

THE ENCYCLOPEDIA OF
LANGUAGE
AND
LINGUISTICS

Volume 5

Editor-in-Chief
R. E. ASHER
University of Edinburgh, UK

Coordinating Editor
J. M. Y. SIMPSON
University of Glasgow, UK

1994



PERGAMON PRESS
OXFORD • NEW YORK • SEOUL • TOKYO

- Aronoff M 1976 *Word-Formation in Generative Grammar*. MIT Press, Cambridge, MA
- Aronoff M 1988 Head operations and strata in reduplication: A linear treatment. *Yearbook of Morphology* 1: 1-16
- Bar-Hillel Y 1953 A quasi-arithmetical notation for syntactic description. *Lg* 29: 47-58
- Bar-Hillel Y 1964 *Language and Information*. Addison-Wesley, Reading, MA
- Bauer L 1988 A descriptive gap in morphology. *Yearbook of Morphology* 1: 17-27
- Hammond M 1990 Subaffixation
- Hoeksema J 1985 *Categorical Morphology*. Garland Press, New York
- Hoeksema J 1988 Head-types in morpho-syntax. *Yearbook of Morphology* 1: 123-27
- Hoeksema J 1991 Categorical morphology and the valency of nouns. In: Aronoff M (ed.) *Morphology Now*. SUNY Press, Albany, NY
- Hoeksema J, Janda R D 1988 Implications of process morphology for categorial grammar. In: Oehrle, et al. (eds.) *Categorical Grammars and Natural Language Structures*. Reidel, Dordrecht
- Kang B 1987 Constraining type-lifting in categorial grammar. In: Miller A, Powers J (eds.) *Proceedings of the 4th Eastern States Conference on Linguistics*. Columbus, OH
- Lambek J 1958 The mathematics of sentence structure. *American Mathematical Monthly* 65: 154-70
- Moortgat M 1985 Functional composition and complement inheritance. In: Hoppenbrouwers G, Seuren P A M, Weijters J (eds.) *Meaning and the Lexicon*. Foris, Dordrecht
- Moortgat M 1988a Mixed composition and discontinuous dependencies. In: Oehrle et al. *Categorical Grammars and Natural Language Structures*. Reidel, Dordrecht
- Moortgat M 1988b *Categorical Investigations 2*. Foris, Dordrecht
- Oehrle R T, Bach E, Wheeler D W (eds.) 1988 *Categorical Grammars and Natural Language Structures*. Reidel, Dordrecht
- Scalise S 1988 The notion of 'head' in morphology. In: *Yearbook of Morphology* 1: 229-45
- Steedman M J 1985 Dependency and coordination in the grammar of Dutch and English. *Lg* 61: 523-68
- Wheeler D W 1988 Consequences of some categorially-motivated phonological assumptions. In: Oehrle et al. *Categorical Grammars and Natural Language Structures*. Reidel, Dordrecht

J. Hoeksema

Morphology, Computational

Computational morphology (CM) is a subbranch of computational linguistics. CM deals with theories and tools for the computational analysis and synthesis (production, generation) of word-forms. It also deals with morphological tagging of texts, i.e., supplying the word-forms with part of speech and other grammatical labels. Statistical methods are becoming increasingly important. Computational morphological descriptions of individual languages are central components in many natural language parsing systems. CM has several practical applications.

1. Background

Early CM (up to the middle 1970s) tended to be both language-dependent and machine-dependent. Not much attention was paid to the problem of designing a language-independent theory of CM. Normally, only restricted subparts of the vocabulary were considered. Because most systems dealt with English, and English morphology is

superficially 'simple,' the importance of morphological analysis in computational natural language systems was underestimated.

An opinion was widely held, and partly still is, that all word-forms of a language are listed as such, with the requisite morphological information, in a dictionary (the Full Listing Hypothesis).

Most work in CM has dealt with written language. In what follows, however, terms like 'phoneme,' 'phonological,' and 'morphophonological' will be used even if terms such as 'grapheme,' 'graphemic,' and 'morphographemic' would often be more descriptive of real practice.

Ron Kaplan and Martin Kay's paper (1981) on treating phonological rules (in the sense of generative phonology) as finite state transducers is a landmark in CM. The central tenets of this paper were published as part of Kay (1983). The basic idea is to translate or compile (morpho)phonological rules into two-tape finite state transducers. One important consequence of this view is that such rules emerge as 'simple' in the sense that they may be modeled by formally simple, well-understood, abstract machines.

