Introduction

Paul Boucher and Marc Plénat

University of Nantes and University of Toulouse-le-Mirail

The title of this collection of articles, *Many Morphologies*, is an allusion not only to the variety of morphological problems discussed, but also to the diversity of formal solutions offered by the authors. The papers examine Polish derivational morphology, French and English compounding, pluralization in Luiseño and Somali, and more, and though almost all of the papers are written in the generative grammar framework, there are considerable differences between the formal methods they apply to the problems at hand. We also include two papers which discuss computational morphology, one an overview of the current state of the field and the other an in-depth study of a current research project.

The title also reflects the very complex nature of morphological phenomena, often misunderstood or over-simplified in morpheme-based approaches. As Joseph Emonds points out at the start of his paper (p. 235): "The main goal of the generative [grammar] enterprise has always been to elaborate a cross-linguistic syntactic theory, yet a phenomenon at the center of the best traditional syntactic descriptions, namely bound morphology, remains largely peripheral to the domain of notable generative achievements." This volume offers original and highly articulated contributions to current research in the field and will certainly help to remedy this state of affairs.

The creation and development of this book grew out of a research project funded by the French Ministry of Research entitled "The Structure of the Lexicon." A Summer School in Morphology held at the Université de Nantes in 1997 and a Morphology Workshop in September 1998 brought together some of the leading figures in American and European morphological studies. In one of her lectures at the Summer School, Hagit Borer argued for the *a priori* "desirability of reducing morphology to syntax," following Rochelle Lieber's (1992: 21) claim that "a truly simple theory of morphology would be one in which nothing at all needed to be added to the theory of syntax in order to account for the construction of words." In this volume (p. 236), Joseph Emonds goes so far as to suggest that it might be just as fruitful to turn the statement around: "one can equally well claim that many phenomena seen as syntactic should be subsumed under properly expanded morphological analysis." Whatever is

ultimately the most satisfactory solution, this tension between word-internal and word-external structure underscores most of the papers presented here.

The opening paper by Anna Maria Di Sciullo and the closing paper by Joseph Emonds explicitly address this question and offer important new perspectives on the goal of unifying these two sub-fields of grammar. Di Sciullo claims that asymmetry is a fundamental property of the primitives, the operations, and the conditions of the grammar. This property, which she argues is shared by both morphological and syntactic objects, is formulated as the Strict Asymmetry Hypothesis: "grammatical relations are asymmetrical." This means that symmetrical relations – bare sisterhood relations or symmetrical c-command relations – are precluded from the grammar. The two basic operations in Di Sciullo's framework are SHIFT, which derives complex categories from more elementary ones, and LINK, which relates categories in derivations and representations. By positing that affixes SHIFT only with root configurations where given asymmetrical relations must hold, Di Sciullo can capture the particular selectional properties of each affix using general properties of derivation. The restrictions on the derivation of *-er* agentive nominals in English, for instance, follow directly from the SAH and the SHIFT operation: *-er* selects a configuration with certain elemental asymmetrical properties, such as the presence of an A-Spec relation in the complement domain of the affix. Di Sciullo argues that this sort of asymmetry holds across the board for derivational affixes, regardless of their categorial feature, and discusses derivations with *-able, -ify,* and *-ize*. Crucially, she shows that, whereas morphological SHIFT relates affixes to asymmetrical root structures and not categories, syntactic SHIFT is not subject to argument structure asymmetry. For example, when a verb merges with a nominal expression in a syntactic derivation, the asymmetrical argument structure of that noun is not visible to the verb that selects it. A syntactic head merges with a category, not with a category which has a particular argument structure.

Emonds takes a radically different approach to the problem of word formation. He claims that unifying syntax and morphology requires re-examining current theories of syntax and paying serious attention to the internal structure of words. Emonds demonstrates that incorporating principles for compounding and lexical insertion into the syntax can "render superfluous much of what is taken as autonomous morphology" (p. 236). His approach consists of redefining the principles of word-internal as well as word-external syntax on the one hand, and correlating the functional/semantic load of morphological items with the level of their insertion into the derivational process on the other hand.

In Emonds' theory of tri-level lexical insertion, grammatical affixes can be inserted at different levels of the derivation with very different grammatical consequences. They may combine lexically with open-class items at the deep level of insertion. At this initial stage of the derivation, only purely semantic features are considered. Grammatical affixes can also be inserted during the syntax, and contribute to LF, or they can be inserted in PF if they don't contribute to LF.

