Word structure parsing

Anna-Maria Di Sciullo

In this paper I will consider the relation of the grammar and the lexicon as well as the incorporation of grammatical principles in a morphological parser. These two topics are related to lexicography/lexicology to the extent that they involve the form and properties of dictionary entries, in particular the representation of categorial and semantic (thematic) information, as well as the properties of a formal lexicon in computational lexicography.

The first section gives a restrictive view of the lexicon and the lexical entries, and specifies the relation of the lexicon to the grammar. The second section introduces the elements, operations and principles which are part of what I will call "word structure grammar". The third section deals with the incorporation of this theory into a morphological parser. The last section shows the consequences for computational lexicography.

1. The lexicon

The view that I will develop limits the lexicon to a list of marked items; that is, items whose properties cannot be entirely derived from the word structure grammar. These items are morphemes, as well as idiomatic expressions. We will refer to them in terms of "listeme", as suggested in Di Sciullo and Williams (1987). Some examples of listemes are:

(1) rudiment, read, blue
- ary, -able, -ion
transmission, ovation
bite-the-dust, heart-breaker

The lexicon of a given language L consists of the set of the listemes of L, each listeme being associated with its specific set of lexical properties, including categorial, contextual, and thematic properties. The categorial properties of a listeme are expressed by means of lexical categories (e.g. N, V, A) which can be further defined in terms of grammatical features; however, we will leave this question aside for now. The contextual property can be stated in terms of contextual features, such as [N_] or [_V], which indicate the categorial selection of the listeme. Thematic properties are expressed in terms of thematic grids, in the sense of Stowell (1981); that is, in terms of sets of thematic roles (AGENT, PATIENT, etc.), one of which is the external argument of the listeme's argument

structure. We underlined the external argument in the following partial lexical entries:

(2) rudiment: N, [__], (R)
-ary: A, [N__], (TH)
-able: A, [V__], (TH)

This view of the lexicon is a way of formalizing, for each listeme, specific lexical information, such as their context of occurrence (contextual features), as well as one aspect of their semantics, the semantic roles of their arguments (thematic grids).

The idea that the lexicon is exclusively the list of listemes, and not the list of all the words of a language, is a way of expressing a basic intuition about our knowledge of the language: the fact that some words are learned whereas others need not be. The former are listed, whereas the latter are derived by the word structure grammar, which we now turn to immediately.

2. Word structure grammar

We will refer to the objects derived by the word structure grammar as "morphological objects". These objects are not listed, as their properties are entirely derivable. Examples of morphological objects are given below:

(3) rudimentary, bluish, readable

The word structure grammar derives the categorial and thematic properties of the morphological object. It consists of a set of listemes, a set of categories (N, V, A), and a set of thematic roles (AG, TH, R, LOC, etc.). The grammar includes a set of operations that i) build morphological trees, ii) identify the category of the root, and iii) calculate the thematic grid of morphological objects. We will distinguish three different operations involved in these processes: structure building operations, percolation, and binding. We will define each of them below. Furthermore, we will assume the existence of morphological principles that activate or block the operations, given their potential over-generating capacity. Three principles will be discussed: the generalized head principle, thematic distinctivity, and thematic discharge. Let us first consider the operations.

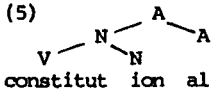
2.1. Operations

2.1.1. Structure building. The structure building operations are very simple; they create binary branching trees, given the general meta-rule (4).

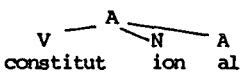
(4) $X^0 \rightarrow X^0 \ X0$
where $X^0 = \{V, N, A\}$

Binary branching structures preserve the morphological properties of listemes at all levels of the morphological tree. (5) is a well formed morphological tree, according to our theory. This is not the case for (6), which is not an instantiation of (4). Furthermore, (5) distinguishes the sub-tree (7) which is a morphological object as well. This is not the case for (6).

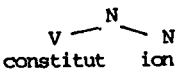
(5)



(6)



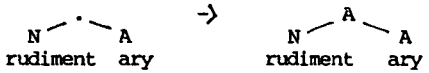
(7)



The fact that the structure of morphological objects is binary branching, follows from a condition on the representation of selectional properties of listemes (see Di Sciullo 1986 for discussion). We will not elaborate upon this point here, since it is not essential to our topic.