2. Two-level Computational Morphology

Building upon the foundation laid by Kaplan and Kay, Kimmo Koskenniemi (1983a, 1983b, 1984) designed his two-level theory of computational morphology (TWOL) which was to become almost a standard in the CM of the late 1980s. Even if TWOL originally was not claimed to be a genuine linguistic theory of morphology, it has possible interpretations in this respect, especially in comparison to orthodox generative phonology of the type proposed in Chomsky and Halle's *Sound Pattern of English* (1968).

The two levels of TWOL are the lexicon (the lexical representations) and the surface word-forms. Both are strings of characters. Lexical representations are strings of phonemes and/or morphophonemes. Surface word-forms are strings of phonemes. The two levels are (or rather, may be) mediated by a set of rules accounting for those morphophonological alternations that are deemed by the linguist to be expressed as rules rather than as, say, Item and Arrangement (1A) type alternation configurations.

The morphotactic structure of the language is captured in the lexicon by a potentially recursive set of pointers to relevant classes of endings. In the most general case, a full lexical entry is a triple:

<form, features, ending-pointer>

where 'form' denotes the base-form chosen, 'features' are morphological, syntactic, semantic, etc.; features to be retrieved when a match is encountered; and 'ending-pointer' is a reference to endings possible in the next morphotactic position.

Thus, consider the Swedish neuter noun *lager* 'store' with the following forms in its paradigm:

	SINGULAR	PLURAL
NOMINATIVE INDEFINITE	lager	lager
NOMINATIVE DEFINITE	lagret	lagren
GENITIVE INDEFINITE	lagers	lagers
GENITIVE DEFINITE	lagrets	lagrens

One possible TWOL description of these forms is the following.

```

LEXICON MAIN
...
lag ERRET;
...

LEXICON ERRET
= "er N NEU INDEF SG" NOMGEN;
= "er N NEU INDEF PL" NOMGEN;
= "er N NEU DEF SG" NOMGEN;
= "er N NEU DEF PL" NOMGEN;
= "er N NEU" MAIN;

LEXICON NOMGEN
= "NOM";
= "GEN";

```

LEXICON MAIN is the main large repository of the free morphemes of the language. Material within double quotes is features to be retrieved. "#" is a designated symbol denoting that the end of the word-form has been reached (there are no more characters to match to the lexical ones). The semicolon ";" denotes the end of the entry.

This description is of IA type and invokes no rules. Thus, when the TWOL program is properly fed with this description and given the word-form *lagrens* to analyze, the first three letters match *lag* in LEXICON MAIN whereupon a reference to LEXICON ERRET is obtained (but no features have yet been retrieved apart from the segments *lag*). Now there are five options available, one of which is *ren*. This is matched; *er* touching the initial double quote is interpreted as segments (for retrieving the base-form); the features "N NEU DEF PL" are retrieved; and a final reference to LEXICON NOMGEN is obtained. This provides as one option word-final *s* which is successfully matched, yielding the feature "GEN" to add to those already retrieved. All surface segments were successfully matched to lexical segments, i.e., the word-form *lagrens* was a legal one (in regard to the current grammar), its base form is *lager*, and its morphological structure "N NEU DEF PL GEN".

Observe the recursive reference back to LEXICON MAIN in LEXICON ERRET. This link caters for productive compound formation.

An alternative description would be to postulate a lexical morphophoneme *E* representing the alternation *e/0* plus a TWOL rule accounting for its realization (default realization is *e*). The character "%" is a morphological feature identifying endings affecting the realization of *E*. Now the lexicons look somewhat different:

```

LEXICON MAIN
...
lagEr ERRET;
...

LEXICON ERRET
= "N NEU INDEF SG" NOMGEN;
= "N NEU INDEF PL" NOMGEN;
= "N NEU DEF SG" NOMGEN;
= "N NEU DEF PL" NOMGEN;
= "N NEU" MAIN;

LEXICON NOMGEN
= "NOM";
= "GEN";

```

```

TWO-LEVEL RULE
=====
E <=> r %
=
r =

```

The TWOL rule with the operator " \Leftrightarrow " (one rule type

out of three originally postulated by Koskenniemi) spells out the necessary and sufficient conditions for the realization. It requires that the lexical character *E* must correspond to surface zero if followed by a character pair consisting of a lexical *r* corresponding to a surface *r*, and this in turn is followed by a character pair consisting of the lexical segment % (another morphophoneme just marking off the ending codas *et*, *en*) the surface realization of which does not matter (=). (Actually, there should be a realization rule for % as well, which is omitted here for brevity.) In other contexts, *E* is realized as default *e*. Forms like *lagren*, *lagrets* will thus be properly accepted and analyzed whereas ungrammatical forms like **lageren*, **lagerets* are properly rejected.