For example, take the well-known problem of the ambiguity between event and result readings for English *-ing* (or *-ment*) nominals. These are not simply two different interpretations, but correlate with different morpho-syntactic properties, such as the ability to be quantified or pluralized, to take adjectives expressing physical properties, to form productive N–N compounds, or to occur with temporal and other adjectival *of*-phrases. These properties can be explained, argues Emonds, by different levels of insertion. Result nominals are formed at the outset of the derivation, at the deep or lexical level. Event nominals, on the other hand, are inserted in the syntax before Spell-Out. The suffix seems to be inert at the deep level for these constructions, which have many verbal properties. Nominalizers like *-ing* can be inserted after syntactic processing of the DP domain, although prior to any syntactic operations affecting the domain containing DP. This ensures that from insertion of *-ing* onwards, the event nominal has an N head and will therefore be selected and distributed like a DP in larger syntactic domains. The third grammatical use of the *-ing* suffix, the gerundive function, can be explained in terms of late or PF insertion. Emonds includes in this level the case-assigning preposition *of*, uninterpreted expletive pronouns, and the dative-marking P *to*. These items, he argues, have in common the fact that they contain no semantic features and moreover that any interpretable cognitive syntactic feature in their lexical entry is optional.

Following Di Sciullo's article are two papers dealing with the semantic structure of intransitive verbs. In the framework of Pustejovsky's Generative Lexicon, Christian Bassac and Pierrette Bouillon examine the internal structure of English verbs exhibiting the middle construction. Then Bożena Cetnarowska investigates tests used to distinguish between unergative and unaccusative predicates, illustrating her discussion with data from English and Polish.

After reviewing the data and the various formal solutions currently proposed in the literature, Bassac and Bouillon show that a Generative Lexicon (GL) analysis accounts for the properties of middle constructions, which otherwise remain mysterious. In this approach, the syntactic behavior of a word can be calculated from its qualia structure, that is, from the sub-part of its lexical sense which links arguments and events together. Instead of simply listing the different senses of a word, the GL approach manipulates the lexical sense structure to derive a given sense in context. Thus, claim Bassac and Bouillon, the polymorphism of verbs showing the middle/transitive alternation can be explained in terms of their complex semantic type. These verbs display event structures with two sub-events, a prominent event encoded as the agentive role and a resultative state encoded as the formal role. The fact that this structure is a complex one allows two different syntactic projections (under certain conditions). The transitive version will be the lexically driven projection of the initial sub-event, whereas the middle variant will be a forced projection of the final resultative state. Given this lexical potential, Bassac and Bouillon carefully delineate the conditions that must be met for the projection of a middle construction.

Cetnarowska investigates the unaccusative/unergative split, focusing on unaccusativity mismatches – cases where a particular verb is identified as unaccusative by virtue of its occurrence in a given syntactic construction, yet fails to satisfy certain additional diagnostics for unaccusativity, such as deriving an adjectival past participle. Early studies of the intransitivity split suggested that processes of derivational morphology, such as *-ee* and *-able* suffixation or *un-* and *re-* prefixation in English, are sensitive to the unaccusative/unergative distinction. However, more recent studies have either questioned the validity of such tests, or have failed to investigate conflicts between predictions of affixation processes and syntactic tests. Cetnarowska argues that in fact the results of syntactic tests for verb classification are not always as clear as has been assumed and should therefore not be regarded as either more reliable or more informative than derivational tests. While gaps in derivational paradigms make it difficult to get clear and unambiguous results, she shows that many of these gaps can be explained if one takes into consideration a certain number of semantic, morphological, and pragmatic restrictions on derivational processes.

The papers by Susan Steele and Jacqueline Lecarme both address the problem of plural formation, but from very different formal perspectives. Again, the syntax–morphology dichotomy is at the heart of the difference in the two approaches.