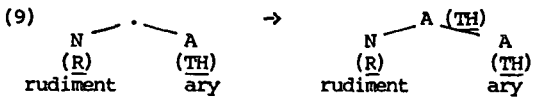
2.1.2. Percolation. The identification of the category of the root in the morphological tree is accomplished through percolation, as follows:

(8)



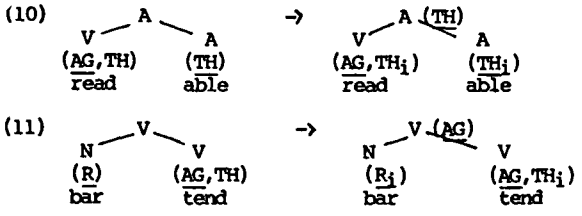
Percolation is also involved in the identification of the external argument of the argument structure of the root, since it is generally the case that the theta role of the categorial head (which we define below) becomes the external argument of the morphological object, as in *rudimentary* for instance:

(9)



2.1.3. Binding. Essentially, binding is a general operation relating two elements such as two thematic roles in a morphological tree. It is necessary to distinguish between two different cases of binding, since cases such as *readable* involve the binding of two identical thematic roles, whereas cases such as *bar-tend* involve the binding of two distinct thematic roles. In the latter case, the internal argu-

ment of the verb *tend* is discharged within the compound and the co-indexed thematic roles cannot percolate. In the former case, the two co-indexed, identical thematic roles, count as a single thematic role which percolates.



2.2. Principles

The operations defined in the preceding sections are elementary operations which may over-generate. We will present three principles limiting over-generation: the generalized head principle, thematic distinctiveness, and thematic discharge principles.

2.2.1. Generalized head principle. If nothing limits percolation, for instance, an incorrect result may arise with respect to the identification of the category of the external argument of the root. Up to this point in our proposal, nothing could prevent this operation from deriving an N rather than an A in structures such as (8), or prevent the *R* thematic role from becoming the external argument in (9). In order to limit the effects of percolation, we will assume the existence of principle (12) and the associate parameter (13) for English.

(12) *Generalized head principle*:
The head, with respect to a grammatical property (categorial, thematic, etc.), is in a fixed position within a morphological object.

(13) *Parameter (English)*:
The head, with respect to a given grammatical property, is the rightmost element with respect to that property.

Principle (12), in conjunction with parameter (13), gives the desired results. In the case of (8), it blocks the percolation of the N categorie to the root node; in the case of (9), it blocks the percolation of the thematic role *R*, thus leaving only one possibility for both the identification of the category and the external argument of the root.

2.2.2. Thematic principles. Let us now consider the binding operation. In 2.1.3., we defined this operation in very general terms in order to account for two cases of binding as exemplified in (10) and (11). According to our definition, binding

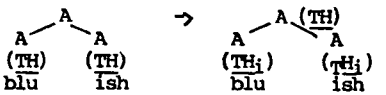
is an operation that co-indexes two elements in the morphological tree. Stated in such general terms, nothing prevents any two elements from being co-indexed. In order to limit the effects of binding, let us assume the existence of principles (14) and (15).

(14) *Thematic distinctiveness*:
Each element of a thematic grid must be thematically distinct.

(15) *Thematic discharge*:
Every thematic role that can be discharged in the tree, must be.

Principle (15) forces binding to apply in (11). Consequently, the co-indexed elements cannot percolate. Principle (14) activates binding in (10) which is not a case of argument satisfaction, but rather the binding of two identical theta roles; in a sense they constitute a single discontinuous thematic role that percolates to the root node. Note that principle (14) also prevents the percolation of more than one thematic role in the following cases:

(16)



3. Word structure parser

In this section we will briefly present the word structure parser that we have formulated, and which incorporates the principles and operations described in the preceding sections.

The general structure of the parser consists of two modules: a segmentor, including a lexicon and a list of listemes with their lexical properties, and a morphological parser. We will focus on argument structure parsing of morphologically complex words, after affix stripping operations are performed by the segmentor SEGMENT.

In brief, SEGMENT calls the lexicon and gives as a result, for a given morphological object, the list of its listemes (root, affix) with their categorial and thematic properties. SEGMENT is written in lisp, and is called by MORPHO PARSE, the morphological parser, when a morphological object is to be parsed.

The structure of MORPHO PARSE consists of a set of procedures that create, attach, and label nodes in order to build binary branching trees, as well as procedures that calculate the thematic grid of morphological objects. The calculus integrates the morphological principles formulated above; that is, the generalised head, thematic distinctivity, and thematic discharge. The parser includes a set of nodes (active node, new node, inactive node) and a two-cell buffer, in order to identify the head with respect to a given grammatical property. No push-down

stack is actually needed for the morphological parse. The new node is equivalent to the top of push-down stack. The program is also written in lisp.

The morphological parser takes a morphological object as its input, and gives as the output the structure of the word: a binary branching tree with its proper categorial and thematic properties. Thus, the morphological objects need not be listed in the lexicon, which is reduced to listemes.