One of the most central features of TWOL morphology is that TWOL rules are compiled into finite state automata. Originally (Koskenniemi 1983a), the automata had to be compiled by hand which required both time and excessive care. However, a rule compiler was developed by Karttunen, et al. (1987) which eliminates the need for hand-coding. Requisite finite state automata are automatically derived from the rules postulated by the linguist.

Run-time automata are executed in parallel. This is an important property which, inter alia, eliminates the need for rule ordering much discussed in generative phonology. Each TWOL rule expresses a singular true fact which is independent of what other rules the linguist might have postulated.

There also are no intermediate representations in excess of the two levels (lexicon and surface). Nothing is derived directionally 'from' anything else. Rather, TWOL rules state allowed and required correspondences between the two levels, such as the contextual restrictions on the correspondence *E/0* above.

Since there is no directionality in the description, a TWOL morphology may be run both as an analyzer and as a synthesizer, i.e., it is possible both to analyze word-forms (retrieve their morphological structure), and to produce (generate) word-forms from the proper string of lexical segments. For example, a correspondence such as:

```

LEXICON: lagEr%ets
SURFACE: lag0r0ets

```

may be viewed both ways. Seen from the surface, the proper lexical form is *lagEr%ets*. Seen from the lexicon, the proper surface form is *lagrets*. (An additional step is required for generating a form directly from a sequence consisting of base-form + morphological features such as "lager N NEU DEF SG GEN".)

TWOL morphology is truly language-independent. Successful descriptions, particularly including full inflectional morphology, have been designed for more than 20 languages, such as Finnish, Estonian, Hungarian, Lappish, Cheremis, Swedish, English, French, Romanian, Russian, Polish, Old Church Slavonic, Sanskrit, Ancient Greek, Swahili, Japanese, and Arabic. The Russian TWOL description is the largest so far developed, and its lexicon contains 70,000 entries. The English lexicon contains more than 50,000 entries, the Swedish lexicon 48,000 entries. In addition to full inflectional (and rule-based derivational) morphology, these descriptions (all developed in Finland) have full capability of analyzing productively formed compounds.

Kay (1987) discusses some important ramifications of TWOL computational morphology, especially in regard to the treatment of nonconcatenative phenomena such as those found in Semitic languages (see *Morphology, Nonconcatenative*). Kay also provides a more unified treatment of morphophonological alternations (see *Morphophonemics*) and morphotactics than is usual in CM.

3. Statistical Approaches to Computational Morphology

Statistical (probabilistic, stochastic) approaches have been developed, especially in the framework of projects aimed at the automatic 'part of speech' tagging of large text corpora. It is reasonable to regard these as belonging to CM insofar as they are concerned with assignment of basic-level (above all inflectional) grammatical information.

The statistical approaches differ from the rule-based ones in several respects. They do not postulate ordinary grammatical rules as they do not set out to characterize grammatically correct words or sentences. Rather, their task is to assign the best possible analysis to any conceivable input, grammatical or not. Often they make the best guess available, normally based on bigram or trigram analysis (i.e., sequences of codes assigned to two or three consecutive word-forms). The best statistical approaches are robust in the sense that they can cope with authentic running text. By definition, they run the risk of making occasional errors.

The basic task of such statistically based analyzers is to assign proper part of speech information and especially to perform homograph separation, i.e., to pick the correct code in ambiguous situations. A good example is the English word *round* which is morphologically at least five ways ambiguous (A, N, V, PREP, ADV). Picking the right code out of this set is obviously not a trivial task.

The classic of statistically based morphological tagging is the TAGGIT program written by Greene and Rubin (1971) for tagging the Brown Corpus of American English (Kučera and Francis 1967). TAGGIT achieved a success rate of some 78 percent in separating the homographs of running text.

