have the feature HYPOTHETICAL are *if, even if, unless,* and *provided that*. In Knott's analysis, subordinating conjunctions that have the feature ACTUAL are *until*, *because*, *before*, *whereas*, *when*, and *while*. The distinction is well drawn that the former provide no basis for determining whether the presupposition or precondition may be true, whereas the latter do seem to indicate that provide greater strength for the causal relationship. All subordinating conjunctions marked with either of these features appear to fall under the **condition** branch of the type hierarchy.

4 Next Stages of Analysis

This is a brief overview of the next steps of the analysis of the subordinating conjunction digraph. Based on the preceding section, feature names and feature values need to be developed so that each category is adequately characterized. Next, components of dictionary definitions are correlated with these feature names and values. (This process has been initiated, with many patterns emerging already and with success quite likely.) As these feature names and values are assigned to individual senses of the subordinating conjunctions, they will be used to break the definitional cycles that remained at the end of section 2, so that a stricter hierarchy will emerge. It should be understood that the subordinating conjunctions have a greater number of senses than

envisioned in either Barker or Knott (based on much more substantial and comprehensive corpus evidence from Merriam-Webster), so that the final hierarchy will emerge as more elaborate and comprehensive. Moreover, the prediction is that the resulting hierarchy will have nodes that are written as synsets (subordinating conjunction concepts that can be labeled and verbalized in more than one way).

Notes

1. The main entries and their definitions (including label numbers, usage labels, and usage notes) were entered into DIMAP, a set of utilities for creating and maintaining dictionaries for natural language processing, available from CL Research (http://www.clres.com). The analyses described herein were completed with functionality available in DIMAP.