Steele positions her study of number inflection in Luiseño in the tradition of Aronoff (1994) and Anderson (1992), who argue for a "processual" view of morphology. Such a theory, in Steele's words, "focuses on, and attempts to account for, the kinds of relationships that can exist between and among stems and words" (p. 82). In this perspective, stems and words involve a phonological part, a semantic part, and a syntactic part. Each of the three parts involves a set of features and associated values. Steele shows that an information-based approach to morphological processes can adequately represent the relationship between the Luiseño plural morph *-um* and various morphological operations. By demonstrating the complexity of these relationships, notably the fact that there is no simple map between the morph *-um* and the feature [pl], she shows that a morpheme-based approach is inadequate and cannot deal with complexity of this sort. Steele solves the particular problem posed by Luiseño plural morphs quite neatly and provides support for the fundamental insight of processual approaches to morphology – that morphology is a set of relationships rather than a set of morphemes.

Lecarme, on the other hand, claims that the properties of Somali plurals are consistent with a purely syntactic approach to word formation. She examines the concept of gender polarity in Somali in the framework of Halle and Marantz's (1993, 1994) Distributed Morphology. The central thesis of this approach is that there is no separate component for lexical operations, no need for a distinction between derivational and syntactic morphology; instead, all morphology is syntactic. Syntax does not operate on words, but on (fully specified) bundles of

formal features. Only at a later stage of the syntactic derivation of a linguistic expression are the bundles of formal features linked with the (underspecified) pieces of phonology: stems, affixes, and words.

In order to sort out the puzzling problems posed by Somali gender polarity, Lecarme first proposes a revised classification of Somali plurals. She redefines the organizing principle behind the classification, grouping the plural forms according to whether they correspond to a "zero suffix" – a change in tonal pattern rather than in form, to a suffix containing a consonant copied from the stem, to a vocalic suffix, or to what she terms complex suffixes. This leads her to propose a new generalization which calls into question the notion of polarity, the idea that "if under certain conditions A become B, B will become A under the same conditions" (Meinhof 1912: 18).

Secondly, Lecarme proposes a solution to the problem that has far-reaching consequences for the very concept of number in nouns, namely, that gender is a feature of the plural suffix itself, rather than something inherited from the noun stem. This obviously challenges traditional views on the boundary between derivational and inflectional morphology, but is not without precedence in the generative framework. For instance, Ritter (1991), in her analysis of Hebrew plural morphology, has claimed that gender is specified both on the noun stem and on the plural affix. Carstens (1991, 1993) has taken a similar approach in her analysis of Bantu nominal class morphology. The Distributed Morphology framework allows Lecarme to come up with a tenable answer to some of the problems raised by these and other studies.

Luigi Burzio's paper addresses a series of questions that differ quite sharply from those discussed in the other papers in this volume. He treats morphology in terms of its relations with phonology and the lexicon rather than with syntax. Following earlier studies by Bybee (1985, 1988, 1995), Burzio proposes to reduce morphology to a set of surface-to-surface relations in the overall context of a system of violable parallel constraints within the Optimality Theory framework.

Burzio argues that word-to-word relations can be defined in terms of a theory of Gradient Attraction, which states that the overall structure of a word is influenced by that of other similar words in the lexicon. The fact that attraction does indeed operate between surface forms is especially clear when a particular derived form carries traces of the influence of other surface forms, as in the case of *remédiable,* which is derived from the base *rémedy*, but whose stress pattern conforms to that of the derived word *remédial*. Burzio's analysis of this type of problem suggests that the word formation rules (WFRs), which have been thought to provide the phonology with inputs since Aronoff 1976, are somewhat redundant. WFRs build relations between words via their underlying representations, but in this analysis the information supposedly provided by these WFRs can be recovered in the surface forms. Along with eliminating WFRs, Burzio forces us to rethink morphology, to consider it as something other than a distinct

module alongside phonology. Concepts such as morpheme and allomorph emerge quite naturally from the new formal framework proposed by Burzio, wherein weak differences tend to be neutralized. Two representations that do not differ in meaning have no reason to differ in form unless there are some unusual circumstances. Paradigmatic uniformity is therefore the rule. On the other hand, since representational entailments, like all constraints, can be violated, allomorphs are always possible when required by the circumstances.

Burzio's paper stands apart from the other papers in the volume in that he argues that morphological rules emerge from a lexicon, which, even though it is constrained by universal constraints, is full of idiosyncrasies, whereas most morphologists try to derive the lexicon from a set of rules or principles. This traditional opposition between rules and lists cuts across the article by Nabil Hathout, Fiammetta Namer, and Georgette Dal. In this article, the authors present the initial results of their research into the semi-automatic generation of a constructional (i.e., derivational) database for French. When completed, this database should include, for each of some 70,000 lexical units: the lemma, its grammatical category, its constructional analysis, as well as its derivational history and a gloss in natural language. Building this database has led the authors to develop two separate programs, DéCor and DériF, which are described in the article and illustrated by the treatment of the suffix *-able*.