A novel approach called Constituent-likelihood Automatic Word-tagging System (CLAWS) was developed for tagging the London-Oslo-Bergen (LOB) Corpus of British English in 1978-83 (see the articles in Garside, et al. 1987). Each word should be supplied with one tag drawn from a set of 133 tags. In the lexicon, each word-form is supplied with a list of its potential parts of speech in descending frequency order (NN1 = singular common noun, JJ = general adjective, vv0 = verbal base-form). This is the step of tag assignment:

```
...
consultant NN1
consummate JJ vv0
contact NN1 vv0
...
```

A module called CHAINPROBS has the important function of marking the most probable sequence of tags (tag selection). The success rate reported is 96-97 percent.

A similar success rate is reported by Church (1988) who uses a linear time dynamic programming algorithm to assign parts of speech to words. The algorithm optimizes the product of lexical probabilities and contextual probabilities.

4. Tools

TWOL implementations have been done in several programming languages such as C, PASCAL, and many dialects of LISP. A commercially available version is thoroughly described by Antworth (1990).

The Beta program developed by Benny Brodda, Stockholm (Brodda 1990), is a versatile tool for many central purposes of CM, e.g., preprocessing and standardizing texts, doing concordances, performing tagging, and doing morphological analysis.

5. Applications

CM, often in conjunction with modern compression technology, has found extensive applications especially in machine translation, information storage and retrieval, text-to-speech synthesis, and in the practice of spelling verification and spelling correction.

Bibliography

- Antworth E L 1990 *PC-KIMMO: A Two-level Processor for Morphological Analysis*. Summer Institute of Linguistics, Dallas, TX
- Brodda B 1990 Corpus work with PC Beta: A presentation. In: Karlgren H (ed.) *Papers Presented to the 13th International Conference on Computational Linguistics*, vol. 3. Yliopistopaino, Helsinki
- Church K 1988 A stochastic parts program and noun phrase parser for unrestricted text. In: *Second Conference on Applied Natural Language Processing: Proceedings of the Conference, 9-12 Feb., 1988*. Association for Computational Linguistics, Austin, TX
- Garside R, Leech G, Sampson G (eds.) 1987 *Computational Analysis of English: A Corpus-based Approach*. Longman, London
- Greene B B, Rubin G M 1971 *Automatic Grammatical Tagging of English*. Department of Linguistics, Brown University, Providence, RI
- Kaplan R M, Kay M 1981 Phonological rules as finite state transducers. Paper presented at the Winter Meeting of the Linguistic Society of America
- Karlsson F 1989 Computational testing of linguistic models in morphology. In: Batori I S, Lenders W, Putschke W (eds.) *Computational Linguistics*. Walter de Gruyter, Berlin
- Karttunen L, Koskenniemi K, Kaplan R M 1987 A compiler for two-level phonological rules. In: Dalrymple M E, *Tools for Morphological Analysis*. Report No. CSLI-87-108. Center for the Study of Language and Information, Stanford University, Stanford, CA
- Kay M 1983 When meta-rules are not meta-rules. In: Jones K S, Wilks Y (eds.) *Automatic Natural Language Parsing*. Ellis Horwood, Chichester
- Kay M 1987 Nonconcatenative finite-state morphology. In: *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics, 1-3 April, 1987*. University of Copenhagen, Copenhagen
- Koskenniemi K 1983a *Two-level Morphology: A General Computational Model for Word-form Recognition and Production*. University of Helsinki, Department of General Linguistics, Helsinki Publications no. 11
- Koskenniemi K 1983b Two-level morphology for morphological analysis. In: Bundy A (ed.) *Proceedings of the Eighth International Joint Conference on Artificial Intelligence, 8-12 Aug. 1983, Karlsruhe, W. Germany*. W. Kaufmann, Los Altos, CA
- Koskenniemi K 1984 A general computational model for word-form recognition and production. In: *Proceedings of Coling-84 Tenth International Conference on Computational Linguistics*

2-6 July, 1984, Stanford University. Association for Computational Linguistics, Austin, TX

Kučera H, Francis W N 1967 *Computational Analysis of Present-day American English*. Brown University Press, Providence, RI

F. Karlsson

Morphology: History

From the Greek philosophers to the generative grammarians, morphology has shifted its focus from the morphological categories to phonological form. Although the Greeks were aware that words are bilateral signs, they considered the word the smallest indivisible element and hence explored semantic, not formal, categories. With the discovery of the formal analyses of the Indian grammarians at the end of the eighteenth century interest began to shift. Humboldt compared languages strictly by their morphological form and grammars throughout the remainder of the nineteenth century reflected his preference. By the middle of the twentieth century, morphological research was so absorbed by allomorphic studies that Bloomfield denied any relevance of semantics to morphology at all.