The parser proceeds from left to right, and is strictly deterministic; that is, it does not simulate non-determinism by allowing parallel derivations or back-tracking. It excludes the possibility of multiple structures for the same word, or the possibility of destroying structure. For each morphological object, the parser derives a unique structure.

The parser may be thought of as a human parser in the sense that it possesses a built-in knowledge of the grammar of word structure. It can just like any human, compute the properties of morphological objects without having learned them, as well as create new morphological objects.

With respect to the relation between linguistic and parsing theories, our work shows that the interaction of these domains gives interesting results for the analysis of morphological objects. The linguistic theory that we have described above, allows for a principled account of the categorial and thematic properties of morphological objects. The word structure parser that we have formulated embodies the operations and principles of the grammar, and gives the correct results in the derivation of the categorial and thematic properties of morphological objects.

4. Consequences

In this section we will point out some of the consequences of our work, briefly presented in sections 2 and 3 for computational lexicography.

As a point of departure, we will stress the fact that there are basic relations between lexicography and lexicology. Let us say that lexicography is applied lexicology, and lexicology is a sub-area of linguistics dealing with the analytic study of the form and meaning of words. Given these general assumptions, our work has clear implications for lexicography/lexicology since it provides a linguistic basis for the identification of some properties of the form and meaning of words. And this can be applied to lexicography.

Our work presents a way of formalizing specific aspects of the form and meaning of words, given the theoretical difference between listemes and morphological objects that we suggested. In particular, information such as the context of occurrence of a given word and the semantic roles of its arguments may be stated in a non-ambiguous fashion. Of course such information is present in ordinary dictionary entries, but not in a formal manner; it can be inferred from a list of examples where the word is used.

Consider for instance the following entries taken from THE OXFORD ADVANCED LEARNER'S DICTIONARY OF CURRENT ENGLISH:

(17) **happiness**: n. the state of being happy; good fortune.

(18) **happy**: adj, (-ier, -iest) 1. Fortunate; lucky; feeling or expressing pleasure, contentment or satisfaction. *Their marriage has been a very happy one.* **as happy as the day is long [as a king]**, very happy. 2. (as a polite formula) pleased. *I shall be happy to accept your invitation.* 3 (of language or conduct) well suited to the occasion; well adjusted to the conditions, as *a happy thought* [*idea, suggestion*].

Furthermore, our work provides a view of the interaction of the grammar with the lexicon allowing for a restricted formulation of the latter. In fact, not all the actual words of a given language must be stated in the lexicon. Only the listemes must be. The lexical properties of morphological objects (categorial and thematic) are derived by the grammar. This allows us to distinguish between the learned versus the derived knowledge of language. Thus our theory is related to the area of psycholinguistics are well.

This view is appealing in the area of computational lexicography for obvious reasons, as it minimizes the lexicon and optimizes the algorithmic aspect of the parser.

The form of the lexical information presented in standard dictionaries such as OALD is not quite appropriate for computational lexicography. Current dictionaries provide no formal access to the thematic properties of a word. Consider again the entries in (17) (18). The semantic properties of these words can only be inferred through human knowledge.

Furthermore, even though current dictionaries do provide categorial and contextual information by means of abbreviations, such as V, N, A, Vt etc., they fail to give direct access to the relations between words with respect to categorial, contextual and thematic properties – for instance the fact that *happy* and *happiness* are categorially and thematically related. These relations are given by the word structure grammar within our theory and they are presented in an algorithmic form in our word structure parser. As a consequence, listemes such as *happy* are listed words, while the formal properties of morphological objects such as *happiness* are derived.

The study of the properties of listemes and morphological objects can be used to define the formal information that must be available. For instance categorial, contextual and thematic properties must be part of the formal elements of lexical entries of an on-line dictionary. As in the following partial entries:

(19) happy: A, [__], (TH), ...

ness: N, [A__], (R), ...

References

Cited dictionary

OXFORD ADVANCED LEARNER'S DICTIONARY OF CURRENT ENGLISH (OALD)
A.S. Hornby et al., London: Oxford University Press (1948/61).

Other literature

Berwick, Robert C./Weinberg, Amy S. (1984), "Parsing Efficiency , Computational Complexity and the Evaluation of Grammatical Theories", in: *Linguistic Inquiry* 13.2: 165–192.

Di Sciullo, Anna-Maria (1986), "Configurational Morphology", ms. UQAM.

Di Sciullo, Anna-Maria/Williams Edwin S. (1987), *On the Definition of Word.* Linguistic Inquiry Monograph 14, Cambridge, MA: MIT Press.

Marcus, Mitchel (1980), *A Theory of Syntactic Recognition for Natural Language.* Cambridge, MA: MIT Press.

Stowell, Timothy (1981), *The Origin of Phrase Structure,* Unpublished MIT Doctoral dissertation, Cambridge, Mass.