The two programs being developed use very different approaches. DéCor is based on the Network Model developed by Bybee (1988, 1995), and uses a statistical approach. Its aim is solely to pair formally similar lexical units that belong to the same referential set; i.e., to relate each derived form, whose status is deduced from the frequency of its final and/or initial sequence, to the corresponding lemma. This is carried out through various sorting functions based on frequency and economy. DéCor can thus be used for any language, or at least for those languages which resort to concatenated morphemes.

DériF, on the other hand, implements linguistic hypotheses in the framework of the constructional morphology for French developed by Danielle Corbin and her team (cf. Corbin 1991, forthcoming). This theory is diametrically opposed to Bybee's model used by DéCor. These linguistic hypotheses allow DériF not only to pair off the derived form and its base, but also to propose a bracketed representation of the former as well as a semantic gloss. As expected, the comparison of the results of the two programs shows the superiority of the second over the first, at least as concerns the pairing operation. DéCor is handicapped in only being able to search for the correct base in the available lexicon, which prevents it, for instance, from identifying the possible form *°perturbable* as the base form of *imperturbable*, or the bound root *sec-* as the base form of *secable*, or even the short, isolated allomorph *buv-* (from the verb *boire*) as the base for *buvable*. However, it is also clear that there is much work to be done before an exhaustive description of the morphology of French is completed in any given theoretical

framework. Toward that goal, these two programs appear to be both compatible and complementary.

The article by Béatrice Daille, Cécile Fabre, and Pascale Sébillot surveys various resources and applications of computational morphology. They begin by describing the existing techniques for parsing and stemming in the natural language processing framework, then go on to demonstrate, using a number of concrete examples, how these techniques can be used to acquire morphological knowledge from corpora or to incorporate such knowledge into a lexical database.

On the side of resources, the authors give several short and insightful descriptions of known lexical databases that provide inflectional or derivational information, such as DELAS and the other databases for French developed by the LADL team, MULTEXT, and the CELEX database, as well as morphological systems such as the stemmer developed by Porter (1980) or the parser built by Karttunen (1983), which implements Koskenniemi's (1983) two-level morphology model. More recent work is also presented, such as programs which aim at automatically extracting rules and/or morphological families from thesauri or from corpora without resorting to linguistic information or by using very little linguistic information. For example, while Porter's stemmer is based on a set of transformational rules like *-ational* → *-ate* (which transforms a word like *relational* into *relate*), Xu and Croft (1998) propose a set of techniques for building morphological families without resorting to linguistic knowledge. The main idea in their work is that in a given corpus, words that should be grouped together in the same family are likely to co-occur frequently. It is therefore possible to form families by grouping those members of a set of candidates that tend to co-occur regularly in the corpus. These techniques can correct the sort of faulty groupings that result from simple formal comparisons. It is unlikely, for example, that *couper* 'to cut' and *coupable* 'guilty' will co-occur frequently within the same segment of text. These techniques can also remove from a given family those candidates whose meaning no longer has a transparent relationship to the base meaning. For instance, *décidément* 'without a doubt' should not be included in the family built up around *décider* 'to decide.' Learning techniques of this sort are very promising for future work in the NLP field.

For applications, the authors highlight linguistic annotation of corpora and lemmatization, building of semantic lexicons, terminology acquisition, detection of term variation, and document retrieval. In all of these fields, morphological cues can be essential. For example, researchers have only recently begun to take serious interest in the verbal variants of nominal terms. Locating such variants usually entails identifying syntactic phrases in the corpus which contain two terms from the same family, one nominal, the other verbal (cf. the conceptual equivalence of *stabiliser les prix* 'stabilize prices' and *stabilisation des prix* 'price stabilization'). Such morphological similarity is, however, far from sufficient, as shown by such pairs as *introduction d'un gène* 'introduction of a gene' and

introduire dans un gène 'introduce in a gene.' Other criteria are needed and the authors discuss two recent studies which shed some interesting light on this problem: Fabre 1998 and Fabre and Jacquemin 2000. These researchers show how predictions based on the conservation of argument structures can be used to achieve a finer-grained detection of variants. This is only one of many examples which should convince the reader of the importance of using morphological information in natural language processing.