1. The Ancient Grammarians

1.1 *The Babylonian Grammarians (ca. 1600 BC)*

The ancient Babylonians left the earliest recorded evidence of morphological analysis. In order to preserve the literary traditions of Sumerian after Akkadian had become the conversational norm, Babylonians of the post-Sumerian era wrote grammatical texts in cuneiform on clay tablets. These texts organize Sumerian adverbs, pronouns, and verbs according to Paradigms (see *Paradigms*), divided neatly by straight lines and translated into Akkadian. Although the paradigms are not named, the texts consistently reflected person, number, tense, mood, and other recognizable morphological categories (see *Babylonian Grammatical Texts*).

1.2 *Pāṇini and the Indian Grammarians*

The Indian grammars from Pāṇini's *Aṣṭādhyāyī* (ca. 500 BC) distinguished derivation and inflection and contained formal rules for grammatical units below the word level, e.g., Pāṇini's affixes (*pratyaya*) and augments, which Kātyāyana called *āgama*. The affixes could be replaced before the surface level or deleted to accommodate 'zero' morphology. 'Empty' realizations were also possible. Several Indian terms continue to prove useful for contemporary morphology, e.g., *bahuvrīhi* and *dvandva* compounds (see *Word-formation: Compounding; Pāṇini; Sanskrit Grammatical Terms*).

2. The Classical Tradition

2.1 *The Greeks*

In *Kratylos*, Aristotle first raised the issue of the arbitrary relation of meaning to sound but the Greek philosophers were primarily interested in the categories of meaning. Protagoras and Plato had separated nouns from verbs, including adjectives among the latter, and begun investigations

into gender. Aristotle advanced this categorization, lumping all other words into the category 'conjunctions' or 'connectors' (*śūndesmoi*), on the basis of their having no meaning in isolation, only grammatical function. The distinction closely parallels the current distinction of lexical and grammatical morphemes.

The Stoics (Diogenes, Laertes, Apollonius) elevated grammar to the level of a major discipline within philosophy. They defined the distinction between 'the signifier' (*tò sēmainon*) and 'the signified' (*tò sēmainómenon*) and first questioned the difference between common and proper nouns. They distinguished four parts of speech: nouns, articles, verbs, and conjunctions, subdividing articles into personal and demonstrative pronouns and the definite article. 'Conjunctions' included all words which bind together the other parts of speech in sentences, e.g., prepositions. The Stoics isolated case and first distinguished the 'upright' case, nominative, from the other, 'oblique' cases. Their names of the central cases remain. The Stoics discovered the disparity between natural number and gender and the corresponding grammatical categories. They also first defined the passive and active voices and established four tenses based on two times and two aspects.

The Alexandrian grammarians changed language study from a subdiscipline of philosophy to an independent 'technical' discipline. They defined categories in terms of the formal characteristics of their inflectional paradigms rather than their semantics. Aristarchus (216-144 BC) and Dionysius Thrax (ca. 170-90 BC) categorized words into the canonical eight parts of speech, but without distinguishing derivation and inflection. Dionysius perfected the description of the verbal conjugation system, classifying the tenses as present, past and future; divided the past into two aspects; and described three voices: active, passive, and middle.

2.2 *The Latin Grammarians*

In *De Lingua Latina* (47-45 BC), Marcus Varro categorized the parts of speech according to whether they reflected case or tense, in a way quite similar to the current generativist \pm mechanism, [\pm Noun, \pm Verb]. However, Varro reached a different conclusion as to the major parts of speech.

- Nouns and adjectives have case but no tense.
- Verbs have no case but tense.
- Participles have case and tense.
- Adverbs, conjunctions, etc. have neither case nor tense.

Varro defined singularis and pluralis tantum nouns, and distinguished potential from attested forms. He noted that *unguentum* 'perfume' has a plural *urguenta* because of the existence of several kinds of perfume. Were similar differences in the kinds of olive oil and vinegar to arise, so would plurals *olea* 'olive oils' and *aceta* 'vinegars.' Varro is perhaps best known for his discussions of the extensive violations (*anomaly*) of derivational regularity, Thrax's *analogia*, e.g., indeclinable nouns, irregular comparatives like *bonum, melius, optimum*, and derivational irregularities like *vinum* 'wine': *vinaria* 'wineshop,' *unguentum* 'perfume': *unguentaria* 'perfume shop' but *caro* 'meat': *laniena* 'butcher's shop' instead of *carnaria*.