There is often a considerable gap between the degree of technical sophistication of morphological theories like the ones presented in this volume and the somewhat rudimentary character of the morphological knowledge used in natural language processing systems. Techniques based on learning strategies often prove to be much more efficient than those using theories and descriptions drawn from formal linguistics. Morphology, as a field of scientific investigation, may not have reached a sufficient point of maturity, nor attained a sufficient degree of empirical coverage to contribute significantly to the implementation of such systems. It turns out in fact that the derivational and inflectional paradigms of a language are much too important in the overall organization of a language to be neglected by automatic processing systems. The recognition of this fact does not come easily to theoreticians who would like to demonstrate the social usefulness of their particular theory. However, it could turn out to be a very useful stimulus for future research. The resources and tools developed for computerized applications could very well contribute significantly to theoretical speculation. The current developments in the field of natural language processing will probably play the same kind of role in advancing the field of morphology as the major technological developments played in the history of the natural sciences. The quantity of data currently available for treatment is far greater than anything that morphologists could dream of ten or twenty years ago. This unprecedented extension of our empirical knowledge will no doubt contribute to the birth of many new morphologies to come.

References

Anderson, Stephen M. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press.

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Cambridge, MA: MIT Press.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, Joan L. 1988. Morphology as lexical organization. In *Theoretical morphology*, ed. Michael Hammond and Michael Noonan, 119–141. San Diego, CA: Academic Press.

Bybee, Joan L. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10.5: 425–455.

Carstens, Vicky. 1991. *The morphology and syntax of Determiner Phrases in Kiswahili*. Ph.D. dissertation, University of California at Los Angeles.

Carstens, Vicky. 1993. On nominal morphology and DP structure. In *Theoretical aspects of Bantu grammar*, ed. S.M. Mchombo, 151–180. Stanford, CA: Center for the Study of Language and Information.

Corbin, Danielle. 1991. Introduction. La formation des mots: structures et interprétations. *Lexique* 10: 7–30.

Corbin, Danielle. forthcoming. *Le lexique construit*. Paris: Armand Colin.

Fabre, Cécile. 1998. Repérage de variantes dérivationnelles de termes. Technical report, Carnets de grammaire, Équipe de Recherche en Syntaxe et Sémantique UMR 5610, CNRS et Université de Toulouse-Le Mirail.

Fabre, Cécile and Christian Jacquemin. 2000. Boosting variant recognition with light semantics. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, 264–270. San Francisco: Morgan Kaufmann Publishers.

Halle, Morris and Alec Marantz. 1993. Distributed Morphology and the pieces of inflection. In *The view from Building 20: Essays in honor of Sylvain Bromberger*, ed. K. Hale and S.J. Keyser, 111–177. Cambridge, MA: MIT Press.

Halle, Morris and Alec Marantz. 1994. Some key features of Distributed Morphology. *MIT Working Papers in Linguistics* 21: 275–288.

Karttunen, Lauri. 1983. KIMMO: A general morphological processor. *Linguistic Forum* 22: 163–186.

Koskenniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Ph.D. dissertation, University of Helsinki.

Lieber, Rochelle. 1992. *De-constructing morphology*. Chicago: University of Chicago Press.

Meinhof, Carl. 1912. *Die Sprachen der Hamiten*. (Abh. des Hamburgischen Kolonialinstituts), Hamburg.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program* 14: 130–137.

Ritter, Elisabeth. 1991. Two functional categories in Noun Phrases: Evidence from Modern Hebrew. In *Perspectives on phrase structure,* ed. S. Rothstein, 37–62. New York: Academic Press.

Xu, Jinxi and Bruce W. Croft. 1998. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 16.1: 61–81.

This is the introduction to *Many Morphologies*, edited by Paul Boucher, published in 2002 by Cascadilla Press. To purchase *Many Morphologies* or to find out more about the book, please visit our web site at http://www.cascadilla.com/manym.html or contact us:

Cascadilla Press
P.O. Box 440355
Somerville, MA 02144
USA

phone: 1-617-776-2370
fax: 1-617-776-2271
e-mail: sales@cascadilla.com
http://www.cascadilla.com